mda

methods, data, analyses

Journal for Quantitative Methods and Survey Methodology

Volume 16, 2022 | 1

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 | Vol. 16(1), 2022, pp. 3-32 DOI: 10.12758/mda.2021.08

Gender and Survey Participation. An Event History Analysis of the Gender Effects of Survey Participation in a Probability-based Multi-wave Panel Study with a Sequential Mixed-mode Design

Rolf Becker

University of Bern

Abstract

In cross-sectional surveys, as well as in longitudinal panel studies, systematic gender differences in survey participation are routinely observed. Since there has been little research on this issue, this study seeks to reveal this association for web-based online surveys and computer-assisted telephone interviews in the context of a sequential mixed-mode design with a push-to-web method. Based on diverse versions of benefit–cost theories relating to deliberative and heuristic decision-making, several hypotheses are deduced and then tested by longitudinal data in the context of a multi-wave panel study on the educational and occupational trajectories of juveniles. Employing event history data on the survey participation of young panelists living in German-speaking cantons in Switzerland and matching them with geographical data at the macro level and panel characteristics at the meso level, none of the hypotheses is confirmed empirically. It is concluded that indirect measures of an individual's perceptions of a situation, and of the benefits and costs as well as the process and mechanisms of the decision relating to survey participation, are insufficient to explain this gender difference. Direct tests of these theoretical approaches are needed in future.

Keywords: Gender; survey participation; nonresponse; event history analysis; societal environment; panel study; web-based online survey; sequential mixed-mode design; push-to-web method

Against the background of decreasing response rates in modern societies with a high level of prosperity, the number of empirical studies on survey participation and nonresponse in the social sciences is increasing (e.g. Leeper, 2019; Beullens et al., 2018; Dutwin & Lavrakas, 2016; Keusch, 2015; Tourangeau & Plewes, 2013; Brüggen et al., 2011; Stoop et al., 2010; Groves & Peytcheva, 2008; Groves, 2006; Groves et al., 2001; de Heer, 1999; Smith, 1995; Goyder & Leiper, 1985; Steeh, 1981). One of the main findings is that the decrease in response rates is generally observed for cross-sectional surveys, while the participation rate within longitudinal studies, such as panel studies, remains high (Becker et al., 2019; Brick & Williams, 2013; Schoeni et al., 2013: 84–85). Among these studies, a constant *gender effect in survey participation and nonresponse* is observed as a social phenomenon (Slauson-Blevins & Johnson, 2016: 428; Keusch, 2015: 186; Busby & Yoshida, 2013; Dykema et al., 2012; Laguilles et al., 2011; Stoop et al., 2010: 10, 20; Couper et al., 2008: 260; Marcus & Schütz, 2005; Patrick et al., 2013; Porter & Whitcomb, 2005; Kwak & Radler, 2002: 259; Curtin et al., 2000: 419; Singer et al., 2000: 180; Green, 1996: 176; Dalecki et al., 1988: 54). Particularly for mail or web surveys, it is frequently found that women are more likely to respond than men (Green, 1996: 176; Becker & Glauser, 2018). Furthermore, women seem to be more likely than men to respond promptly after the invitation to take part in an online survey (Becker, 2021; Becker et al., 2019). Finally, over the last several years, it has been found for different survey modes and survey topics that the gender effect on the rate of survey participation remains clear, even though response rates are declining overall (Slauson-Blevins & Johnson, 2016). However, it is still not known whether the gender difference in participation rates changes across surveys in a multi-wave panel.

Acknowledgements

The data for the first seven waves of the panel study are available as Scientific Use Files at FORS in Lausanne and can be found in the online catalogue under the reference number 10773 (https://forsbase.unil.ch/project/study-public-overview/15802/0/). The data of Wave 8 will be available in 2021. These scientific use files include the paradata of the fieldwork periods.

For helpful comments on an earlier version, I wish to thank the anonymous reviewers and the *mda* editor Sabine Häder. The author is responsible for the remaining shortcomings.

Funding

The DAB panel study is substantially financed by the State Secretariat for Education, Research and Innovation (SERI). The interpretations and conclusions are those of the author and do not necessarily represent the views of the SERI.

Direct correspondence to

Rolf Becker, University of Bern, Department of Sociology of Education,
Fabrikstrasse 8, CH–3012 Bern, Switzerland
E-mail: rolf.becker@edu.unibe.ch

In this respect, Green (1996: 176) states there is too little research on gender as it relates to surveys to reach a conclusion. However, there are several "ad-hoc explanations" of the effects of gender on response rates. Slauson-Blevins and Johnson (2016), for example, emphasize that the lower inclination of men to take part in scientific surveys might cause the decrease in their rate of response. "Gender differences in survey participation are partially attributable to the difficulty of contacting male participants rather than outright refusals to participate (…). Yet while survey researchers, often conclude that gender differences exist, there has been little effort to conceptually understand this difference" (Slauson-Blevins & Johnson, 2016: 428). Another explanation that seems plausible attributes gender disparities in survey response to differences in socialization regarding norms around helping, or differences in susceptibility to social influence. External circumstances, such as access to the Internet in a prosperous country like Switzerland, do not contribute to the explanation of these gender disparities since there is no "digital divide" in Internet use across the genders (BFS, 2021).

Focusing on self-administered survey modes, such as an online questionnaire or administered computer-assisted telephone interview (CATI), the question is still unsolved regarding *why* gender has been found to play a significant role in survey participation and response to questionnaires or interviewers, with women responding in greater proportion than men (Porter & Whitcomb, 2005). Likewise, it is unclear *why* female and male online panelists are motivated differently (Slauson-Blevins & Johnson, 2016; Göritz & Stieger, 2009). Are there gender-specific motivations, resources, and circumstances that drive male and female invitees to participate in different ways? Against the theoretical background of a diverse variety of rational action theories that take heuristic decision-making process into account, the *main question* asked in this empirical contribution is as follows: Are gender differences in survey participation a fundamental phenomenon or are they epiphenomenal to other factors, such as social origin and class-related socialization in terms of educational level and achievement? In other words: are gender differences a singular phenomenon observed for cross-sectional surveys or in early waves in a multi-wave panel study? Do gender disparities in surveys disappear when we control for a number of covariates, which correlates with the propensity toward survey participation?

To answer these research questions, longitudinal data on survey participation are needed. The optimal type of longitudinal data would be the observation of a target person's survey participation across their life course, including time-variant information on their resources, circumstances and preferences. Since such data combining individual information on target persons with survey paradata are rare, the measurement of survey participation in a panel study provides a suboptimal surrogate. Therefore, data on the survey participation of female and male panelists collected since 2012 in an event history design are utilized in this contribu-

tion. This type of data, collected in the context of a multi-wave panel study on educational and occupational trajectories of juveniles born around 1997 and living in German-speaking cantons in Switzerland, makes it possible to analyze gender differences in overall survey nonresponse, the development of these gender differences during the fieldwork, and changes in them across surveys for a single target sample familiar with Internet and mobile devices.

In the remainder of this contribution, the next section outlines the theoretical background, as well as the hypotheses to be tested. The following section comprises a description of the data, design, statistical procedures, and the variables. After then, the empirical findings are presented. The final section gives a summary and conclusion.

Theoretical Background

Despite the fact that there is no theoretical vacuum regarding survey methodologies, Singer (2011: 379) concluded that, although various theories of survey participation exist, we know comparatively little about *why* individuals are willing or are not willing to participate, and about *how* they decide to take part in (or refuse to take part in) a scientific survey. Although different versions of rational action theories – such as social exchange theory (Dillman, 2000), the theory of subjective utility (Becker & Mehlkop, 2011), leverage-salience theory (Groves et al., 2000) or the social-psychological approach on habitual-heuristic action (Groves et al., 1992) – all assume that survey participation is based on a deliberative assessment of the benefits and costs of survey participation, or on an automatic-spontaneous decision, there is no comparative test of these approaches in theoretical and empirical respects. However, such a systematic test is needed to confirm Singer's (2011: 388) conclusion that the general benefit-cost theory of survey participation can be seen as a synthesis of principles derived from these other theories (Goyder, Boyer, & Martinelli, 2008). Thus, it is unclear whether gender differences in survey participation can be explained by a target person's reasoned judgment that the benefits of acting outweigh the costs, or by an almost instantaneous cognitive procedure with the help of heuristics (Singer, 2011: 381).

According to these approaches, survey participation is a function of subjective perceived costs and benefits of survey participation, as well as the subjective expected probability of successful realization of perceived benefits p: $f(sp) = p \cdot B - C$. The decision regarding participation or refusal is based on a subjective assessment of subjective expected utilities *SEU(.)* of different alternatives *da*, other than survey participation *sp*. If $SEU(sp) = p_{sp} \cdot B_{sp} - C_{sp} > SEU(da) = p_{da} \cdot B_{da} - C_{da}$, it is likely an eligible individual will indeed take part in the survey.

However, what gender-specific costs or, in particular, gender-specific benefits of survey participation might there be? According to Singer (2011), a potential

respondent's decisions depend mainly on benefits, not on costs or perceptions of risk or harm. Why should female target persons systematically perceive increased benefits resulting from survey response than their male counterparts? A plausible answer might be that it is the *gender-specific expectations of success probability* that result in gender differences regarding the benefits of survey participation. These expectations might be based on an individual's skills – such as literacy or computer skills – and the related confidence in their own abilities (e.g. persistence; decisiveness; internal or external control beliefs). In theoretical respects, this assumption is based on empirical evidence that girls and young women have better educational achievement, higher language proficiency and more advanced literacy and educational attainment than boys and male adolescents (DiPrete & Buchmann, 2013; Beck et al., 2010; Becker & Müller, 2011; Buchmann et al., 2008). However, it has to be taken into account that there is a stark correlation between educational success, educational attainment and social origin among the genders. These educational advantages in favor of female target persons indicate that cognitive burden, uncertainty in interview situations and insufficient language ability and proficiency are much lower for women compared to men. Since the transaction costs of survey response are relatively lower for women, they are more likely to take part in a scientific survey than male potential respondents.

Hypothesis 1: Controlling for social origin, educational attainment and achievements (indicating language proficiency and language ability, as well as educational success and motivation, mostly in favor of women) are positively correlated with survey participation. Due to the advantage of female panelists in educational success and achievements, the gender effect becomes insignificant when these dimensions are taken into account.

Furthermore, the correlations between gender, educational attainment, social origin and the rural-urban divide in educational opportunity are evident (Glauser & Becker, 2016; Sixt, 2013). In sum, according to Green (1996), education, intelligence and achievement, as well as socioeconomic status and living in rural areas, were found to correlate positively with survey response rate (Becker, 2021; Groves & Couper, 1998; Dalecki et al., 1988: 54).

Hypothesis 2: By controlling additionally for regional opportunity structures in terms of a potential respondent's place of residence in a rural or urban area, the gender difference in survey participation diminishes in multivariate estimations.

Success in educational attainment is correlated with favorite educational returns in an individual's working life. Although women profited from educational expansion (e.g. Becker & Mayer, 2019; Becker & Müller, 2011), recent research has argued that the opportunity costs of survey participation are higher for men, as women are more likely to stay at home (Stoop et al., 2010: 20). Women are therefore more likely to be reachable and ready to take part in a survey. But this line of reasoning might not be valid for self-administered web-based online interviews, since

for these invitees can decide themselves whether and when they will start completing the questionnaire – e.g. after the working day, at the weekend or at another point in time suitable for them. This might be also true for CATI due to widespread availability of mobile devices. For juveniles in particular, it is confirmed that most possess a smartphone instead of a fixed phone line (BFS, 2021).

If there are gender-based preferences for survey modes – i.e. that men are more likely to prefer computer-assisted web-based interviews (CAWI) (due to their technical affinity) and women the CATI mode (due to their language abilities) – it could be assumed that there is no gender difference of response in surveys with a sequential mixed-mode design. Since it is often observed, even for online surveys, that women tend to respond earlier than men after survey launch (e.g. Göritz, 2014; Göritz & Stieger, 2009), a sequential mixed-mode design offering CAWI and CATI modes could have the potential to compensate for the gender-based likelihood of participation in different survey modes in the long run of the fieldwork period.

Hypothesis 3: In a sequential mixed-mode design, the gender difference in survey participation diminishes across the running fieldwork period and the offered survey modes.

Furthermore, Green (1996) argues that gender differences may exist in survey response due to differences in (primary) socialization regarding differences in susceptibility to social influence or helping norms. These aspects correspond with findings by Porst and von Briel (1995: 11) that, besides personal and situative aspects, women are more likely to respond to surveys due to altruism (Porst & von Biel, 1995: 15). In line with theoretical arguments on gender-based secondary and tertiary socialization across the life course, it seems that girls and women display a different social character than boys and men (e.g. Grunow, 2013). For example, compared to their male counterpart, they are more likely to have learned to carry out a task – such as the request of another person or completing a questionnaire – in an autonomous, precise and persistent way (e.g. Quenzel & Hurrelmann, 2010). According to Green (1996: 181), women are therefore more communicative and interested in sharing opinions with others.

Hypothesis 4: Gender differences in survey participation disappear in multivariate estimations when controlling for personality traits and individual beliefs, indicating at least some facets of gender-specific socialization.

For male youth and adults oriented toward traditional "*masculinity norms*" or the "*male breadwinner model*", it has been observed they are less interested in tasks such as reading and writing, as well as in constructive communication with other persons and authorities (e.g. Hadjar, 2011). On the one hand, this again means that personality traits (such as persistence and decisiveness or internal and external control beliefs) could help explain the gender differences in survey participation (e.g. Porst & von Briel, 1995). One the other hand, it seems that the gender differences result from the low propensity toward survey participation observed for male

target persons oriented toward the traditional gender stereotypes and gendered life courses.

Hypothesis 5: Gender differences in survey participation are statistically dissolved by taking a panelist's orientation toward gender roles into account, as well as their personality traits and other individual skills.

Data, Design, Variables and Statistical Procedures

Data set

The empirical analysis is based on longitudinal data of a probability-based multiwave panel study about the determinants of educational choice and training opportunities (for details, see Becker et al., 2020). This project started in 2012. The last survey was realized in May/June 2020. Data and paradata were collected in a sequential mixed-mode design with a push-to-web method (see also: Kreuter, 2013). The first mode was an online survey, followed by a CATI and, in a selected number of surveys, a paper-and-pencil interview (PAPI) by mail survey. The initial target population comprised eighth-graders in the 2011/12 school year (born around 1997), who were enrolled in regular classes in public schools in German-speaking cantons of Switzerland. The panel data are based on a random and 10 per cent stratified gross sample of 296 school classes, out of a total universe of 3,045 classes. A disproportional sampling of school classes from different school types, as well as a proportional sampling of school classes regarding share of migrants within schools, was applied. At the school level, a simple random sample of school classes was chosen. The initial probability sampling was based on data obtained from the Swiss Federal Statistical Office (FSO) (for details, see Glauser, 2015).

In the first three waves, the contacted panelists ($n \approx 3{,}800$) were interviewed in the context of their school class. After leaving the compulsory school, the panelists were pursued individually after the fourth wave. Each of the eligible and contactable panelists was invited for the surveys, even when they had skipped a wave. To improve the response rate, the panelists received unconditionally prepaid material incentives or cash in hand (Becker et al., 2019). Across the panel waves, the overall response rate was constant at about 80 per cent (*Table 1*). The response rate for online survey increased from 46 per cent in Wave 4 to 76 per cent in Wave 8, while the response rate for the CATI decreased from 38 per cent to 5 percent.

The proportion of women among the invitees was rather constant, at 50 per cent in Waves 4 and 5, 51 per cent in Wave 6 and 52 per cent in the Waves 7 and 8 at the start of survey launch. At the start of the risk time for CATI, about 47 per cent of the nonrespondents, i.e. invitees who had not taken part in the CAWI before, were female in Wave 4. Their share decreased to 43 per cent in Wave 6 and remained constant for the recent waves.

Table 1 Samples and response in the DAB panel

	Wave 4 Oct–Nov 2014	Wave 5 Jun–Aug 2016	Wave 6 May–Jun 2017	Wave 7 May–Jun 2018	Wave 8 May–Jun 2020
Sample size					
Contactable individuals	2,655	2,799	2,712	2,488	2,492
Type of survey					
Online survey	yes	yes	yes	yes	yes
CATI survey	yes	yes	yes	yes	yes
PAPI survey	no	no	yes	yes	no
Incentive	voucher	voucher	pen	money	money
Realized interviews					
Individuals	2,235	2,228	2,053	1,957	2,016
of whom: online	1,227	1,329	1,375	1,645	1,884
CATI	1,008	899	597	287	132
PAPI	0	0	81	25	0
Response rate in %					
Contactable individuals	84%	80%	76%	79%	81%
Online	46%	47%	51%	66%	76%
CATI	38%	32%	22%	12%	5%
PAPI	–	–	3%	1%	–

Source: DAB (own calculation)

For the analysis of gender effects on survey participation, the empirical analysis focuses on the online and CATI modes only since the number of participants in the PAPI mode was rather low (106 cases out of 13,145 target persons across six panel waves, i.e. a response rate of 3% in Wave 6 and 1% in Wave 7). The observation window was standardized to 52 days for methodological reasons, such as comparability between waves, low number of participants after seven weeks of fieldwork and right-censored data due to survey nonresponse. In the case of both survey modes, non- and under-coverage was rather low for this sample. About 93 per cent of the Swiss population has access to the Internet and they mostly use this medium every day of the week. Each of the young interviewees in this panel study had daily access to the Internet or possessed a telephone or other mobile device (BFS, 2021).

In total, 13,145 complete cases were available for analyzing gender-specific patterns of participation in at least one of the five surveys. Since time stamps – collected automatically by the online survey software *Unipark* or by the CATI software – indicated exact time references for the invitation sent by email or SMS and

the start of a panelist's response, it was possible to calculate the exact duration of episodes since survey launch on a daily basis (Becker, 2021; Durrant et al., 2013). For the analysis of participation in the CATI by nonrespondents in the initial survey mode, the waiting time was calculated on a daily basis by the difference between the invitation to the CATI mode and the data of the telephone interview. For the invitees who did not take part at all until the end of the fieldwork period, i.e. the censored cases, the waiting time was 52 days. The number of skipped events was negligible.

The distribution of these waiting times from invitation until an individual started the survey participation as a *stochastic event* was analyzed using the techniques and procedures of *event history analysis* (Blossfeld, Rohwer, & Schneider, 2019). This means that episodes of survey participation are the units to be analyzed. In this respect, it was possible simultaneously to analyze an individual's intention to participate in the survey and the timing of when they did so. At the aggregate level, the development of the response rate was observed across different points in time during the field period.

For our purpose, this data set provides additional advantages due to the survey design. Multiple waves, for example, ensure that a constant gender effect on survey participation is not random. These waves are associated with different prepaid incentives, but with the same features of survey management; it is therefore possible to reveal if a gender effect depends on the type of an incentive by controlling for cover letters (including the incentive), digital invitation and reminders. The sample consisted of members belonged to a single birth cohort. Therefore, their survey participation did not depend on the age of the panelists. The survey topic – their own educational and occupational trajectory – was a general one and not related directly to gender. The number of items on gender-related issues, such as gender-based socialization or gender-based inequalities, was rather limited because the primary task of the panel was the reconstruction of their educational trajectories and careers in the labor market. Each of the target persons was involved in training or employment so different states and time constraints in this regard did not matter for survey participation. In respect of sponsorship and authority, it was emphasized in the advance letter that the project was in receipt of a grant from a governmental agency and was conducted by the same researchers at a Swiss university. Furthermore, for the sampling, 106 regions – characterized by a certain spatial homogeneity and reflecting small partially cross-cantonal labor market areas with functional orientation toward centered and peripheral opportunities and living standards, in addition to urbanicity, population density and a lack of social cohesion (Couper & Groves, 1996: 174) – were considered (Glauser & Becker, 2016: 20). This allowed for an analysis of the rural-urban divide in gender-specific survey participation in terms of Internet access and living conditions. In sum, the data allowed for a dynamic longitudinal analysis by considering the macro, meso and micro level –

i.e. social environmental attributes, survey characteristics and respondent attributes – at the same time.

Dependent and independent variables

The *dependent variable* was the *time-dependent likelihood of participation* in the CAWI. In general, the participation rate was defined by the ratio of contacted target persons who completed the questionnaire or the telephone interview (RR1 according to AAPOR, 2016: 61; Tourangeau & Plewes, 2013: 11; Bethlehem, 2009: 213; Singer, 2006: 637). This variable was coded in the following way: "1" for participation in online survey, "2" for participation in the CATI mode and "0" for nonresponse or incomplete response. Across the five panel waves, a maximum of 0.1 per cent of respondents canceled their completion of the questionnaire in a survey.

The main *independent variable* was a panelist's *gender.* It was considered as a dummy variable, with men as the reference category. Another *covariate* was the individual's *educational level*, indicated by the *school type* in which they were enrolled in their compulsory schooling. The school type was a proxy for the individual's appreciation of the utility of social-scientific research and information-gathering activities associated with their education (Groves & Couper, 1998: 128). The following school types were distinguished along their basic, extended and advanced requirements: low, intermediate and academic level. The target person's *achievement* was measured by the (z-standardized) grade point average (GPA) in German taught at school; this covariate indicated their *cognitive resources* and *language proficiency* (Wenz, Al Baghal & Gaia, 2021). Using a dummy variable, it was controlled for that German was the first language, indicating the target person's *language ability* (Kleiner, Lipps & Ferrez, 2015). By the way, this indicator measured the impact of migration background – net of German mother tongue, educational level and social origin – on survey response (Kalter, Granato & Kristen, 2007). *Social origin* was taken into account as a proxy for the socioeconomic conditions in which the target persons grew up, including welfare, integration and environment (Groves & Couper, 1998: 30). This was indicated by the well-established class scheme suggested by Erikson and Goldthorpe (1992). Personal characteristics – such as *persistence*, *internal and external control belief* and *decisiveness* – were controlled for (Marcus & Schütz, 2015; Saßenroth, 2013). They were extracted

from a number of items by factor analysis.[1] The *gender role models* for women and men were considered – after their extraction by factor analysis – for the indication of gender-specific socialization.[2]

Another covariate was the *current panel wave*, indicating the effect of different prepaid incentives, such as vouchers (worth 10 Swiss Francs), a ballpoint pen (worth 2 Swiss Francs) or cash (10 Swiss Francs), as well as the panelist's experiences with this panel. The *opportunity structure of the region* in which the panelists live was measured by macro data on regional levels delivered by the Swiss FSO. In order to reduce complexity and to control for the high correlation of regional contextual characteristics, factor scores were extracted from these data (for details: Glauser & Becker, 2016). The 106 regions in the German-speaking cantons were characterized by a certain spatial homogeneity and reflected the principle of small, partially cross-cantonal labor market areas with functional orientation toward centered and peripheral opportunities and living standards, in addition to urbanicity, population density and a lack of social cohesion.

Statistical procedures

Overcoming the limits of comparative-static estimations of survey response, the techniques and statistical procedures of *event history analysis* were utilized (Bloss-

1 They were measured in the first and second waves. *Persistence* was measured by the respondent's agreement with the following five statements: "I do not like unfinished business"; "If I decide to accomplish something, I manage to see it through"; "I complete whatever I start"; "Even if I encounter difficulties, I persistently continue"; and "I even keep at a painstaking task until I have carried it through". The *control beliefs* were measured by six items indicating the respondent's internal and external locus of control, as suggested by Jakoby and Jacob (1999): "I like to take on responsibility"; "Making my own decisions instead of relying on fate has proved to be good for me"; "In the case of problems and resistance, I generally find ways and means to assert myself"; "Success depends on luck, not on performance"; and "I feel like I have little influence over what happens to me". *Decisiveness* was based on a question about the respondent's decision certainty: "Life is full of decisions that need to be taken. Which of the six statements apply to you?" The wordings of these statements were: "I am really unsure as to what I should decide and often waver back and forth"; "Others unsettle me in my decision"; "After making a decision, I have great doubts as to whether I really made the right decision"; "It is very hard for me to decide because there are so many possibilities"; and "When I make a decision, I stick to it". For each of these items, the agreement itself consisted of a scale of discrete values from 1 for "I strongly disagree" to 5 for "I strongly agree". In order to reduce complexity and to avoid multicollinearity, three factors – persistence, control beliefs and decisiveness – were extracted by factor analysis (*Table A-1* in the Appendix).

2 The items of gender role stereotyping are measured in the third wave. Separately for the genders, the respondents were asked their subjective view of whether it was interesting for women or men to be employed, to earn much income, to be successful in their career, to have children, to take care of the household and to be responsible for childcare. The possible answers ranged from "1" for complete rejection to "5" for complete agreement (*Table A-2* in the Appendix)

feld et al., 2019). In this contribution, the aim is to model the likelihood of survey participation – that is, the hazard rate – as a stochastic and time-variant function of individual resources, the settings of the survey and societal factors. This rate $r(t)$ is defined as the marginal value of the conditional probability of such an event occurring – namely the start of completing the questionnaire in a web-based online survey – in the time interval $(t, t + \Delta t)$ given that this event has not occurred before time t (Blossfeld et al., 2019: 29). Using this statistical procedure, it is possible to reveal impacts of x for the probable occurrence of survey participation as the event y to be investigated: $\Delta X_t \rightarrow \Delta \Pr(\Delta Y_{t'})$, whereby $t < t'$ by taking the timing of events into account.

Due to the *sequential mixed-mode design*, specialties of the timing of events had to be considered for the bivariate and multivariate analyses. In the sequential mixed-mode design of the DAB panel study, access to the online mode was possible for each of the invitees during the complete field period. Nonrespondents were asked to take part in the CATI mode about two weeks after survey launch. There was then a competing risk of taking part in one of the two offered modes, which are mutually exclusive during an overlapping risk period. A competing risk is an event – such as participation in one of the two survey modes – that either hinders the occurrence of the primary event of interest (e.g. participation in the online survey instead of CATI) or that modifies the chance that this event (e.g. participation in CATI) will occur (Noordzij et al., 2013: 2670). When eligible panelists prefer one mode or another, the unchosen mode cannot be realized at another point in time due to censoring. Panelists who have not started completing the questionnaire have the "*chance*" to take part in the CATI or online mode at a point in time that is convenient for them.

In the case of competing risks, the traditional survival analysis is inadequate for methodological reasons. Therefore, the *cumulative incidence competing risk method* was used to describe panelist participation patterns across the field period. For example, the *cause-specific cumulative incidence function* (CIF), which is the probability of survey participation before the end of field period, was estimated to reveal the risk of choosing one of the competing survey modes (Lambert, 2017). The CIF describes the incidence of the occurrence of an event while taking competing risks into account (Austin & Fine, 2017: 4293).

Furthermore, *parametric regression procedures* were used to estimate the impact of independent variables on the likelihood of interesting events. For this purpose, the *subdistribution hazards approach* by Fine and Gray (1999) was seen as the most appropriate method to use for analyzing competing risks (see also Schuster et al., 2020; Noordzij et al., 2013). By taking competing risks into account, the coefficients estimated by the *stcrreg* module implemented in the statistical package *Stata* could be used to compute the cumulative incidence of participation in one of the survey modes and to depict the hazards in a CIF plot (Austin & Fine, 2017).

Finally, non-parametrical procedures were utilized to describe gender differences in survey participation. On the other hand, the parametrical procedures were used in the sense of residual analysis. This means it was necessary to test whether the gender difference in survey participation "disappeared" when controlling for theoretically based variables. In this way, it was possible to decide if the gender difference was fundamental or based on other factors correlated with gender, such as educational level, language proficiency or socialization.

Empirical Results

Description of gender-specific participation rates

If one measures both the timing and the quantity of participation in the online survey across several panel waves, the gender disparity becomes obvious. In the left-hand panel in Figure 1, it is apparent for each of the panel waves that the likelihood of survey participation in terms of cumulative incidences was significantly higher for female panelists than for their male counterparts. In particular, the differences in speed and rates increased in the initial stage after the survey launch. After about two weeks, the development of these incidences was rather similar for both genders.

For Waves 4 and 5, the same patterns different for genders were observed for the cumulative incidence of participation in the consecutive CATI mode since it was offered to the nonresponding panelists (right-hand panel in *Figure 1*). While there were no gender differences in taking part in this mode for Wave 6, men were more likely to respond to CATI than women in the both most recent Waves 7 and 8.

It is evident for the CAWI mode that women started to complete the questionnaire earlier than male panelists. Across each of the waves, after 10 days, 50 per cent of the female panelists had taken part in the online survey, while after 15 days half of the male risk sample had completed the online questionnaire.[3] The situation was different for the CATI mode. For women, the median value for the CATI mode was 15 days since this survey mode was offered to the panelists; this value was 16 days for their male counterparts. For this survey mode, except for Waves 5 and 7, there were no systematic gender differences in participation. In sum, while 62 per cent of the female panelists took part in the online survey and 39 per cent of female nonrespondents, who do not responded in the CAWI mode yet, took part in the

3 For the CAWI mode, each of the tests – such as the *Wilcoxon-Breslan-Gehan* test, sensitive at the beginning of the process time, or the *Generalized Savage Log-rank* test, stressing increasing differences at the end of the process time – provided significant differences between the compared units, such as gender and waves (Blossfeld et al., 2019: 83). The null hypothesis that the timing and quantity of survey participation do not differ across genders and waves must therefore be rejected for the initial survey mode.

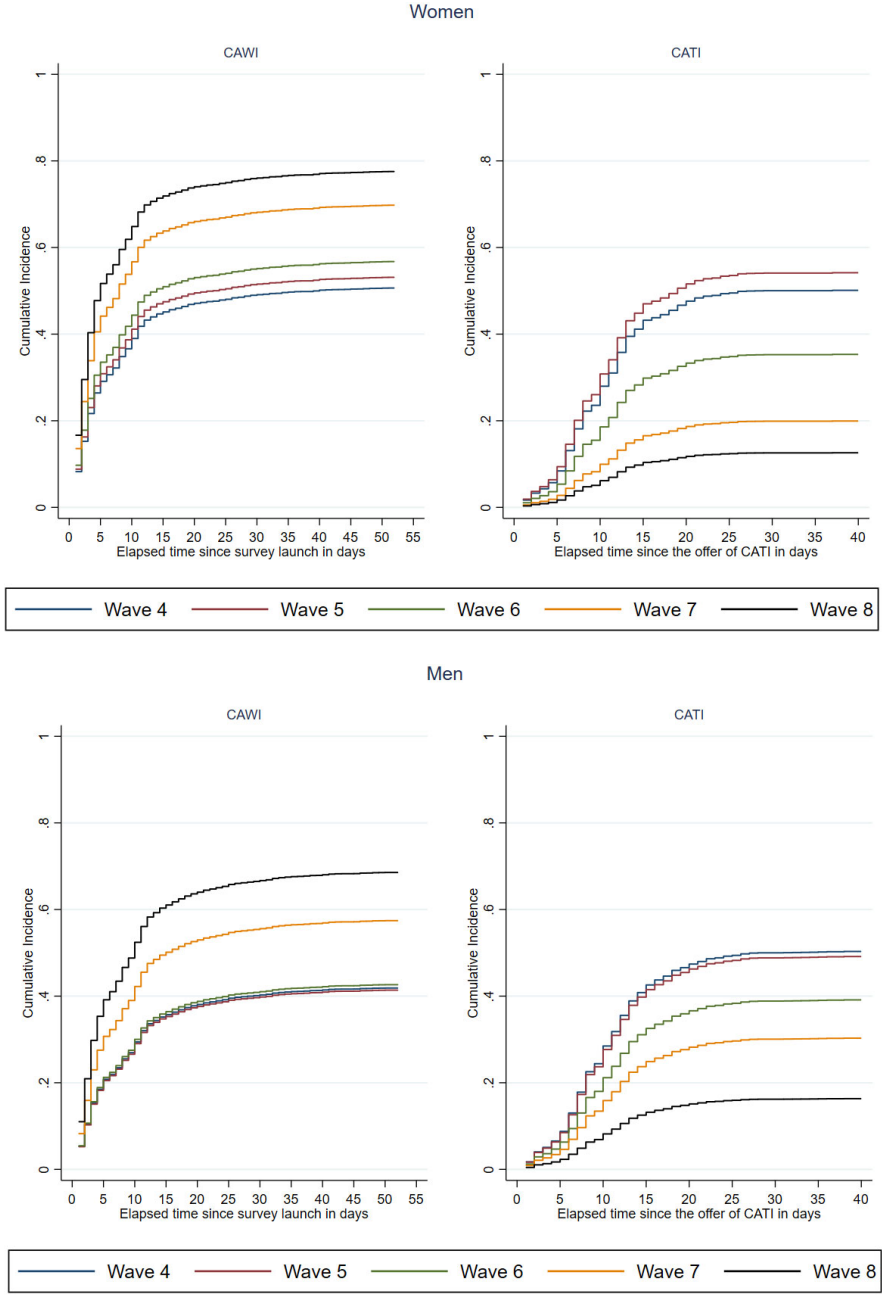

Figure 1 Gender disparities in survey participation across panel waves

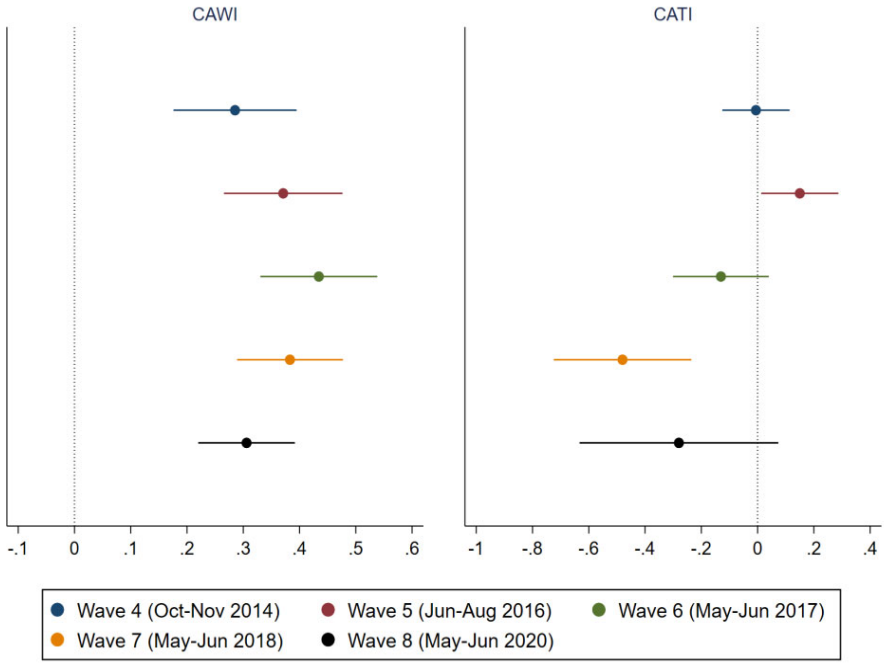

Figure 2 Gender disparities in survey participation across survey modes (estimated by competing risk model)

CATI mode, 51 per cent of the male panelists completed the online questionnaire and 40 per cent of them who did not respond in the initial survey mode took part in the CATI. The overall participation rate across the fieldwork period of 52 days was 82 per cent for women and 76 per cent for male panelists.

Finally, this gender disparity in survey participation was confirmed for each of the waves by multivariate analysis. For each of the waves, so-called β-coefficients for gender estimated by competing risk analyses are depicted in *Figure 2*.[4] In sum, this finding again provides evidence for gender disparities of participation different for the survey modes offered in a sequential mixed-mode design with a push-to-mode strategy. They are significant for each of the waves in the CAWI mode. For the CATI mode, it was observed that young women were more likely to participate at the survey, while there was a reverse gender disparity in Wave 7. Overall, this finding does not confirm *Hypothesis 3* proposing that gender disparities in survey participations diminish in a sequential mixed-mode design across the surveys and

4 The whiskers in the plot present the 95-% confidence interval of the coefficients. If they cross the vertical zero line, these effects are insignificant.

panel waves. Finally, it became obvious that an extension of the fieldwork period did not always result in decreasing gender differences in survey participation. In spite of three digital reminders in the CAWI mode and a sequence of reminders in the CATI mode after three call attempts, the participation rate declined completely to zero after four or five weeks.

Parametric analysis of gender-related participation rates

Utilizing a competing risk model, stark and statistically significant gender differences in survey participation were confirmed again for the initial online mode (*Table 2*). On average, by controlling for panel wave and regional opportunity structure, the inclination of female panelists to take part was () 53 per cent higher than men's (Model 1.1).

For the CATI mode, however, there were no significant gender differences in survey response (Model 1.2). Among the nonrespondents to whom the CATI mode was offered, there was no gender disparity in the timing and quantity of survey response for the entire number of panel waves. Furthermore, as described above, the differentiation of survey participation again made it obvious that participation in the CAWI increased across the panel waves, while the propensity toward response in the CATI mode decreased for recent panel waves. Finally, the effect of regional opportunity structure was only significant for the initial survey mode, where the response rates were lower in urban areas compared to the rural context. Living in urban areas resulted in a lower rate and speed of survey participation after the survey launch. This result does not confirm *Hypothesis 2*, since gender disparities remain constant.

Additionally, there was an impact of social origin on survey participation (Models 2.1-2.2). The selectivity of survey participation in terms of social origin was characterized by the fact that panelists from the middle and upper social classes had a greater inclination to complete the questionnaire than children of less skilled and unskilled blue-collar workers. Working class children were more likely to postpone their response and not take part in the CAWI or CATI than panelists from the other social classes. In contrast to *Hypothesis 1*, the gender differences in survey response were not explained by the social origin of panelists.

Panelists with a high educational level were more likely to take part in the online surveys than individuals enrolled in lower secondary schools with basic requirements. High language proficiency and ability in German language was correlated with starting early to complete the online questionnaire, while these achievements and skills were insignificant for participation in the CATI. By controlling for social origin, educational level and language, there were still gender disparities in survey participation in the initial survey mode. Therefore, *Hypothesis 1* is not in line with these findings.

Table 2 Gender and participation in different panel waves of the DAB panel study

Survey mode	CAWI	CATI	CAWI	CATI	CAWI	CATI
Models	1.1	1.2	2.1	2.2	3.1	3.2
Gender						
Female	0.356 (0.022)***	–0.046 (0.038)	0.273 (0.023)***	–0.052 (0.038)	0.290 (0.023)***	–0.038 (0.039)
Waves (Ref.: Wave 4)						
Wave 5	0.034 (0.039)	0.035 (0.046)	0.045 (0.040)	0.063 (0.046)	0.046 (0.040)	0.064 (0.046)
Wave 6	0.109 (0.039)**	–0.396 (0.053)***	0.124 (0.039)**	–0.359 (0.054)***	0.125 (0.039)**	–0.359 (0.054)***
Wave 7	0.502 (0.037)***	–0.848 (0.066)***	0.505 (0.037)***	–0.822 (0.067)***	0.507 (0.037)***	–0.822 (0.067)***
Wave 8	0.757 (0.036)***	–1.476 (0.092)***	0.772 (0.036)***	–1.442 (0.092)***	0.774 (0.036)***	–1.441 (0.092)***
Macro factor						
Regional opportunity structure	–0.037 (0.011)**	–0.016 (0.018)	–0.046 (0.012)***	–0.016 (0.019)	–0.045 (0.012)***	–0.014 (0.019)
Social origin (Ref.: Upper service class*)*						
Lower service class			0.012 (0.039)	–0.067 (0.067)	0.007 (0.039)	–0.068 (0.067)
Routine non-manual employees			–0.011 (0.037)	–0.029 (0.062)	–0.008 (0.037)	–0.034 (0.062)
Farmers, small proprietors			–0.066 (0.054)	–0.002 (0.092)	–0.061 (0.054)	–0.002 (0.092)
Foremen, skilled manual workers			–0.153 (0.042)***	–0.152 (0.067)*	–0.143 (0.042)***	–0.146 (0.067)*

Survey mode	CAWI	CATI	CAWI	CATI	CAWI	CATI
Models	1.1	1.2	2.1	2.2	3.1	3.2
Semi- and unskilled manual workers			–0.139 (0.057)*	–0.027 (0.094)	–0.137 (0.057)*	–0.027 (0.093)
Missing value			–0.140 (0.045)**	–0.224 (0.075)**	–0.138 (0.045)**	–0.226 (0.075)**
School type (Ref.: Basic requirements*)*						
Extended requirements			0.508 (0.032)***	0.146 (0.045)**	0.494 (0.032)***	0.132 (0.045)**
Advanced requirements			0.915 (0.038)***	0.240 (0.066)***	0.896 (0.038)***	0.217 (0.066)**
Missing value			0.382 (0.042)***	0.002 (0.067)	0.373 (0.042)***	–0.007 (0.067)
Language						
Language proficiency (GPA)			0.128 (0.013)***	–0.032 (0.021)	0.124 (0.013)***	–0.037 (0.021)
Language ability (German vs. others)			0.225 (0.034)***	0.060 (0.052)	0.228 (0.034)***	0.052 (0.052)
Personality traits						
Persistence			0.054 (0.013)***	–0.015 (0.020)	0.057 (0.013)***	–0.010 (0.020)
Control belief			0.035 (0.012)**	0.015 (0.020)	0.036 (0.012)**	0.017 (0.020)
Decisiveness			0.049 (0.012)***	0.046 (0.019)*	0.045 (0.012)***	0.042 (0.019)*
Gender role orientation						
Female role model					–0.086 (0.014)***	–0.045 (0.023)*

Survey mode	CAWI	CATI	CAWI	CATI	CAWI	CATI
Models	1.1	1.2	2.1	2.2	3.1	3.2
Male role model					0.054 (0.015)***	−0.013 (0.022)
Number of episodes	13,145	6,898	13,145	6,898	13,145	6,898
Number of events	7,460	2,744	7,460	2,744	7,460	2,744
Number of competing risks	2,923	1,392	2,923	1,392	2,923	1,392
Number of censored cases	2,762	2,762	2,762	2,762	2,762	2,762
Wald chi^2 (d.f.)	1,016.2 (6)	446.7 (6)	2178.6 (20)	510.6 (20)	2222.8 (22)	516.6 (22)

* $p<0.05$; ** $p<0.01$; *** $p<0.001$; β-coefficients, estimated by competing risk model (in brackets: robust standard error; clustered for individual units).

Source: DAB (own calculations)

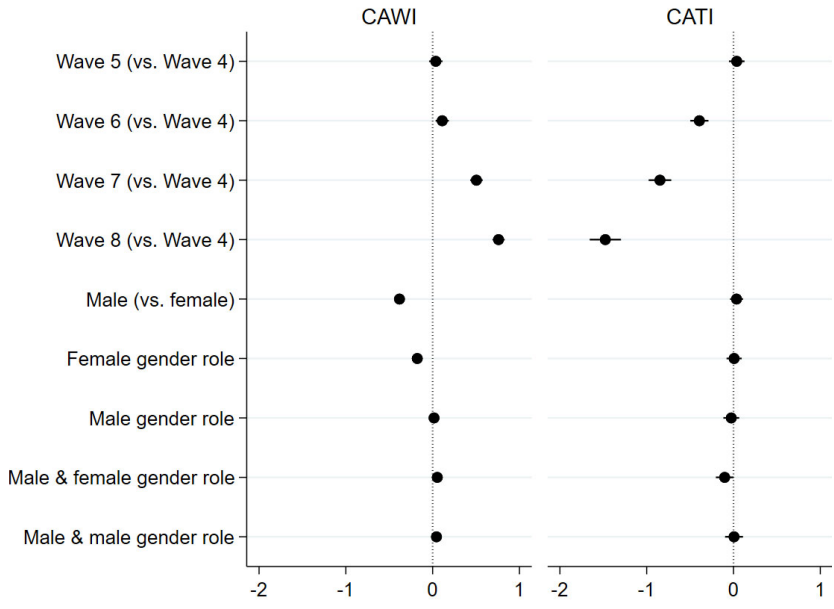

Figure 3 Impact of gender role orientation on survey participation (estimated by competing risk model)

The effects of personal traits were also straightforward. Panelists with high persistence, distinct primary or internal control belief and pronounced decisiveness were more likely to take part in the CAWI than individuals with external or secondary control belief, or individuals who were indifferent or characterized by remissness (Model 2.1). It was also found that panelists who postponed their response were more likely to take part in the CATI provided they had pronounced decisiveness (Model 2.2). However, since the gender difference was still significant, *Hypothesis 4* – stressing that personality traits explain the gender differences in survey participation – is not supported empirically.

While a panelist's orientation toward a traditional female role model made them less likely to respond to an invitation to the CAWI (Model 3.1) and to the CATI (Model 3.2), it was obvious that panelists who agreed with the "male breadwinner model" were more likely to be motivated in (early) survey participation (Model 3.1). However, if interaction effects of gender and gender role orientation are taken into account, by controlling for the same variables as in the models 3.1 and 3.2, there was no significant effect of them on survey response (*Figure 3*). This was true for the online mode (left-hand panel) as well as for the CATI mode (right-hand panel). Overall, these interaction effects on response were very small and did not dissolve the gender differences in survey participation at all. Therefore, *Hypothesis 5* is not

confirmed empirically, proposing that the effects of gender role orientation on survey response explain the gender differences in survey participation.

Finally, this issue was also true for the interactions effect of gender and each of the other covariates considered in model estimations, such as panel experience, social origin, educational level, language ability and proficiency, and regional opportunity structure. Each was insignificant; therefore, they are not reported or discussed in detail.

Discussion

In the dynamic analysis of the likelihood and timing of survey participation, the empirically evident gender differences could not be discounted by taking theoretically proposed processes and mechanisms into account, at least indirectly. Even if factors at the macro level of societal environment, at the meso level of survey characteristics and at the micro level of an interviewee's resources, abilities and beliefs were considered in the event history analysis, the gender effect on the timing and quantity of survey participation remained significant. None of the different hypotheses considering gender-based processes and mechanisms at each of the analytical levels was confirmed empirically. It seems that there are unobserved heterogeneities in gender-specific preferences and circumstances, and those perceptions of benefits and costs of responses in surveys of a multi-wave panel are not taken into account in a way that would support the assumptions of the theory of subjective expected utility and the heuristic logic of habitual action regarding scientific surveys.

Summary and Conclusions

The manifest aim of this empirical analysis has been to contribute to an evidence-based explanation of systematic gender disparities in survey participation. The latent aim is to relaunch this issue as a matter of interest in the research on survey methods. Regarding survey methodology, this research issue is still notoriously under-investigated in contemporary survey methodology (Becker, 2021: 20; Green, 1996). Thus, the question to be answered by this analysis was *why* we continuously observe differences between the genders regarding survey participation and, in particular, in its timing and quantity. Why are female target persons more likely to take part in social-scientific surveys than men?

Utilizing event history data on the likelihood of young panelists participating in surveys within a single-cohort and multi-wave panel study conducted in German-speaking cantons of Switzerland, the analysis has attempted to explain the gender differences in survey participation by hypotheses deduced from an advanced version of reasoned action theory and heuristic decision-making (Singer, 2011). According to Green (1996), it is assumed that, among other influences, the

gender difference is mainly based on gender-specific abilities, skills and achievements, which can be indicated by an invited potential respondent's language proficiencies and abilities, as well as by their educational success and attainment. Since girls and women have become advantaged in this respect due to educational expansion, it seemed plausible that a male target person's lower propensity toward survey participation might be correlated with their educational level and skills, resulting in gender disparities of participation. Even when social origin – providing a direct influence on an individual's educational achievement and attainment – was taken into account, the gender differences remained constant in each of the surveys. In the panel with a sequential mixed-mode design and a push-to-web-method, the gender differences were obvious for the initial online mode. However, even when other influences (such as personality traits, agreement with traditional gender roles or living in a rural or urban region) were taken into account, the gender disparities remained unsolved.

This result could be based on some limitations of this contribution. First of all, there is no elaborated theory explaining gender differences in survey participation. Ad-hoc arguments dominate a coherent explanation. Ideally, this theory should be a special case of a rational action approach. Second, the data provided less information on the mechanisms relevant for explaining response in general (e.g. benefit-cost calculation) that should be integrated into the statistical models. There was also a lack of information regarding different circumstances for genders that were essential for assessing the likelihood of survey participation. Third, the target population was limited to juveniles of a single birth cohort living in a small area in a small country. However, it could be argued that an explanation of gender-based survey response should be universal.

While none of the different hypotheses was confirmed empirically (and have not been confirmed in previous studies), and since the residual analysis conducted in the context of a multi-wave panel study on the educational and occupational trajectories of juveniles born around 1997 was not successful at all, the search for an empirically tested answer on the association between gender and survey participation must continue. Future studies may more profitably address the incremental effects of gender by directly measuring individual preferences, expectations and motivations, as well as perceived benefit-cost balance and everyday life. The social mechanisms emphasized in the wide variety of rational action theories and approaches to heuristic decision-making must also be observed directly with systematic reference to gender. As a by-product of such an endeavor, the different theories attempting to explain survey participation per se could be tested. Such a comparative test of theories on survey participation is overdue.

References

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

Austin, P.C., & Fine, J.P. (2017). Practical recommendations for reporting Fine-Gray Model Analyses for competing risk data. *Statistics in Medicine*, 36(27), 4391–4400.

Beck, M., Jäpel, F., & Becker, R. (2010). Determinanten des Bildungserfolgs von Migranten im Schweizer Bildungssystem. In G. Quenzel & K. Hurrelmann (Eds.), *Bildungsverlierer – Neue Ungleichheiten*. (pp. 313–337). Wiesbaden: VS.

Becker, R., Möser, S., & Glauser, D. (2019), Cash vs. vouchers vs. gifts in web surveys of a mature panel study – Main effects in a long-term incentives experiment across three panel waves. *Social Science Research*, 81, 221–234.

Becker, R., & Glauser, D. (2018). Are prepaid monetary incentives sufficient for reducing panel attrition and optimizing the response rate? An experiment in the context of a multi-wave panel with a sequential mixed-mode design. *Bulletin of Sociological Methodology*, 137, 74–95.

Becker, R. (2021). The Effects of a Special Sequential Mixed-Mode Design, and Reminders, on Panellists' Participation in a Probability-Based Panel Study. *Quality and Quantity*, 55, 1–26.

Becker, R., & Mehlkop, G. (2011). Effects of prepaid monetary incentives on mail survey response rates and on self-reporting about delinquency – Empirical findings. *Bulletin of Sociological Methodology*, 109, 5–25.

Becker, R., & Müller, W. (2011). Bildungsungleichheiten nach Geschlecht und Herkunft im Wandel. In A. Hadjar (Ed.), *Geschlechtsspezifische Bildungsungleichheiten* (pp. 55–75). Wiesbaden: VS Verlag für Sozialwissenschaften.

Glauser, D., & Becker, R. (2016). VET or general education? Effects of regional opportunity structures on educational attainment in German-speaking Switzerland. *Educational Research in Vocational Education and Training*, 8, 1-25

Becker, R., & Mayer, K. U. (2019). Societal Change and Educational Trajectories of Women and Men Born between 1919 and 1986 in (West) Germany. *European Sociological Review*, 35(2), 147–168.

Becker, R., Glauser, D., & Möser, S. (2020). Determinants of educational choice and vocational training opportunities in Switzerland – Empirical analyses with longitudinal data from the DAB panel study. In N. McElvany, H. G. Holtappels, F. Lauermann, A. Edele, & A. Ohle-Peters (Eds.), *Against the Odds – (In)Equity in Education and Educational Systems* (pp. 125–143). Münster: Waxmann.

Bethlehem, J. (2009). *Applied Survey Methods*. Hoboken, N.J.: Wiley.

Beullens, K., Loosveldt, G. Vandenplas, C., & Stoop, I. (2018). Response rates in the European Social Survey: increasing, decreasing, or a matter of fieldwork efforts? *Survey Methods: Insights from the Field*. Retrieved 4 July 2020, from the Survey Methods website: https://surveyinsights.org/?p=9673.

BFS (Bundesamt für Statistik – Federal Office for Statistics) (2021). Internetnutzung. Retrieved on May 5, 2021, from the website of the Swiss Federal Statistical Office: https://www.bfs.admin.ch/bfs/de/home/statistiken/kultur-medien-informationsgesellschaft-sport/informationsgesellschaft/gesamtindikatoren/haushalte-bevoelkerung/internetnutzung.html.

Blossfeld, H.-P-, Rohwer, R., & Schneider, T. (2019). *Event History Analysis with Stata*. London, New York: Routledge.

Brick, M.J., & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *AAPSS (The ANNALS of the American Academy of Political and Social Science)*, 645(1), 36–59.

Brüggen, E., Wetzels, M., de Ruyter, K., & Schillewaert, N. (2011). Individual differences in motivation to participate in online panels. The effect on response rate and response quality perception. *International Journal of Market Research*, 53(3), 369–390.

Buchmann, C., DiPrete, T.A., & McDaniel, A. (2008). Gender inequalities in education. *Annual Review of Sociology*, 34, 319–337.

Busby D. M., & Yoshida, K. (2013). Challenges with online research for couples and families: evaluating nonrespondents and the differential impact of incentives. *Journal of Child and Family Studies*, 24(2), 505–513.

Couper, M.P., & Groves, R.M. (1996). Social environmental impacts on survey cooperation. *Quality & Quantity*, 30(May), 173–186.

Couper, M.P., Kapteyn, A., Schonlau, M., & Winter, J. (2011). Noncoverage and nonresponse in an Internet survey. *Social Science Research*, 36(1), 131–148.

Couper, M.P., Singer, E., Conrad, F.G., & Groves, R.M. (2008). Risk of disclosure, perceptions of risk, and concerns about privacy and confidentiality as factors in survey participation. *Journal of Official Statistics*, 24(2), 255–275.

Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the Index of Consumer Sentiment. *Public Opinion Quarterly*, 64(4), 413–428.

Dalecki, M.G., Ilvento, T.W., & Moore, D.E. (1988). The effects of multi-wave mailings on the external validity of mail surveys. *Journal of the Community Development Society*, 19(1), 51–70.

de Heer, W. (1999). International response trends: results of an international survey. *Journal of Official Statistics*, 15(2), 129–142.

Dillman, D.A. (2000). *Mail and Internet Surveys. The Tailored Design Method*. New York: Wiley.

DiPrete, T.A., & Buchmann, C. (2013). *The Rise of Women. The Growing Gender Gap in Education and What it Means For American Schools*. New York: Russell Sage Foundation.

Durrant, G.B., D'Arrigo, J., & Steele, F. (2013). Analysing interviewer call record data by using a multilevel time event history modelling approach. *Journal of the Royal Statistical Society*, 176(1), 251–269.

Dutwin, D., & Lavrakas, P. (2016). Trends in telephone outcomes, 2008–2015. *Survey Practice*, 9(3), 1–9.

Dykema J., Stevenson, J., Klein, L., Kim, Y., & Day, B. (2012). Effects of e-mailed versus mailed invitations and incentives on response rates, data quality, and costs in a web survey of university faculty. *Social Sciences Computer Review*, 31(3), 359–370.

Erikson, R., & Goldthorpe, J.H. (1992). *The Constant Flux. A Study of Class Mobility in Industrial Societies*. Oxford: Clarendon Press.

Fine, J.P., & Gray, R.J. (1999). A Proportional Hazards Model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446), 496–509.

Glauser, D. (2015). *Berufsausbildung oder Allgemeinbildung*. Wiesbaden: VS Springer.

Göritz, A. S. (2014). Determinants of the starting rate and the completion rate in online panel studies. In M. Callegaro, R. Baker, J. Bethlehem, A.S. Göritz, J.A. Krosnick, & P.J.

Lavrakas (eds.), *Online Panel Research: A Data Quality Perspective*. (pp. 154–170). New York: John Wiley & Sons.

Göritz, A.S., & Stieger, S. (2009). The impact of the field time on response, retention, and response completeness in list-based web surveys. *International Journal of Human-Computer Studies*, 67(4), 342–348.

Goyder, J., & Leiper, J.M. (1985). The decline in survey response: a social values interpretation. *Sociology*, 19(1), 55–71.

Goyder, J., Boyer, L., & Martinelli, G. (2008). Integrating exchange and heuristic theories of survey nonresponse. *Bulletin de Méthodologie Sociologique*, 92(1), 1–14.

Green, K.E. (1996). Sociodemographic factors and mail survey response. *Psychology & Marketing*, 13(2), 171–184.

Green, K.E. (2014). Reluctant respondents. Differences between early, late and nonresponders to a mail survey. *Journal of Experimental Education*, 59(3), 268–276.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5), 646–675.

Groves, R.M., & Couper, M.P. (1998). *Nonresponse in Household Interview Surveys.* New York: John Wiley & Sons.

Groves, R.M., & Peytcheva, E. (2008). The Impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, 72(2), 167–189

Groves, R.M., Cialdini, R.B., & Couper, M.P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, 56(4), 475–495.

Groves, R.M., Dillman, D.A., Eltinge, J.L., & Little, R.J.A. (2001). *Survey Nonresponse.* New York: Wiley & Sons.

Groves, R.M., Singer, E., & Corning, A. (2000). Leverage–saliency theory of survey participation. Description and an illustration. *Public Opinion Quarterly*, 64(3), 299–308.

Grunow, D. (2013). Gender-based division of labour and socialisation as a matter of biography. *Zeitschrift für Soziologie der Sozialisation und Erziehung*, 33(4), 384–398.

Hadjar, A. (2011). Einleitung. In A. Hadjar (Ed.), *Geschlechtsspezifische Bildungsungleichheiten* (pp. 7–20). Wiesbaden: Springer VS.

Kalter, F., Granato, N., & Kristen, C. (2007). Disentangling recent trends of the second generation's structural assimilation in Germany. In S. Scherer, R. Pollak, G. Otte, & M. Gangl (eds.), *From Origin to Destination. Trends and Mechanisms in Social Stratification Research* (pp. 214–245). Frankfurt am Main: Campus.

Keusch, F. (2015). Why do people participate in web surveys? Applying survey participation theory to Internet survey data collection. *Management Review Quarterly*, 65(January), 183–216.

Kleiner, B., Lipps, O., & Ferrez, E. (2105). Language ability and motivation among foreigners in survey responding. *Journal of Survey Statistics and Methodology*, 3(3), 339–360.

Kreuter, F. (2013). Improving Surveys with Paradata: Introduction. In F. Kreuter (Ed.), *Improving Surveys with Paradata* (pp. 1–9). Hoboken, NJ: Wiley.

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), 257–273.

Laguilles, J. S., Williams, E.A., & Saunders, D.B. (2011). Can lottery incentives boost web survey response rates? Findings from four experiments. *Research in Higher Education*, 52(5), 537–533.

Lambert, P.C. (2017). The estimation and modeling of cause-specific cumulative incidence functions using time-dependent weights. *Stata Journal*, 17(1), 181–207.

Leeper, T.J. (2019). Where have all the respondents gone? Perhaps we ate them all. *Public Opinion Quarterly*, 83(S1), 280–288.

Marcus B., & Schütz, A. (2005). Who are the people reluctant to participate in research? Personality correlates of four different types of nonresponse as inferred from self- and observer ratings. *Journal of Personal*, 73(4), 959–984.

Noordzij, M., Leffondré, K., von Stralen, K.J., Zocali, C., Dekker, F.W., & Jager, K.J. (2013). When do we need competing risks methods for survival analysis in nephrology? *Nephrology Dialysis Transplantation*, 28(11), 2670–2677.

Patrick, M.E., Singer, E., Boyd, C.J., Cranford, J.A., & McCabe, S.E. (2013). Incentives for college student participation in web-based substance use survey. *Addictive Behavior*, 38(3), 1710–1714.

Porst, R., & Briel, C. v. (1995). Wären Sie vielleicht bereit, sich gegebenenfalls noch einmal befragen zu lassen? Oder: Gründe für die Teilnahme an Panelbefragungen. (ZUMA-Arbeitsbericht, 1995/04). Mannheim: Zentrum für Umfragen, Methoden und Analysen.

Porter, S.R., & Whitcomb, M.E. (2005). Non-response in student surveys: the role of demographics, engagement and personality. *Research in Higher Education*, 46(2), 127–152.

Quenzel, G., & Hurrelmann, K. (2010), Geschlecht und Schulerfolg: Ein soziales Stratifikationsmuster kehrt sich um. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 62(1), 61–91.

Saßenroth, D. (2013). *The Impact of Personality on Participation Decisions in Surveys. A Contribution to the Discussion on Unit Nonresponse*. Springer VS, Wiesbaden.

Schoeni, R.F., Stafford, F., McGonagle, K.A., & Andreski, P. (2013). Response rates in national panel surveys. *AAPSS (The ANNALS of the American Academy of Political and Social Science)*, 645(1), 60–87.

Schuster, N.A., Hoogendijka, E.O., Koka, A.A.L., Twiska, J.W.R., & Heymansa, M.W. (2020). Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risk analysis. *Journal of Clinical Epidemiology*, 122(1), 42–48.

Singer, E. (2006). Introduction: Nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5), 637–645.

Singer, E. (2011). Toward a benefit–cost theory of survey participation: evidence, further tests, and implication. *Journal of Official Statistics*, 27(2), 379–392.

Singer, E., van Hoewyk, J., & Maher, M.P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64(2), 171–188.

Sixt, M. (2013). Wohnort, Region und Bildungserfolg. Die strukturelle Dimension bei der Erklärung von regionaler Bildungsungleichheit. In R. Becker & A. Schulze (eds.), *Bildungskontexte*. (pp. 457–481). Wiesbaden: Springer VS.

Slauson-Blevins, K., & Johnson, K.M. (2016). Doing gender, doing survey? Women's gatekeeping and men's non-participation in multi-actor reproductive surveys. *Sociological Inquiry*, 86(3), 427–449.

Smith, T. (1995). Trends in non-response rates. *International Journal of Public Opinion Research*, 7(2), 157–171.

Steeh, C.G. (1981). Trends in nonresponse rates, 1952–1979. *Public Opinion Quarterly*, 45(1), 40–57.

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving Survey Response: Lessons learned from the European Social Survey*. New York: John Wiley & Sons.

Tourangeau, R., & Plewes, T.J. (2013). *Nonresponse in Social Science Surveys*. Washington, D.C.: National Academics Press.

Wenz, A., Al Baghal, T., & Gaia, A. (2021). Language proficiency among respondents: implication for data quality in a longitudinal face-to-face survey. *Journal of Survey Statistics and Methodology*, 9(1), 75–93.

Appendix

Table A–1: Varimax-rotated three factor structure of personality traits items

Items (value range: 1 = disagree – 5 = agree)	Factor 1 Persistence	Factor 2 Control belief	Factor 3 Decisiveness
Persistence			
I hate to leave something unfinished.	0.5882	0.0287	–0.0029
When I have made up my mind, I manage to keep it up.	0.7192	–0.0581	–0.0455
What I‘ve started I‘ll finish.	0.7353	–0.0640	0.0038
Even when I encounter difficulties at work, I persist in it.	0.7373	–0.0528	–0.0649
Even with a tedious task, I don‘t give up until I‘m done.	0.7264	–0.0167	–0.0152
Control belief			
I am happy to take responsibility.	0.5238	–0.1482	0.2326
It has proven to be good for me to make decisions on my own instead of turning to fate.	0.5408	–0.1443	0.1740
When there are problems and resistance, I usually find ways and means to assert myself.	0.5807	–0.1498	0.1827
Success depends on luck, not performance.	0.0124	0.0878	0.8260
I feel like I have little control over what happens to me.	–0.0059	0.2084	0.7648
Decisiveness			
I am very unsure of how to decide and often fluctuate back and forth.	–0.0487	0.8017	0.0809
I let other people confuse me in my decision.	–0.0871	0.7768	0.1132
After a decision, I have great doubts as to whether I have really made the right decision.	–0.0607	0.8109	0.0931
There are so many options that I have a hard time deciding which one to choose.	–0.0161	0.7772	0.0549
When I have made a decision, I hold on to it.	0.4613	–0.0437	–0.0111

	N	Minimum	Maximum
Persistence	3,680	–5.0302	2.4407
Control belief	3,680	–3.0764	3.1386
Decisiveness	3,680	–4.4313	2.5783

Table A–2: Varimax-rotated one factor structure of gender role items

Items (value range: 1 = disagree – 5 = agree)	Mean	SD	Minimum	Maximum	Factor Gender role
Female gender role: I think it's important for a woman…					
to be employed.					0.6173
to earn much money.					0.7766
to have a successful career.					0.7675
to have children.					0.5039
to take care of the household.					0.4947
to be responsible for childcare.					0.5105
Female gender role orientation	–3.54e-09	0.9655	–3.8443	2.5000	
Male gender role: I think it's important for a man…					
to be employed.					0.7244
to earn much money.					0.8087
to have a successful career.					0.8004
to have children.					0.5809
to take care of the household.					0.4666
to be responsible for childcare.					0.5724
Male gender role orientation	–2.51e-09	0.9655	–4.6210	1.9810	

Table A-3: Descriptive statistics (all respondents across five waves)

	N	%	Mean	SD	Minimum	Maximum
Gender	13,145	51.0			0	1
Waves	13,145					
Wave 4	2,654	20.2			0	1
Wave 5	2,799	21.3			0	1
Wave 6	2,712	20.6			0	1
Wave 7	2,488	18.9			0	1
Wave 8	2,492	19.0			0	1
Regional opportunity structure	13,145		0.2218	0.9804	–1.6488	3.6225
Social origin (EGP)	13,145					
I	1,863	14.2			0	1
II	2,453	18.7			0	1
IIIa/b	3,173	24.1			0	1
IVa/b/c	805	6.1			0	1
V/VI	2,132	15.2			0	1
VIIa/b	704	5.4			0	1
Missing values	2,024	15.4			0	1
School type	13,145					
Basic requirements	3,383	25.7			0	1
Extended requirements	5,467	41.6			0	1
Advanced requirements	2,068	15.7			0	1
Missing values	2,227	16.9			0	1
Language proficiency (z-standardized GPA)	13,145		–0.0992	0.9089	–3.3773	1.3327
Language ability (German vs. other languages)	13,145	85.5			0	1
Persistence	13,145		0.0182	0.9192	–5.0302	2.4407
Control belief	13,145		–0.0235	0.9449	–3.0764	3.1386
Decisiveness	13,145		0.0361	0.9264	–4.4313	2.5783
Female gender role	13,145		–3.54e-09	0.9656	–3.8443	2.5000
Male gender role	13,145		–2.51e-09	0.9656	–4.6210	1.9810

methods, data, analyses | Vol. 16(1), 2022, pp. 33-50 DOI: 10.12758/mda.2021.09

Do Web and Telephone Produce the Same Number of Changes and Events in a Panel Survey?

Oliver Lipps[1,2] *& Marieke Voorpostel*[1]

[1] *FORS, Switzerland*

[2] *University of Bern*

Abstract

Measuring change over time is one of the main purposes of longitudinal surveys. With an increase in the use of web as a mode of data collection it is important to assess whether the web mode differs from other modes with respect to the number of changes and events that are captured. We examine whether telephone and web data collection modes are comparable with respect to measuring changes over time or experiencing events. Using experimental data from a two-wave pilot of the Swiss Household Panel, we investigate this question for several variables in the domain of work and family.

We find differences for the work-related variables, with web respondents more likely to report changes. These differences do not disappear once the socio-demographic composition of the sample is taken into consideration. This suggests that these differences are not driven by observed different characteristics of the respondents who may have self-selected into one or the other mode. Contrary to work-related variables, a termination of a relationship was more common in the telephone group. This shows that one mode does not necessarily measure more change or events than another; it may depend on the variable in question. In addition, the difference in the protocol mattered: a web respondent in a household that participated fully by web sometimes differed from a web respondent in a household that had a household interview by phone. Nonetheless, the telephone group differed more from the various web protocols that the web protocols among themselves.

With more household panel surveys introducing web questionnaires in combination with more traditional face-to-face and telephone interviews, this study adds to our understanding of the potential consequences of mixing modes with respect to longitudinal data analysis.

Keywords: reporting change, reporting events, mode effects, household panel, mixed-mode, measurement, selection

One of the main purposes of longitudinal surveys is to measure change over time. For example, many studies in the social sciences focus on changes in circumstances and the occurrence of events in people's lives and assess their consequences for the individuals experiencing them. By following people over time, it becomes possible to analyze how a wide variety of changes such as for example a change in employment situation or civil status affects people's lives in a multitude of ways (see Chandola & Zhang, 2018; Choi, Chung & Breen, 2020; Leopold, 2018; Rözer et al., 2020 for recent examples).

To reduce costs while keeping response rates and representativeness on an acceptable level, an increasing number of longitudinal studies rely on web as one of the modes of data collection (Voorpostel, Lipps & Roberts, 2021). This is also the case for long-running household panel studies: traditionally often relying on face-to-face (e.g., the UK Understanding Society (UKHLS), the Household, Income and Labour Dynamics in Australia (HILDA) Survey) or telephone interviews (e.g., the US Panel Study of Income Dynamics (PSID), the Swiss Household Panel (SHP)) as their main mode of interview. While most switch already participating households to web at a later wave (e.g., UKHLS), some use web already from the first wave of interview, as is the case for the latest refreshment sample of the SHP.

With an increasing role of the web mode in longitudinal studies it is important to understand to what extent longitudinal data collected include a comparable number of events in different modes. If data collected with one mode produces fewer events and changes over time than data collected with another mode, this affects the analytical potential of such data and should be taken into consideration when deciding upon a design for a longitudinal study.

Whereas there is increasing research attention to measurement differences by mode of specific target variables, both in cross-sectional and longitudinal surveys, we know very little about the extent to which modes vary in how they capture changes over time in longitudinal surveys. As the measurement of intra-individual change is the main purpose of longitudinal surveys, it is of great importance to assess the relationship between survey mode and the measurement of change over time.

Comparing telephone to web, we argue that the same factors that drive mode differences in measurement of target variables may also drive differences in the measurement of change and event occurrence over time. Web and telephone are two modes that differ in important ways. With respect to survey *participation*, web and telephone differ in coverage, reachability of respondents, and their willingness to participate (De Leeuw, 2018; Nagelhout et al., 2010). As certain transitions and

Direct correspondence to
Oliver Lipps, FORS, Switzerland and University of Bern
E-mail: oliver.lipps@fors.unil.ch

events tend to be more common in specific subgroups of the population, a different sample composition may produce different reporting of change over time.

With respect to *measurement*, an important difference between telephone and web is the presence of an interviewer in telephone interviews. Interviewers affect different aspects of the survey data collection process (Brady & Blom, 2017). Interviewers on the one hand increase data quality as they can guide the respondent through complicated questions and burdensome parts of the questionnaire, motivate respondents to complete the task and may check whether (intended or unintended) reported or not reported changes are plausible. Reduced effort by web respondents is evidenced by the fact that item nonresponse tends to be higher in web surveys (Groves et al., 2011) although findings regarding satisficing behavior in web surveys is mixed (Bowyer & Rogowski, 2017; Fricker et al., 2005; Chang & Krosnick, 2010). On the other hand, the presence of an interviewer tends to increase socially desirable responding (Chang & Krosnick, 2010). The mode of interview also affects responses through variation in other characteristics, such as the pace of the interview, presentation (visual or auditive), and the timing of the interview (Christian, Dillman & Smyth, 2008). These differences in reporting may lead to different rates of change and event occurrence measured in telephone and web surveys.

We formulate the following two research questions: (1) Do telephone and web respondents differ in the likelihood of reporting status changes and events in the work and family domain? And (2) Does any difference persist after controlling for differential sample composition by mode? As this is a first exploration of this topic, we refrain from formulating hypotheses on the specific events. Rather, we assess whether the mode in which a survey is administered is associated with the frequency with which respondents report specific changes in circumstances and event occurrence, and if so, in a second step, whether these differences remain after controlling for known differences between the modes in sample composition. If differences by mode remain, this gives some indication of different response behavior by mode. Although this remains speculative as there is no population data on such changes and it will not be possible to validate reported changes, it does suggest that the mode of interview has consequences for longitudinal analyses of the studied changes and events that go beyond sample composition with respect to socio-demographic characteristics.

We examine several common changes and events in the work and family domain and include events and changes that have received research attention. More precisely, we include the following events and changes: change in employment situation (employed, unemployed or inactive), change in jobs, experience of unemployment, change in partnership status, civil status or household size, termination of a relationship, death of a close person, and residential moves.

Data and Method

Data

Design of the Swiss Household Panel mode experiment

For this study we use data from a two-wave pilot for the Swiss Household Panel (SHP) comparing telephone to web. The SHP is a longitudinal household study that follows randomly sampled households in Switzerland over time since 1999. The SHP interviews all household members on an annual basis, predominantly by telephone (Tillmann et al., 2016). In preparation of the third refreshment sample which was launched in 2020, a mode experiment conducted in 2017-2018 compared the standard telephone-based recruitment and fieldwork strategy with two web alternatives.

In the SHP, each household assigns a household reference person (HRP), who completes the household grid and the household questionnaire (household level) in each wave. Based on the household grid, the HRP and all household members of at least 14 years old complete an individual questionnaire (individual level). The standard SHP protocol involves telephone interviews on the household level, and with all household members to complete an individual questionnaire, also by telephone. In the mode experiment this group was referred to as the *telephone group.* The first web alternative tested was a mixed-mode protocol combining a telephone interview with the HRP on the household level, with the HRP and household members completing their individual questionnaires via web (*mixed-mode group*). The second web alternative tested was a web-only protocol using web for the grid, the household, and all individual questionnaires (*web-only group*) (see Voorpostel et al., 2020).

The sample for the study was a simple random sample of individuals which was stratified by region, drawn from a sampling frame based on population registers maintained by the Swiss Federal Statistical Office. The households of the sampled individuals were randomly assigned to one of the three experimental groups. The sampled individual was approached first as a HRP.[1]

The sampling frame included landline telephone numbers for 60 percent of the sampled individuals. In both the telephone group and the mixed mode group, face-to-face and web were offered as alternatives if no telephone number was available and to initial refusals. HRPs in the *web-only group* (3) received a login code with their invitation letters and completed all questionnaires by web. Household

1 An exception was made for the web group: if the sampled person was a young adult child presumably living with their parents (deduced from auxiliary frame data), a parent was selected at random to act as the HRP instead. In both waves, in all three treatment groups household members were free to select an alternative HRP than the one initially approached.

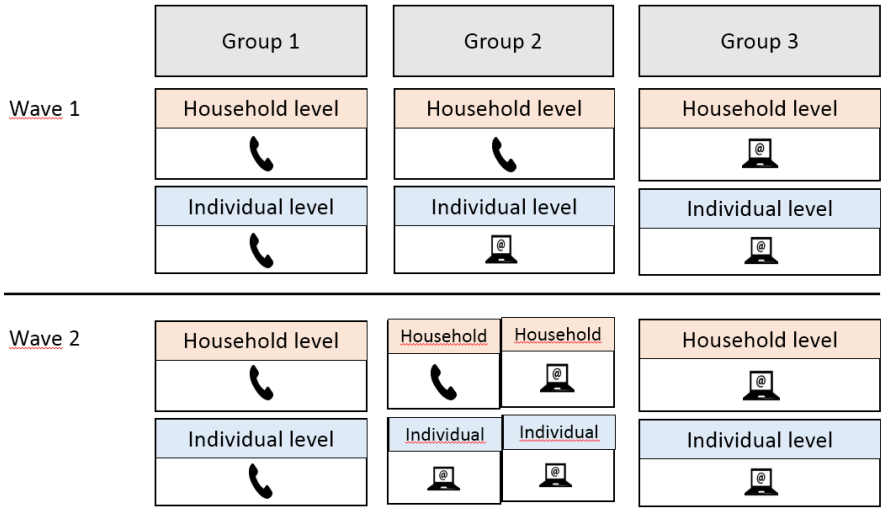

Figure 1 Illustration of the research design for the two-wave pilot study of the SHP_IV (adapted from figure 1 in Voorpostel et al., 2020)

members in the *mixed mode* (2) and the *web-only group* (3) received login codes for their individual questionnaires after the HRP had provided information on the household composition. Upon request, respondents could be interviewed by telephone. In both experimental groups, two reminders were sent two weeks apart to decrease nonresponse to the web questionnaire. If a telephone number was available, the second reminder was replaced by a telephone contact.

Wave 2 followed the same protocols, but with 30 percent of the mixed-mode group switched to the protocol of the web-only group (*mixed-mode-to-web group*). This means that while 30% of the mixed-mode group switched on the household level from the telephone to the web, the remaining 70% kept the telephone on the household level. Due to splitting the sample at wave 2, the mixed-mode group started out with a larger sample size (2192 households) at wave 1 than the telephone group (790 households). As response rates tend to be lower in web surveys, the web group was also larger than the telephone group (1213 households). Figure 1 illustrates the research design.

Response rates in the first wave on the household level varied between 47% for the web group and 53% for the telephone group (the mixed-mode group obtained 52%). Of all household members included in the grid of participating households 69% (n=707) participated in wave 1 in the telephone group, 67% (n=1798) in the mixed-mode group, and 62% (n=879) in the web group. All households that completed at least the grid in the first wave and that had not left the study were reapproached at wave 2. Wave 2 also included 42 newly formed households from

split households. Response rates on the household level in wave 2 were 77% for the telephone group (332 households) and the mixed-mode group (621 households), 74% for the mixed-mode-to-web group (263 households), and 76% (459 households) for the web group. Individual level participation in wave 2 was 73% (n=570) in the telephone group, 72% (n=1006) in the mixed-mode group, 75% (n=460) in the mixed-mode-to-web group and 76% (n=807) in the web group.

Analytical sample

As we analyse changes on the household level and on the individual level, we define analytical samples for households and individuals. We include only households and individual respondents who answered in the assigned mode. These comprise in the first wave 328 households including 603 individuals who participated by telephone in the telephone group (excluding 44 households (comprising 65 household members) who participated by face-to-face and 39 web respondents), 800 households (by telephone; excluding 79 households who participated by face-to-face) and 1579 individuals (by web; excluding 24 face-to-face respondents and 195 telephone respondents) in the mixed-mode group, and 349 households including 792 individuals who participated by web in the web group (excluding 74 households who participated by telephone with 87 household members who participated by telephone). In the second wave, these figures amount to 274 households (460 individuals) in the telephone group, 482 households (776 individuals) in the mixed-mode group, 211 households (431 individuals) in the mixed-mode-to-web group, and 342 households (713 individuals) in the web group. We imputed all independent variables used in the regression analyses using chained equations implemented in the iterative chain equations (ice) procedure in Stata (Royston and White 2011) and disregard cases with missing values for the dependent variables. All analyses are done using Stata 16 SE.

Measures

Dependent variables: changes and events

We examine nine dependent variables in the domains of work and family. All dependent variables refer to changes or event occurrence but were measured in different ways and sometimes refer to different time points. Most variables are based on questions asked to the respondent and some are constructed variables from multiple questions and may include information provided by the household reference person (which is then verified by the respondent). Changes and events were measured either directly by asking the respondent about them or indirectly by comparing the response provided in both waves. Except for change in the number of household members and whether the household moved, all variables are measured on

the individual level. For a change in jobs only employed respondents were included and for the experience of unemployment only respondents who were employed for at least one month in the year prior to the interview entered the analyses. Table 1 provides an overview of the dependent variables and gives details of how they are measured. As a result of these differences, the number of observations included in the analyses varies. All dependent variables in the regression analyses are dichotomous, indicating whether an event or change occurred.

Independent and control variables

The main independent variable is the survey mode. We include the survey mode by distinguishing between the different experimental groups (telephone group (reference), mixed-mode group, mixed-mode-to-web group, web group). The mixed-mode group used telephone on the household level and web on the individual level. The mixed-mode-to-web group only refers to wave 2 (hence it equals 0 for all households in wave 1 and for households not part of the mixed-mode-to-web group in wave 2). This group completed all questionnaires by web and included those households that were moved from the mixed-mode group to the web group in the second wave. For the models estimating the dependent variables that were measured at both waves we pool observations from both waves and include a dummy variable indicating whether the observation came from wave 1 or wave 2 (1=wave 1, 2=wave 2). For the dependent variables on the individual level, we include whether the respondent was the HRP or another household member (1=HRP, 0=other respondents).

The regression models further control for the following socio-demographic variables associated with survey participation and panel attrition (Roberts & Vandenplas, 2017; Voorpostel et al., 2020). First, whether the household has a registered landline (information from the registry data, 1=yes, 0=no). The remaining control variables were measured in the survey, but consistency with information from the registry was very high (Voorpostel et al., 2020), indicating that there was hardly any measurement error in these variables: gender (1=male, 0=female), age in categories (14-30, 31-49, 50-60, 61-92), first nationality (Swiss, neighboring country, other country), education (1=tertiary level, 0=lower than tertiary level). Descriptive statistics for all dependent and independent variables are included in the Appendix.

Table 1 Overview of dependent variables

	Question formulation/other details	Person asked	Measurement of change/ event occurrence
Work domain			
Change in employment status	Constructed from detailed variables about working, distinguishing between working, unemployed and inactive	Individual	Change between waves
Change in jobs[a]	"Since (month, year), have you changed jobs or employers?"	Individual	Reported in both waves
Experience of at least one episode of unemployment[b]	"We are going to review the months since (month, year) and for each month you should tell me whether your main activity was: full-time employee, part-time employee, full-time self-employed, part-time self-employed, unemployed, retired, training/education, housework, or any other situation?"	Individual	Reported in both waves
Family/household domain *Individual level*			
Change in partnership status	"Do you have a partner?" (yes, living together/ yes, not living together/ no	Individual	Change between waves
Change in civil status since last wave	Civil status of household members is provided by HRP in grid questionnaire, respondent confirms in individual questionnaire	HRP Individual	Change between waves
Termination of relationship	"Since (month-year), has a close and important relationship ended?" (yes/no)	Individual	Reported in both waves
Death of close other	"Since (month-year), has a person closely related to you died?" (yes/no)	Individual	Reported in both waves

	Question formulation/other details	Person asked	Measurement of change/ event occurrence
Household level			
Change in household size[c]	Constructed from information on household composition	HRP	Change between waves
Residential move[d]	"Did you move to another accommodation since (date of the last interview)?"	HRP	Reported in second wave

[a]) Only employed respondents answered these questions
[b]) Only respondents who experienced a change in employment situation answered these questions. We included respondents who worked at least one month.
[c]) In case a household splits, we define this change to be missing for the newly established household.
[d]) In case a household splits and both new households remain in the study, the HRP of both households are asked whether they moved or not.

Results

We first explore bivariate differences by mode in the reporting of changes and events. Table 2 presents the distribution of the dependent variables by mode. The table shows the percentage of respondents who reported the change or event, who reported no change or event, and who had a missing value on the item, meaning they replied with "don't know" or "no answer" (item nonresponse, INR). A clear pattern that emerges for all variables, and that is in line with previous studies, is that the respondents who replied by web had a higher percentage of INR.

When we disregard the INR and only include substantive responses, we find significant differences only in the work domain, where web respondents were more likely to report a change in jobs or a change in employment status. For none of the other events and changes we find significant differences between telephone and web.

Tables 3.1 to 3.3 present the results of the regression models using linear probability regression models predicting probability of experiencing the event or the change. We control for experimental group (which take the complete experimental design into account), wave and whether the respondent is the HRP (base model) and add in a second step all independent variables to assess whether significant mode effects change upon controlling for selection. The experimental group determines the mode on each level (household or individual) such that including mode is not necessary.

In the domain of employment (change in jobs, experience of unemployment, change in employment status), we find significant effects for the experimental groups for all three dependent variables in the multivariate models, although effect sizes are modest. These significant effects remain unaltered after controlling for the composition of the sample. The distinction between the experimental groups reveals that the differences are not only related to the mode (as analyzed in Table 2) but vary by the combination of modes used on the household and individual level in different ways. Table 2 shows that a change in employment status is more often reported in the web group than in the telephone group (with a significantly higher probability of .04). Yet, while comparable in magnitude, Table 3.2 shows that the difference to the telephone group (between .03 and .04) is only significant (5%) for the web group. To simplify interpretation, we find from the models in Table 3.2 predicted probabilities of a changed employment status in the controlled model of 10.4% in the telephone group, 13.0% in the mixed mode group, 14.4% in the mixed mode to web group, and 14.8% in the web group. For a change in jobs (Table 3.1) it is the opposite: respondents in the mixed-mode groups are more likely to report job changes (the probability is .06 higher) than respondents in the telephone group, whereas the web group does not differ significantly. Another association emerges for the experience of unemployment (Table 3.1): the mixed-mode-to-web group,

Table 2 Distribution of the variables measuring change and event occurrence by mode and wave (significant (5% level) differences by mode in bold)

		Wave 1				Wave 2 change between waves			
		N	No (%)	Yes (%)	INR (%)	N	No (%)	Yes (%)	INR (%)
Change in employment situation#	Tel.					428	**90.0**	**10.0**	0.0
	Web					1640	**86.0**	**14.0**	0.0
Change in jobs	Tel.	383	**91.1**	**8.1**	0.8	290	**91.4**	**8.6**	0.0
	Web	1642	**83.5**	**13.1**	3.4	1287	**82.0**	**14.5**	3.6
>=1 month unemployed*	Tel.	434	97.2	2.8	0.0	319	97.5	2.5	0.0
	Web	1681	96.5	3.5	0.0	1350	96.7	3.3	0.0
Change in partnership status	Tel.					427	93.0	7.0	0.0
	Web					1619	93.0	5.9	1.1
Change in civil status	Tel.					428	97.9	2.1	0.0
	Web					1640	98.4	1.5	0.1
Termination of close relationship	Tel.	603	93.5	6.5	0.0	460	91.1	8.9	0.0
	Web	2371	91.3	6.8	1.9	1920	88.4	8.9	2.7
Death of closely related person	Tel.	603	75.1	24.9	0.0	460	78.5	21.1	0.4
	Web	2371	76.4	21.5	2.1	1920	77.3	20.2	2.4
Change in household size#	Tel.					713	91.4	8.6	0.0
	Web					483	88.8	11.2	0.0
Residential move	Tel.					756	94.6	5.4	0.0
	Web					553	91.5	6.1	2.4

Note: Significance calculated by means of t-tests excluding possible item nonresponse (INR). Probability tests provided almost identical values to the t-tests.
constructed variable, therefore no missing values.
* Of the 12 months before the first interview (or the months since the last interview in the 2018 survey) a valid response about the type of employment (working, inactive, unemployed) is given for 92% of respondents in the telephone mode and 84% in the web mode. Only these valid responses are taken into account here.

Table 3.1 Regressions results: Coefficients from linear probability (OLS) models (marginal effects), individual level, dependent variable measured at both waves

	Change in jobs		Experience of unemployment		Termination of relationship		Death of close other	
Model	base	controlled	base	controlled	base	controlled	base	controlled
Registered landline		-0.0254*		-0.000478		0.0120		0.0106
Male		0.00207		-0.00725*		-0.0356**		-0.0194
Age 31-49 (Ref.: 18-30)		-0.150**		-0.0121**		-0.103**		0.0171
Age 50-60		-0.200**		-0.0189**		-0.124**		0.0522**
Age 61-92		-0.182**		-0.0209**		-0.137**		0.0660**
Neighboring country (Ref.: Swiss)		0.0772**		0.0185**		-0.0101		-0.0278
Other country		-0.0192		0.00293		-0.00804		-0.0581*
Tertiary education		0.0292*		-0.00223		-0.0172*		0.00299
Reference person	-0.0423**	-0.00369	-0.000678	0.00301	-0.0270**	0.0124	0.000695	-0.0149
Wave	-0.0187	-0.0179	-0.00373	-0.00372	0.0252*	0.0285**	-0.0232	-0.0263
web group (Ref. : telephone)	0.0168	0.0190	0.00926*	0.00959*	-0.0181*	-0.0183*	-0.0156	-0.0129
mixed mode group	0.0640**	0.0670**	0.0102	0.0114	0.00135	-0.00134	0.0219	0.0207
mixed mode to web group	0.0620*	0.0657**	0.0210**	0.0221**	-0.0136	-0.0140	-0.00609	-0.00194
Constant	0.160**	0.272**	0.0108	0.0238**	0.0609**	0.142**	0.253**	0.239**
N (Observations)	3,497	3,497	3,700	3,700	5,257	5,257	5,256	5,256
R-squared	0.008	0.061	0.003	0.012	0.005	0.047	0.001	0.008

Note: ** p<0.01, * p<0.05

Table 3.2 Regressions results: Coefficients from linear probability (OLS) models (marginal effects), individual level, dependent variable measured as change between waves

	Change in employment status		Change in partnership status		Change in civil status	
Model	base	controlled	base	controlled	base	controlled
Registered landline		0.0131		-0.0119		-0.0214**
Male		-0.0227		-0.00926		-0.00459
Age 31-49 (Ref.: 18-30)		-0.162**		-0.168**		-0.00137
Age 50-60		-0.161**		-0.180**		-0.00582
Age 61-92		-0.119**		-0.196**		-0.0150
Neighboring country (Ref.: Swiss)		0.00288		-0.0259		-0.0114
Other country		0.0767*		-0.0633**		-0.00499
Tertiary education		-0.0308		0.0177		0.00845
Reference person	-0.0519**	-0.00622	-0.0169	0.0288**	0.0207**	0.0202**
web group (Ref. : telephone)	0.0436*	0.0436*	-0.00777	-0.0139	-0.00283	-0.00574
mixed mode group	0.0255	0.0260	-0.0119	-0.0183	-0.00302	-0.00461
mixed mode to web group	0.0386	0.0401	-0.0179	-0.0236	-0.00653	-0.0100
Constant	0.131**	0.229**	0.0801**	0.216**	0.00896	0.0322**
N (Observations)	2,068	2,068	2,029	2,029	2,067	2,067
R-squared	0.009	0.048	0.002	0.091	0.007	0.018

Note: ** p<0.01, * p<0.05

Table 3.3 Regressions results: Coefficients from linear probability (OLS) models (marginal effects), household level, dependent variable measured as change between waves (household size) or at wave 2 (residential move)

	Change in household size		Residential move	
Model	base	controlled	base	controlled
Registered landline		-0.0409*		-0.0565**
Male		0.00827		0.0230
Age 31-49 (Ref.: 18-30)		-0.0796*		-0.0673**
Age 50-60		-0.0201		-0.110**
Age 61-92		-0.0899*		-0.123**
Neighboring country (Ref.: Swiss)		-0.0508		-0.0220
Other country		-0.0584		0.0147
Tertiary education		0.0305		0.00611
web group (Ref. : telephone)	0.0284	0.0198	0.000175	-0.0123
mixed mode group	-0.00458	-0.00183	-0.0180	-0.0157
mixed mode to web group	0.0157	0.00556	-0.00744	-0.0177
Constant	0.0885**	0.173**	0.0657**	0.187**
N (Observations)	1,196	1,196	1,296	1,296
R-squared	0.002	0.024	0.001	0.051

Note: ** p<0.01, * p<0.05

and to a lesser extent the web group are more likely to report unemployment than the telephone group with a probability that is .02 (mixed-mode-to-web group) and .01 (web group) higher, whereas the mixed-mode group does not differ significantly from the telephone group. In sum, changes and events in the domain of employment are more often reported by web respondents, although within the web respondents some variation by experimental group exists (i.e., if the household level is answered by web or telephone in one or both waves).

Among the changes and events in the family domain (change in partnership status, change in civil status since last wave, termination of relationship, death of close other), we find little evidence of differences in reporting by mode. Only for the termination of a relationship we find that respondents in the web group reported this event less frequently than telephone respondents, although the size of the effect was small (-.02). For the two dependent variables on the household level, a change

in the number of household members and whether the household moved, we find no difference by experimental group.

Conclusion

Using the two-wave pilot of the Swiss Household Panel collected in 2017 and 2018, we examined whether there were any differences between the use of telephone and web as a mode of data collection with respect to the reporting of changes over time or the experience of events. Although there is a growing body of research indicating measurement differences by mode, mode differences in longitudinal measurement have so far not received much attention (but see, e.g., Brown & Hancock, 2015). As a first exploratory step, this study assesses differences in reporting by telephone and web mode for several variables in the domain of work and family. These variables either measure change or event occurrence directly by asking the respondent about it (e.g., the experience of unemployment), or by capturing differences in response in the two waves (e.g., a change in civil status).

We find differences by experimental groups that used different modes for the work-related variables, with web respondents somewhat more likely to report changes and events compared with telephone respondents. Moreover, these differences do not disappear once the socio-demographic composition of the sample is taken into consideration, suggesting that it is not driven by observed differences in characteristics of the respondents ending up in each mode due to differences in coverage or the likelihood of a respondent to answer in one or the other mode (nonresponse error). Although other characteristics not included in the study could play a role, these findings suggest that there may be differences in response behavior. Yet, these differences are relatively modest, and are also not simply a clear mode effect: the difference in the protocol matters in the sense that not all protocols including web on the individual level differed from the telephone protocol. We find no clear pattern here: for a change in employment status the web group differed from the telephone group and for a change in jobs the mixed-mode groups differed from the telephone group. The difference between web and telephone is, however, larger than the differences among the different web protocols. The differences between the web protocols can be an artifact due to varying sample sizes, or possibly the mode on the household level matters for responses given on the individual level. This deserves further exploration in future research.

Finally, whereas the employment changes and events were more common in the web group, the termination of a relationship was more common in the telephone group than in the web group. This shows that one mode does not necessarily measure more change or events than another, this may be depending on the variable in question.

We looked in this study only at a limited number of events and changes. As not all changes and events were reported more frequently by web respondents, we cannot generalize to other domains. Future research should incorporate other events and changes. Another limitation is the possibility that although we controlled for several socio-demographic variables and only analyzed respondents who answered in their assigned mode, there may still be uncontrolled selection in the two modes. Also, slightly different initial non-response or attrition across modes may have resulted in somewhat different samples.

In conclusion, although some differences by experimental group emerged, they were small with no clear pattern across work and family variables. For employment status variables, we find evidence that longitudinal data collected by web would produce a higher number of changes and events that respondents report. This finding further underlines the differences between web and telephone as a mode of data collection. Therefore, as web and telephone differ in important ways, longitudinal analyses of data collected in these two modes in a mixed-mode design should always incorporate the mode to obtain valid conclusions.

References

Bowyer, B. T., & Rogowski, J. C. (2017). Mode matters: evaluating response comparability in a mixed-mode survey. *Political Science Research and Methods*, *5*(2), 295-313. https://doi.org/10.1017/psrm.2015.28

Brown, M., & Hancock, M. (2015). National Child Development Survey. 2013 Follow-up: A Guide to the Datasets. London: Institute of Education.

Christian, L. M., Dillman, D. A., & Smyth, J. D. (2008). The effects of mode and format on answers to scalar questions in telephone and web surveys. *Advances in Telephone Survey Methodology*, 12, 250-275.

Chandola, T., & Zhang, N. (2018). Re-employment, job quality, health and allostatic load biomarkers: prospective evidence from the UK Household Longitudinal Study, *International Journal of Epidemiology*, 47(1), 47–57. https://doi.org/10.1093/ije/dyx150

Chang, L., & Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized questionnaires. An experiment, *Public Opinion Quarterly*, 74(1), 154–167, https://doi.org/10.1093/poq/nfp090

Choi, S., Chung, I., & Breen, R. (2020). How marriage matters for the intergenerational mobility of family income: Heterogeneity by gender, life course, and birth cohort. *American Sociological Review*, *85*(3), 353-380. DOI: 10.1177/0003122420917591

Christian, L. M., Dillman, D. A., & Smyth, J. D. (2008). The effects of mode and format on answers to scalar questions in telephone and web surveys. *Advances in telephone survey methodology*, *12*, 250-275.

De Leeuw, E. D. (2018, August). Mixed-mode: Past, present, and future. *Survey Research Methods,*12(2), 75-89. https://doi.org/10.18148/srm/2018.v12i2.7402

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*(3), 370-392. https://doi.org/10.1093/poq/nfi027

Groves, R.M., Fowler, F.J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. 2011. *Survey Methodology*. Hoboken: John Wiley and Sons.

Leopold, T. (2018). Gender differences in the consequences of Divorce: A study of multiple outcomes. *Demography* 55**,** 769–797. https://doi.org/10.1007/s13524-018-0667-6

Nagelhout, G. E., Willemsen, M. C., Thompson, M. E., Fong, G. T., Van den Putte, B., & de Vries, H. (2010). Is web interviewing a good alternative to telephone interviewing? Findings from the International Tobacco Control (ITC) Netherlands Survey. *BMC Public Health*, *10*(1), 1-10. https://doi.org/10.1186/1471-2458-10-351

Roberts, C., & Vandenplas, C. (2017). Estimating components of mean squared error to evaluate the benefits of mixing data collection modes. *Journal of Official Statistics*, *33*(2), 303-334. http://dx.doi.org/10.1515/JOS-2017-0016

Royston, P., & White, I. R. (2011). Multiple imputation by chained equations (MICE): implementation in Stata. *Journal of Statistical Software*, *45*(4), 1-20.

Rözer, J. J., Hofstra, B., Brashears, M. E., & Volker, B. (2020). Does unemployment lead to isolation? The consequences of unemployment for social networks. *Social Networks*, *63*, 100-111. https://doi.org/10.1016/j.socnet.2020.06.002

Tillmann, R., Voorpostel, M., Kuhn, U., Lebert, F., Ryser, V. A., Lipps, O., ... & Antal, E. (2016). The Swiss household panel study: Observing social change since 1999. *Longitudinal and Life Course Studies,* 7(1), 64-78. http://dx.doi.org/10.14301/llcs.v7i1.360

Voorpostel, M., Kuhn, U., Tillmann, R., Monsch, G.-A., Antal, E., Ryser, V.-A., Lebert, F., Klaas, H.S., & Dasoki, N. (2020). Introducing web in a refreshment sample of the Swiss Household Panel: Main findings from a pilot study. *FORS Working Paper Series,* paper 2020-2. Lausanne: FORS.

Voorpostel, M., Lipps, O., & Roberts, C. (2021). Mixing modes in household panel surveys: Recent developments and new findings. In: P. Lynn (Ed.) *Advances in longitudinal survey methodology.* 204-226. Hoboken, NJ: Wiley. https://doi.org/10.1002/9781119376965.ch9

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis, *Journal of Survey Statistics and Methodology*, 5(2), 175–211. https://doi.org/10.1093/jssam/smw024

Appendix

Descriptive statistics (all variables are binary 0/1 variables). Item nonresponse excluded

	N Individual level	Mean Individual level	N Household level	Mean Household level
Change in employment situation	2068	13.2	-	-
Change in jobs	3497	13.1	-	-
Episode of unemployment	3784	3.3	-	-
Change in partnership status	2029	6.2	-	-
Change in civil status	2067	1.6	-	-
Termination of relationship	5257	7.8	-	-
Death of close other	5256	21.8	-	-
Change in household size	-	-	1196	4.1
Residential move	-	-	1296	5.8
Telephone group	5354	19.9	2786	21.6
Both mixed-mode groups (1st wave)	5354	29.5	2786	28.7
Mixed-mode group (2nd wave)	5354	14.4	2786	17.2
Mixed-mode-to-web groupa (2nd wave)	5354	8.1	2786	7.6
Web group	5354	28.1	2786	24.8
Wave	5354	1.45	2786	1.47
Registered landline	5354	68.9	2786	68.4
Male	5354	48.3	2786	43.7
Age 14-30 (for HRP min is 18)	5354	25.6	2786	7.0
Age 31-49	5354	30.7	2786	34.6
Age 50-60	5354	22.0	2786	26.5
Age 61-92	5354	21.7	2786	32.0
Swiss	5354	86.9	2786	87.0
Neighboring country	5354	6.9	2786	7.4
Other country	5354	6.1	2786	5.6
Tertiary education	5354	27.9	2786	31.4
Reference person	5354	47.4	-	-

[a]) The mixed-mode-to-web group is 1 for the observations from the second wave of the households that were moved to the web protocol, and 0 otherwise.

methods, data, analyses | Vol. 16(1), 2022, pp. 51-76 DOI: 10.12758/mda.2021.12

Exploratory Likert Scaling as an Alternative to Exploratory Factor Analysis. Methodological Foundation and a Comparative Example Using an Innovative Scaling Procedure

Thomas Müller-Schneider

Universität Koblenz-Landau, Campus Landau

Abstract

Identifying the dimensional structure of a set of items (e.g., when studying attitudes) is an important and intricate task in empirical social research. In research practice, exploratory factor analysis is usually employed for this purpose. Factor analysis, however, has known problems that may lead to distorted results. One of its central methodological challenges is to select an adequate multidimensional factor space. Purely statistical decision heuristics to determine the number of factors to be extracted are of only limited value. As I will illustrate using an example from lifestyle research, there is a considerable risk of fragmenting a complex unidimensional construct by extracting too many factors (overextraction) and splitting it across several factors. As an alternative to exploratory factor analysis, this paper presents an innovative scaling procedure called *exploratory Likert scaling*. This methodologically based technique is designed to identify multiple unidimensional scales. It reliably finds even extensive latent dimensions without fragmenting them. To demonstrate this benefit, this paper takes up an example from lifestyle research and analyzes it using a novel R package for exploratory Likert scaling. The unidimensional scales are constructed sequentially by means of bottom-up item selection. Exploratory Likert scaling owes its high analytical potential to the principle of multiple scaling, which is adopted from Mokken scale analysis and transferred to classical test theory.

Keywords: dimensional analysis, classical test theory, multiple scaling, exploratory factor analysis, exploratory Likert scaling

A fundamental task and activity of empirical social research involves measuring latent dimensions and assessing their content by means of related indicators. Attitudes, action patterns, preferences, motives, and abilities are typical areas of such dimensions—also referred to as latent constructs or dispositions. Empirically, latent dimensions are inferred from item response patterns by employing specific statistical techniques. Major questions and issues in data analysis and methodology concern the dimensionality of a given domain (or universe of items) such as political attitudes or lifestyle preferences: Is the phenomenon in question structured by only one dimension or by several, and if so, how many dimensions are to be meaningfully distinguished in a certain theoretical context? How might one determine the dimensional structure of a set of items in number and content, and how might one then construct scales for measuring the identified dimensions? Understanding the latent dimensional structure of the data in question is essential to achieving conceptual clarity (Rose, 2014, pp. 21–45).

Among practitioners of social science research, a two-step approach of dimensional analysis predominates, which can also be found in relevant textbooks on data analysis. This approach begins by exploring and determining the dimensional structure of a set of items by means of factor analysis. The items of each of the extracted dimensions are then subjected to an item analysis in order to construct Likert scales according to classical test theory (e.g., Fromm, 2012; Kopp & Lois, 2014). The following article deals with the first step, that is, the exploration of dimensional structures. As for exploratory factor analysis (EFA), it is well known that determining the number of factors to be extracted may be a "knotty issue," as DeVellis (2012, p. 127) puts it. Finding an adequate multidimensional solution is still a crucial methodological challenge of EFA. Underextraction on the one hand and overextraction on the other may lead to substantial misinterpretation. This contribution presents an innovative scaling procedure that can serve as a useful alternative to EFA. I refer to this procedure as exploratory Likert scaling (ELS).

Since exploratory Likert scaling, unlike exploratory factor analysis, is based on the concept of multiple unidimensionality, this article begins with a methodological foundation of exploratory dimensionality analysis in the social sciences. Subsequently, the problem of EFA that pertains here is highlighted and illustrated with an example from lifestyle research. As it turns out, there is an imminent risk of splitting complex latent dimensions across the multidimensional factor space. This renders a gainful and appropriate application of EFA technically complicated and demanding in research practice. Against this backdrop, exploratory Likert scaling is outlined. It is shown to be a straightforward technique of multiple unidimen-

Direct correspondence to
Thomas Müller-Schneider, Universität Koblenz-Landau, Campus Landau
E-mail: tms@uni-landau.de

sional scaling based on classical test theory. Each scale is constructed by employing a "bottom-up" clustering procedure using item discrimination as fundamental criterion. The same example of lifestyle research is then used again to illustrate the analytical potential of ELS for the identification of multiple unidimensional constructs.

The Objective: Identifying Multiple Unidimensional Scales

From a methodological perspective, exploring the dimensional structure of a complex data set is anything but trivial. The starting point for further considerations is a unidimensional item response pattern. In general, unidimensionality (aka homogeneity) means that the components of a test or scale measure the same underlying property (latent dimension). In the context of classical test theory (CTT), unidimensionality implies a high degree of interrelatedness among items (Green et al., 1977; Heidenreich, 1984, p. 370; Nunnally, 1978, p. 274). The concept of internal consistency as a measure of reliability is essentially based on inter-item correlations. In addition, highly associated items are each correlated with the total score of all other items. This corrected item–total correlation (or item discrimination) is considered to be an indicator of the relationship between an item and the true score of the latent dimension in question (DeVellis, 2006, p. 52). Likert scaling, which uses Cronbach's alpha as a measure of internal consistency and the corrected item–total correlation as the criterion for item analysis, is not only the most common application of CTT but also by far the dominant scaling method in the social sciences. It is in this specific sense (i.e., with reference to CTT) that the term *Likert scaling* is used here.

The correlational approach to building unidimensional scales can also be found in item response theory. In Mokken scale analysis[1] for dichotomous items, measures of item discrimination and overall scale homogeneity are derived from the coefficient H_{ij}, which is equivalent to the corrected Phi (Phi/Phi_{max}) in 2×2 tables (Stokman & van Schuur, 1980, p. 23). In contrast to the correlational approach, the Rasch model uses the principle of local independence to define unidimensionality. In practical application, however, the model is not convincing. In the social sciences and especially in sociology it is rarely even used as a scaling method (international educational assessment studies such as PISA are something of an exception). Furthermore, and more fundamentally, it has been found to be unsuitable for assessing unidimensionality (Hattie, 1985; Stelzl, 1979). The option to relax the

1 Mokken scale analysis can be regarded as a nonparametric probabilistic version of Guttman scaling (van Schuur, 2003).

assumption of local independence, which was implemented in the context of model improvements (e.g., TenVergert et al., 1993), does not diminish the severe criticism of a problematic concept of unidimensionality.

The characterization of a unidimensional data structure as a definable group of highly correlated items can be readily extended to complex data structures. An item set with two or more underlying dimensions contains a corresponding number of item groups (or clusters) with specific properties: Within these groups the respective items are highly correlated; between groups the items are only poorly correlated or are not correlated at all. In the case of small and well-ordered correlation matrices, simple data inspection is sufficient to identify homogeneous item clusters. Upon expanding the matrices and increasing their complexity, the limits of this visual method are reached very quickly, so that specific multivariate techniques are required for the exploration, determination, and interpretation of dimensional structures. Basically, any technique capable of identifying homogeneous groups (clusters) of items in correlation matrices, such as factor or cluster analysis, is applicable. In this context, it should be noted that factor analysis can be described as a multidimensional extension of CTT (Fischer, 1974, p. 77).

The identification of suitable multivariate techniques of exploratory dimensionality analysis also requires clarifying the appropriate concept of dimensionality. One has to distinguish between multidimensionality and multiple unidimensionality (Jacoby, 1991, p. 35). Multidimensionality does not simply mean that an area of interest cannot be adequately captured by a single dimension. Additionally, multidimensionality entails the concept of locating objects (variables, individuals) simultaneously within an *n*-dimensional space. In factor analysis, for example, each item is defined multidimensionally by its loadings (correlations) on every factor of the selected solution. Multiple unidimensionality, in contrast, signifies that a complex data structure is represented by two or more unidimensional scales. Each item can be characterized by its relation to the respective set of scale items. To emphasize the difference, complex data structures with more than one underlying dimension can be described by multiple unidimensional scales (latent constructs) as opposed to a single multidimensional solution (Jacoby, 1991, p. 36).

So, then, what is the objective of exploratory dimensionality analysis? Is it to find a single multidimensional solution or multiple unidimensional scales? Research practice yields mixed messages that differ with respect to two typical steps of dimensional analysis. In the first step, an exploratory factor analysis (a multidimensional procedure) is conducted to determine the number of dimensions along with their associated items. In the second, the obtained multidimensional information is used to develop unidimensional scales (i.e., Likert scales according to CTT criteria) (e.g., Kopp & Lois, 2014). This common two-step practice with its switch from multidimensionality to multiple unidimensionality reflects the prominent methodological status of unidimensional constructs. Following McIver

and Carmines (1981), "social scientists should strive to develop and use unidimensional concepts because they are more susceptible to theory-relevant research" (p. 14). They subsequently state:

> Multidimensional concepts, on the other hand, typically hamper such research because they are too ambiguous in terms of their meaning, too difficult to measure in a clear and precise manner, and too theoretically oriented themselves. Their complexity and ambiguity renders them less optimal for the development and assessment of social-science theories. In other words, using unidimensional scaling models to measure unidimensional concepts puts the measurement strategy on the same analytical level (p. 14).

Matters get more complicated by the fact that strict unidimensionality at the level of single items is only of ideal-typical nature. For example, a response to the item "reading a book" may be influenced by different latent dimensions, such as an inclination to enjoy high culture as well as a domestic leisure orientation. The fact that a single item may underlie more than one latent dimension does not, however, imply that the measurement concept of unidimensionality is invalid. This is because unidimensionality refers to the level of the overall scale and thus to the common core of meaning of all scale items. Single items are useful in building a unidimensional scale only to the extent that they tap into this common core (Nunnally, 1978, p. 274). If too strongly affected by one or more "interfering dimensions", an item has to be removed from the scale in question.

Against the methodological background presented and in accordance with common research practice, the objective in exploratory dimensionality analysis is to identify multiple unidimensional scales—or, with respect to the most common scaling model in the social sciences, to identify multiple Likert scales.

The concept of multiple unidimensionality does not disqualify exploratory factor analysis as a helpful first-step tool for the exploration of dimensional structures.[2] However, the multidimensional model of EFA is prone to certain problems that may result in distorted results. To avoid these problems, this paper suggests an alternative that is directly connected to the objective of identifying multiple Likert scales. Before turning to this alternative, exploratory Likert scaling, I will highlight one of the key challenges for exploratory dimensionality analysis that arises from the multidimensional model of factor analysis.

2 Due to model extensions, Rasch scaling can also now be used to define multidimensional structures (Cheng et al., 2009). Unlike EFA, however, it is not even a helpful first-step tool, as it is inherently unsuitable for the *exploration* of dimensional structures.

A Key Challenge for Exploratory Factor Analysis: Selecting an Adequate Factor Space

The term *exploratory factor analysis* is not used consistently in the literature, so its definition should be briefly clarified. First of all, strictly speaking, a distinction has to be made between two analytical models, namely, principal component analysis (PCA) and factor analysis "proper" (FA), the latter of which is based on the model of common factors (Fabrigar et al., 1999, p. 275). In addition to its confirmatory variant (CFA), which is not relevant here and therefore not considered, factor analysis (FA) can also be used for the dimensional exploration of complex correlation matrices. This type of procedure is called exploratory factor analysis. In social science research (especially in German-speaking countries), however, the term exploratory factor analysis is also used when principal component analysis (PCA) is employed as a statistical method for exploring dimensional structures. It is in this latter sense that I will speak of exploratory factor analysis hereafter. It should be noted that the common factor model and PCA differ significantly in their analytical basis but usually lead to equivalent results—even with regard to the number of factors or components considered (Wolff & Bacher, 2010, p. 349; Velicer & Jackson, 1990). It should also be stated that the problems encountered in the context of EFA do not arise for confirmatory factor analysis, as the number of factors in CFA is predetermined by theoretical considerations. However, the social science applications of factor analysis are mostly exploratory in nature because they are typically not focused on hypothesis testing but on the dimensional interpretation of complex item batteries.

EFA is undoubtedly a useful tool to discover multiple dimensions in terms of homogenous item clusters: "Variables that are correlated with one another but largely independent of other subsets of variables are combined into factors" (Tabachnick & Fidell, 2007, p. 607). In PCA, factors (used here synonymously with principal components) are extracted stepwise from the correlation matrix in such a way that they explain the maximum of the (remaining) variance among all items. Therefore, the first factor accounts for the most variance, and each successive factor accounts for a decreasing portion of the item variance. All factors are uncorrelated with each other. This iterative extraction procedure is used to construct an n-dimensional space of orthogonal factors, whereby the maximum number of extracted factors corresponds to the number of items entered in the analysis (aka the full component model). As the objective of exploratory dimensionality analysis is to identify clusters of highly intercorrelated items, the full component model is obviously worthless. A major task of EFA, therefore, is to determine a more parsimonious solution with an adequate number of factors being extracted. To this end, a variety of procedures (or stopping rules for further extraction) have been suggested (e.g., Peres-Neto et al., 2005; Hoyle & Duvall, 2004). The most relevant in

research practice are based on the eigenvalue (Wolff & Bacher, 2010). This measure represents the amount of variance explained by each factor and equals the number of items in the full component model. The best-known and most commonly applied stopping rule for factor extraction, at least in the social sciences, is the so-called Kaiser rule or eigenvalue-greater-than-one rule (the default option in SPSS). As the name of the rule indicates, it states that a factor is to be extracted (retained) as long as its eigenvalue is greater than one. Another criterion is derived from the visual inspection of a scree plot, which represents factors with respect to their eigenvalue in a downward curve. According to the scree test (Cattell 1966), the point before the curve levels off (the "elbow") denotes the number of factors to be retained as significant. A less well-known procedure is the parallel test, which compares randomly generated eigenvalues with empirical ones. It needs to be stressed that the rules mentioned here typically do not lead to the same results; moreover, they (especially the scree test) may at times be ambiguous and therefore open to interpretation (Ledesma et al., 2015; Tabachnick & Fidell, 2007, p. 644–646; Wolff & Bacher, 2010, p. 343). To put it more generally, determining the number of factors is a fundamental issue in factor analysis and remains a major challenge that has come under considerable debate (Peres-Neto et al., 2005; Ledesma et al., 2015; Hoyle & Duval, 2004). Problems with (orthogonal or oblique) factor rotation[3] are also substantial (Sakaluk & Short, 2017) but not directly relevant to the present topic and can thus be neglected in the context of this article. Regardless, the selection of the number of factors is at any rate more critical than the selection of a rotation method (Tabachnick & Fidell, 2007, p. 644).

With respect to an appropriate number of factors to extract, factor analysis is susceptible to misspecifications, and this might lead to biased results in terms of identifying relevant latent dimensions. Such misspecifications can take the form of underextraction (i.e., extracting too few factors) or overextraction (i.e., extracting too many factors). Whereas it is widely agreed that underextraction leads in many cases to more severe distortions than overextraction (de Winter & Dodou, 2016; Fava & Velicer, 1996, p. 908; Wood et al., 1996), the extracting of surplus factors is presumably the more common problem. This is due to the fact that overextraction usually occurs in cases where the popular Kaiser criterion is employed, especially in combination with a large number of variables (Zwick & Velicer, 1986; Fava & Velicer, 1992, p. 388). It is assumed that a small number of extra factors may do little harm, but substantial overextraction results in the severe problem of factor fission (Cattel, 1978, p. 168; Fava & Velicer, 1992, p. 389; Wood et al., 1996). Factor fission (or factor splitting) denotes the phenomenon that items belonging to a common latent dimension are dispersed across different factors. An older study

3 In orthogonal rotation, extracted factors remain statistically independent (uncorrelated). In oblique rotation, the factors are allowed to correlate.

with real data sets even documented a massive change in the factor structure as the number of extra factors extracted was increased (Levonian & Comrey, 1966).

I will illustrate the problem of factor splitting by an example from lifestyle research. It will be picked up again in the following section for comparison with exploratory Likert scaling. The focus of the example is on three well-known schemes of everyday aesthetics identified and theorized by Schulze (1992), which are the high-culture scheme, the trivial scheme, and the tension scheme. Along with age and education, these three aesthetic patterns of everyday life are among the constituent characteristics of five social milieus. These aesthetic schemes are complex cross-situational response tendencies that are considered here as theoretically relevant dimensions (unidimensional constructs). The schemes were obtained by means of exploratory analysis from a very broad range of individual preferences and action tendencies within a total of 110 relevant items (Schulze, 1992, pp. 595–598). All items were measured on a five-point response scale. For the following analyses, the original data of the study with a total of 1,024 interviews were used. First, a unidimensional secondary analysis of each of the three schemes was performed to obtain rather homogeneous scales with items having discriminatory power of at least 0.45. The results are presented in Table 1. As can be seen, all three scales, as measured by Cronbach's alpha, have a very high internal consistency.

These three dimensions (unidimensional scales) are now contrasted with the results of an exploratory factor analysis (PCA), which was computed for the same data with the usual specifications. Using the eigenvalue-greater-than-one rule, 25 factors were extracted (see appendix, Table A1) and then subjected to varimax rotation. What matters here is not the individual results of the factor analysis but the mapping of the three unidimensional scales in the multidimensional factor space. As Table 1 shows, both the trivial scheme and the tension scheme are fully captured by a single factor each (factor 1 and 2). Almost all scale-specific items show high factor loadings. Only pub attendance (tension scheme) stands out with a lower loading on factor 2 and a simultaneous loading on factor 8. This, however, is completely unproblematic, provided that the item would be included in a subsequent unidimensional scale analysis.

A substantially different picture emerges for the high-culture scheme. First, it should be noted that factor 3 reflects significant manifestations of this aesthetic orientation. A number of relevant items (e.g., classical music and literature, visiting exhibitions and museums) show high factor loadings. Nevertheless, the problem of factor fission is evident. Fragments of the latent dimension and single items are scattered across the multidimensional factor space. For example, the fragment that primarily addresses literature about the inner life (self-awareness and psychological problems) is found on factor 7. Only the item that captures the preference for poetry also loads on factor 3. Further, the marker items of factor 9, which revolve around private educational inclination, have no discernible connection to factor 3. Reading

Der Spiegel loads exclusively on factor 14, and reading *Die Zeit* shows no substantial loading on any of the 25 factors at all. Engagement with literature has also broken out of the high-culture scheme and is located on factor 12 (with classical and modern literature building a "bridge" to factor 3). Overall, it must be noted that the high-culture scheme in the totality of its meanings is not evident in the exploratory factor analysis presented here.

Table 1 Everyday aesthetic schemes: Results from unidimensional scaling and exploratory factor analysis

Dimension	Items (preference for, interest in, inclination to…)	Corrected item–total correlation	Factor and loadings	
			FA 1	
Trivial scheme	*Heimat* films[1]	0.70	0.76	
	Shows/quizzes (TV)	0.64	0.62	
	Popular theater (TV)	0.73	0.75	
	Local broadcasts	0.51	0.47	
	Nature broadcasts	0.46	0.41	
	Light music	0.57	0.62	
	German hits	0.67	0.71	
	German folk songs	0.79	0.76	
	Bavarian folk music	0.79	0.79	
	Brass music	0.78	0.78	
		α = 0.91		
			FA2	FA8
Tension scheme	Pop and rock music (TV)	0.73	0.70	
	Rock music	0.77	0.72	
	Oldies (e.g., The Beatles)	0.56	0.61	
	Reggae music	0.68	0.72	
	Soul music	0.69	0.77	
	Pop music	0.81	0.78	
	Folk music	0.60	0.68	
	Blues music	0.58	0.69	
	Attending concerts (rock, pop, jazz)	0.63	0.55	
	Going to the movies	0.62	0.51	
	Going to a pub	0.48	0.36	0.40
	Going to a discotheque	0.54	0.44	
		α = 0.91		

Table 1 continued

Dimension	Items (preference for, interest in, inclination to…)	Corrected item–total correlation	Factor and loadings				
			FA3	FA7	FA9	FA12	FA14
High-culture scheme	Classical music	0.62	0.76				
	Contemporary classical music	0.51	0.64				
	Classical concerts	0.56	0.68				
	Theater (TV)	0.49	0.70				
	Newspaper: culture section	0.50	0.48				
	Visiting exhibitions/ galleries	0.55	0.46				
	Poems	0.58	0.45	0.45			
	Self-awareness literature	0.55		0.80			
	Psychological problem literature	0.59		0.76			
	Writing (e.g., diary)	0.46		0.38			
	Classical literature	0.75	0.56	0.31		0.33	
	Modern literature	0.70	0.39			0.38	
	Books on social/ political issues	0.63		0.38		0.37	
	Book reading	0.50				0.57	
	Courses, education	0.50			0.72		
	Language learning	0.47			0.71		
	Professional training (at home)	0.54			0.67		
	Reading *Der Spiegel*	0.51					0.51
	Reading *Die Zeit*	0.52					
		$\alpha = 0.91$					

[1] Sentimental films in an idealized rural setting

Note: For clarity, only factor loadings greater than 0.3 are reported (for "Reading *Die Zeit",* there was no factor loading with an absolute value greater than 0.3).

However, factor analysis does not fundamentally fail to represent fragments and individual items of the high-culture scheme within one single factor. One only has to reduce the number of extracted factors in such a way that the relevant construct (i.e., the high-culture scheme) is not split up and becomes visible in its entirety. The Kaiser criterion was not used for this purpose; rather a series of analyses with a gradually decreasing and predetermined number of factors was computed. Reducing the number of factors from eight to seven yielded the expected switch in factor structure, which is to say, the entire dimension of the high-cultural orientation was mapped onto a single (the first) factor. The second factor represents the tension scheme and the third the trivial scheme. The remaining four factors reveal further aspects of everyday preferences, which have to do with sports, shopping, maintaining social contacts, and domestic activities. This solution with seven

factors is roughly in the range considered by the scree plot (five or six factors, depending on its interpretation). The solutions with three to six factors also represent each of the three relevant aesthetic patterns in a single factor. In social science research, it is quite common not to adhere too strictly to potentially problematic or ambiguous statistical criteria in the search for an appropriate *n*-dimensional factor space but rather to consider a range of conceivable solutions. This procedure is not only completely in line with the above methodological considerations for exploratory dimensionality analysis (identifying multiple unidimensional constructs) but also explicitly advised (Wolff & Bacher 2010, p. 343), and for good reason. Ultimately, it comes down to the interpretability and scientific usability of factors. As Tabachnick and Fidell (2007) concisely put it, "A good PCA … 'makes sense'; a bad one does not" (p. 608).

At this point, the question arises as to whether there is a different and possibly more suitable method to identify multiple unidimensional constructs than to try out a range of conceivable *n*-dimensional factor solutions. Exploratory Likert scaling offers a useful alternative to this strategy.

Exploratory Likert Scaling

Exploratory Likert scaling is an innovative method for discovering clusters of internally well and externally poorly correlated items within a given data set. It is based on a generalizable scaling procedure that works according to the crystallization principle and was originally proposed by Mokken (1971) for a step-by-step scale construction. The nucleus of crystallization is the maximally homogeneous "two-item scale" of a data set, which is then gradually extended by "bottom-up item selection" to a scale that meets the conditions of the monotone homogeneity model (Hemker et al., 1995, p. 342; Sijtsma et al., 1990, pp. 181–183). By taking the corrected item–total correlation (item discrimination according to CTT) as a coefficient of scalability, the crystallization principle can be applied immediately to the construction of a Likert scale. I suggest the following algorithm on the basis of bottom-up Mokken scale analysis:

1. Find the two items with the highest positive correlation.[4] Consider this pair of items as the potential crystallization nucleus of a Likert scale and calculate their total score.
2. From the remaining items, select the one that correlates most highly with the total score (i.e., has the highest item discrimination). Expand the scale nucleus by this item and recalculate the total score (with *n*+1 items).

4 The algorithm can also account for negative correlations (reversed items), although this is not relevant in the present context.

3. Repeat the process of step 2 until a predefined lower bound (minimum item–total correlation) is reached. The bottom-up item selection is then completed for this scale.

The use of the (corrected) item–total correlation as a criterion for scale extension not only aims at finding items of a latent dimension that are as discriminative as possible but serves at the same time to establish the internal consistency of the emerging scale in an optimal way. Higher item–total correlations of the selected items will result in a higher average inter-item correlation and thus also in a higher value for Cronbach's alpha (Lord & Novick, 1968, pp. 330–331). After the construction of the first scale is completed with step three of the algorithm, the search for additional scales begins. This is accomplished by means of "multiple scaling" (Mokken, 1971, pp. 194–195; Sijtsma et al., 1990, p. 185), which means that the entire scaling process is iterated. The algorithm therefore has to be extended by a fourth step:

4. Try to create a further scale from the remaining item pool by repeating steps 1 to 3. Then start again with step 4 and continue the process until no new scale nucleus can be found (specified by the minimum item–total correlation).

Multiple scaling according to the crystallization principle enables a sequential identification of groups of internally highly correlated items and thus of multiple unidimensional scales. Multiple scaling is also appropriately seen as "sequential clustering" of items (van Abswoude et al., 2004). Especially with regard to the main objective here, the exploration of the dimensional structure of an item pool using this procedure may, however, lead to problematic results. Depending on the value for the specified item–total correlation, this can be expected to obscure the dimensional structure of the data. The corresponding problem is already familiar from multiple Mokken scaling (Sijtsma & Molenaar, 2002, p. 80). A value close to zero would merge (almost) all items into a single scale, even if two or even more dimensions clearly underlie the data. In the context of EFA, one would use the term *underextraction*. If, on the other hand, one chooses a rather high value for the minimum item discrimination, the above algorithm would split unidimensional scales into a number of fragments. This is equivalent to the problem of factor fission. However, the problem can be easily solved, and above all in a way that optimizes the exploratory potential of multiple scaling. The process of multiple scaling is divided into two steps wherein a high value for the minimum item discrimination is deliberately set in the first step in order to search for very homogeneous kernels of potential scales (search procedure). These kernels then serve as starting sets for the second scaling step (extension procedure), in which the minimum item–total correlation is significantly lowered and overlapping scaling is allowed. Overlapping scale construction means that each item can be assigned not only to the first but also to all subsequent potential scale kernels. Each of them has the opportunity, so to speak,

to collect all scalable items. Now the exploratory potential of the entire scaling procedure, or exploratory Likert scaling, becomes visible: Each latent dimension can be determined reliably and completely even if only one scale fragment of the respective dimension was identified in the first step of the scaling procedure. What remain are only single non-scalable items or item groups that cannot be interpreted in a meaningful way or have no scientific use in the given context.

I will now contrast the factor analysis discussed above with an exploratory Likert scaling based on the same items. For this purpose, the R package "elisr" (Bißantz, 2021) was used. This package was developed on the author's initiative specifically for exploratory Likert scaling. All 110 items of the everyday aesthetic preferences were included in the search procedure. For the construction of potential scale kernels, an item–total correlation of 0.60 was set as the lower bound (in principle, it is reasonable to start several runs with varying lower bounds to ensure that kernels of all relevant dimensions are found). Table 2 presents the potential kernels in the order of their construction.

The software reports the average inter-item correlation and Cronbach's alpha as descriptive measures of internal consistency as well as the corrected item–total correlation that is essential for scale construction. To be precise, one should speak of a marginal item–total correlation, since this value reflects the item discrimination that is found at the moment when the scale is extended by the item in question (with subsequent scale expansion, this value may change). As can be seen in Table 2, a total of nine potential kernels is found at a minimum item–total correlation of 0.60, with some of them consisting of only a two-item scale. The first line of the respective item lists shows the two items that were fused first. Due to the specification of the search procedure, all kernels show a very high internal consistency (measured by the average inter-item correlation or Cronbach's alpha). The three relevant dimensions (the high-culture, trivial, and tension scheme) are all represented by scale fragments. The high-culture scheme even appears in five fragments with different contents (scales 3, 5, 6, 7, and 8), whereby the two items signaling an interest in opera (scale 3) are curiously not included in the overall scale presented in Table 1 above. The contents of the remaining two kernels (scales 2 and 9), which indicate a preference for information about sports and for domestic pursuits, were also identified in the EFA.

In the second scaling step (the extension procedure), the minimum item–total correlation was substantially decreased in order to allow each scale kernel to be extended with relevant items. The value of the minimum item–total correlation is now no longer oriented towards the search for very homogeneous scale kernels but rather towards the still acceptable item discrimination with respect to an overall scale. In this context, the value of 0.3 is often mentioned, but content-related aspects should also be taken into account. So as not to generate scales that were too extensive for reasons of clarity, a lower bound of 0.40 was selected in the present

Table 2 Results of the search procedure

Scales and steps	Items	r_{itm}	$\bar{r}$	α
Scale 1				
1	Brass music ǁ Bavarian folk music	0.87	0.87	0.93
2	German folk songs	0.78	0.79	0.92
3	Popular theater (TV)	0.66	0.70	0.91
4	*Heimat* films[1]	0.69	0.67	0.91
5	German hits	0.64	0.63	0.91
6	Shows/quizzes (TV)	0.61	0.59	0.91
Scale 2				
1	Sports (newspaper) ǁ Sports (TV)	0.80	0.80	0.89
2	Sports magazines	0.62	0.65	0.85
Scale 3				
1	Opera (music) ǁ Opera (TV)	0.78	0.78	0.88
Scale 4				
1	Pop music ǁ Rock and pop (TV)	0.76	0.76	0.86
2	Rock music	0.78	0.74	0.90
3	Soul music	0.63	0.66	0.89
4	Reggae music	0.66	0.62	0.89
Scale 5				
1	Self-awareness ǁ Psychological problem literature	0.76	0.76	0.86
Scale 6				
1	Classical literature ǁ Modern literature	0.69	0.69	0.82
2	Books on social/political issues	0.61	0.60	0.82
Scale 7				
1	Courses/education ǁ Professional training (at home)	0.66	0.66	0.79
Scale 8				
1	Classical concerts ǁ Classical music (preference)	0.62	0.62	0.76
Scale 9				
1	Cleaning up ǁ Tidying	0.62	0.62	0.76

[1] See footnote 1, Table 1.

Note: Minimal item–total correlation for the search procedure = 0.6; r_{itm} = marginal item–total correlation; $\bar{r}$ = average inter-item correlation; α = Cronbach's alpha.

exploratory analysis. Of the nine scales extended in the second scaling step, four are documented in this paper (scales 3 and 5 below in the text and scales 1 and 4 in the appendix, Table A2). The trivial scheme (extended scale 1) is now represented by 12 items. Drawing a line below the item "nature broadcasts" yields the precise set of ten items listed in Table 1. The two remaining items (preference for comedy movies and light fiction) also belong to the extended scale because the minimum item–total correlation chosen for the extension procedure (0.40) is lower than that for the scales compiled in Table 1 (which is 0.45). The same can be said for the tension scheme. In the expansion process of scale kernel 4, three additional items (visiting a night club, meeting in the city, interest in sci-fi/fantasy on TV) were included in addition to those shown in Table 1.

The scale extensions that affect the high-culture scheme are of particular interest and warrant closer scrutiny. The most important result can be seen in the fact that the scheme crystallizes completely at all its scale fragments found in the first step. Contrary to the EFA, no splitting of the latent dimension occurs. This is exactly what is ensured by the extension procedure in the second scaling step. If we look at extended scale 5 (Table 3), for example, we can see that the scale kernel, which is about self-awareness and dealing with psychological problems, is first expanded to include indicative topics (e.g., classical music and literature) and then educationally relevant content. Again, if one were to draw a boundary line at a marginal discriminatory power of 0.45, one would find all items of the high-culture scheme from Table 1. The same holds for the extended scales 6 and 7 (not documented in the appendix), whereby scale 6 starts with indicative high-culture topics and scale 7 with education-specific content. The extended scale 3, with its crystallization nucleus of the two preferences for opera (music, TV), also collects all relevant items, but the education-specific content is now included only below a minimum item discrimination of 0.45.[5] This demonstrates that the lower bound for exploratory purposes should be set rather lower than higher. Items that are borderline in terms of content or statistics can be excluded again for the final scales at a later stage. A final item selection is warranted in any case, given that, as already mentioned, the item–total correlation of an item can change its value in the course of the expansion procedure. Thus, the (marginal) item–total correlation of the two opera items, which is very high when they are merged to form a scale nucleus (0.78, identical to the bivariate correlation), falls below 0.45 in the further extension process. This is the reason why the two items were not included in the high-culture scale reported in Table 1. Extended scale 8 (not documented in the appendix) also contains all items of the high-culture dimension.

5 The items "Courses, education" and "Language learning" both reach the threshold of 0.45, but only after the item "Professional training" is included, which has a marginal discrimination power below this value (0.44).

As for the remaining two scale kernels from the search procedure (scales 2 and 9), these were enlarged by two and four items, respectively (not documented in the appendix). Although these scales are easy to interpret, they remain fragmentary (at least in the analyzed data set and for the specified minimum item–total correlation of 0.40). The extended scale 2 with a total of five items still focuses on sports, and scale 9, with its six items, revolves all around topics stereotyped as female (domestic chores, fashion, cosmetics). It is important to note that these two scales (fragments) neither influence nor even interfere with the bottom-up and sequential construction of the three relevant dimensions. Exploratory Likert scaling is not affected by irrelevant items or scale fragments. This, however, does not apply to the same extent to EFA. First of all, the irrelevant fragments "build" factors with an eigenvalue greater than one and are thus involved in determining the *n*-dimensional

Table 3 Results of the extension procedure: extended scales 3 and 5

Scales and steps	Items	r_{itm}	$\bar{r}$	α
Scale 3				
1	Opera (music) || Opera (TV)	0.78	0.78	0.88
2	Classical music	0.60	0.63	0.84
3	Concerts with classical music	0.59	0.57	0.84
4	Theater (TV)	0.58	0.54	0.85
5	Classical literature	0.56	0.51	0.86
6	Contemporary classical music	0.56	0.49	0.87
7	Poems	0.51	0.46	0.87
8	Modern literature	0.53	0.45	0.88
9	Newspaper: culture section	0.52	0.43	0.88
10	Visiting exhibitions/galleries	0.53	0.42	0.89
11	Books on social/political issues	0.49	0.41	0.89
12	Psychological problem literature	0.47	0.40	0.89
13	Self-awareness literature	0.48	0.38	0.90
14	Book reading	0.47	0.38	0.90
15	Reading *Die Zeit*	0.45	0.36	0.90
16	Reading *Der Spiegel*	0.45	0.36	0.90
17	Professional training (at home)	0.44	0.35	0.91
18	Courses, education	0.45	0.34	0.91
19	Language learning	0.45	0.33	0.91
20	Writing (e.g., diary)	0.43	0.33	0.91
21	Documentaries (TV)	0.43	0.32	0.91
22	Newspaper: politics section	0.43	0.31	0.91
23	Jazz music	0.42	0.31	0.91

Table 3 continued

Scales and steps	Items	r_{itm}	$\bar{r}$	α
Scale 5				
1	Self-awareness ‖ Psychological problem literature	0.76	0.76	0.86
2	Books on social/political issues	0.48	0.55	0.79
3	Modern literature	0.57	0.52	0.81
4	Classical literature	0.66	0.52	0.84
5	Poems	0.57	0.50	0.86
6	Classical music	0.52	0.47	0.86
7	Classical concerts	0.53	0.45	0.87
8	Visiting exhibitions/galleries	0.52	0.43	0.87
9	Contemporary classical music	0.52	0.42	0.88
10	Newspaper: culture section	0.53	0.41	0.88
11	Theater (TV)	0.54	0.40	0.89
12	Book reading	0.49	0.39	0.89
13	Reading *Die Zeit*	0.47	0.38	0.90
14	Reading *Der Spiegel*	0.48	0.37	0.90
15	Professional training (at home)	0.47	0.36	0.90
16	Courses, education	0.47	0.36	0.90
17	Language learning	0.46	0.35	0.91
18	Writing (e.g., diary)	0.46	0.34	0.91
19	Jazz music	0.43	0.33	0.91
20	Documentaries (TV)	0.41	0.33	0.91
21	Newspaper: politics section	0.42	0.32	0.91
22	Opera (music)	0.41	0.31	0.91
23	Opera (TV)	0.43	0.31	0.91

Note: Minimal item–total correlation for the extension procedure = 0.4; r_{itm} = marginal item–total correlation; $\bar{r}$ = average inter-item correlation; α = Cronbach's alpha.

factor space (at least according to the greater-than-one rule). In any case, the fragments must be represented in the selected *n*-dimensional space. This cannot leave relevant factors completely unaffected because they have to be mapped in the same multidimensional factor space too. Whether insignificant scales (or fragments) lead to distortions in exploratory factor analysis is difficult to answer in general, if only because the results also involve substantial subjective decisions by the researcher. It must be added here, however, that the very existence of irrelevant item clusters makes it difficult in principle to speak of a "true" dimensionality or a "true" number of factors, contrary to what is sometimes found in the literature (e.g., Fava & Velicer, 1996, p. 908; Wood et al., 1996).

The explanatory scaling process to identify relevant latent constructs is followed by a second step of dimensional analysis: item analysis to construct the final scales. This second step is "business as usual" and outside the focus of this contribution (see Introduction). Nevertheless, some procedural remarks might be helpful. The final item analysis is based on all items of an extended scale that was selected for representing a latent dimension. Items found to have too little discriminatory power (corrected item–total correlation) must be removed from the scale in question. Usually, a respondent's scale value is then computed as the summated score of all items included in the scale. But how does one deal with overlapping items? Following the concept of multiple unidimensionality, each item should be assigned to one scale only (according to statistical or content criteria). This should also be done in order not to overestimate the correlation between the final scales on grounds of multiply allocated items.

Conclusion and Discussion

This contribution has focused on exploratory dimensionality analysis in the social sciences. It began with methodological considerations on dimensional structures in complex data sets and their empirical identification. It was noted that, with reference to CTT, structures with more than one latent dimension are empirically reflected in a corresponding number of clusters with internally well and externally poorly correlated items. This contribution then further elaborated that the main objective of exploratory dimensional analysis is to find multiple unidimensional constructs as opposed to a single multidimensional solution. With reference to the common research practice of unidimensional scaling within the framework of CTT, this means identifying multiple Likert scales.

Since exploratory factor analysis is a genuinely multidimensional procedure, it is not designed to identify multiple unidimensional structures. Instead, the technique searches for a single *n*-dimensional (orthogonal) factor space to adequately represent multiple item clusters. One of the main methodological challenges of EFA is to determine the number of factors that span this *n*-dimensional space. If the most frequently used statistical criterion, the eigenvalue greater-than-one rule, is applied, it is widely acknowledged that one has to reckon with overextraction and factor fission. As was illustrated by the example provided from lifestyle research, this risks capturing extensive latent dimensions only in the way of disconnected fragments and to completely overlook single items scattered in the overextracted *n*-dimensional space. The best practice of an exploratory dimensionality analysis by means of factor analysis, at least in the social sciences, is therefore not to rely primarily on ambiguous statistical criteria but (as is often done anyway) to check a range of conceivable solutions and then decide on the interpretability and scientific

significance of the factors found. The respective items of each factor can then be subjected to item analysis in order to construct unidimensional Likert scales.

Exploratory Likert scaling is a useful alternative for analyzing dimensional structures. This novel method belongs to the multiple unidimensional scaling approach, as it has already been implemented in Mokken scale analysis. Compared to the statistically complex EFA, exploratory Likert scaling (ELS) is a straightforward and completely different technique. It does not require a predefinition of a multidimensional (orthogonal) space and is better suited for identifying multiple unidimensional constructs than EFA owing to its bottom-up item selection and sequential clustering. A two-step multiple scaling strategy that combines an initial search for homogeneous scale kernels with their subsequent expansion not only avoids a methodological problem with sequential clustering but also optimizes the exploratory potential of the scaling procedure. Starting with any fragment of a unidimensional construct, the procedure interlinks all relevant contents of the construct. No splitting will occur, even if the unidimensional construct is complex in terms of the underlying empirical association structure. Also, large numbers of items do not pose any difficulties for the multiple unidimensional scaling approach. Especially in exploratively demanding data situations—large numbers of items, and high degrees of complexity but with unidimensional associations between items nevertheless—ELS is superior to EFA, the latter of which may quickly become confusing or even misleading in the case of substantial overextraction.

In the literature, there have been proposals on how to optimize factor analysis in order to make better decisions on the number of factors to retain. Lawrence and Hancock (1999), for example, state that "[t]he implementation of more precise factor extraction decision heuristics is essential" (p. 569). Referring to Zwick and Velicer (1986), they point to the minimum average partial procedure and parallel analysis as "extremely promising alternatives" (p. 569) to conventional practice. Ledesma et al. (2015) suggest enhancements of the scree test in the hope of providing better tools to determine the number of factors to retain. However, for identifying multiple unidimensional constructs, the approach of optimizing statistical (formal) criteria to define the number of factors is only of limited value. One reason for this is that the number of factors cannot be totally objectified on the basis of statistics alone. Apart from simulation purposes, there is, as mentioned above, no absolutely "true" dimensionality of a set of items, at least in the field of social sciences. Above all, the attempt at statistical optimization proceeds in the *wrong direction*. From the methodological point of view of multiple unidimensional scaling, the main problem of EFA is that an *n*-dimensional space has to be defined at all. The multidimensional approach creates unnecessary statistical complexity in exploratory dimensionality analysis, which in turn may lead to misspecifications and inappropriate results.

Multiple scaling can in principle also be performed using hierarchical clustering methods, as has already been suggested for Mokken scaling (van Abswoude et

al., 2004). The dendrogram visualizes the fusion process and can be interpreted similarly to exploratory Likert scaling in terms of bottom-up scaling. With an appropriate fusion algorithm, a hierarchical agglomerative cluster analysis of items, as was demonstrated in a case study for Mokken scaling (Müller-Schneider, 2001), leads to substantially the same results as a two-step sequential scale construction (i.e., a search and extension procedure). Nevertheless, there are reasons to prefer exploratory Likert scaling. As the number of items increases, the dendrogram becomes less clear, which considerably impairs the visual analysis of bottom-up item selection and the dimensionality of the data. In addition, and more importantly, exploratory Likert scaling with its characteristic coefficients is, unlike cluster analysis, directly integrated into the analytical framework of dimensional analysis. Item–total correlation determines the constitution as well as the extension of a scale kernel, and at each step, the internal consistency of the resulting scale can be precisely traced by the average item correlations and Cronbach's alpha.

Besides the reliable identification of multiple unidimensional constructs, there is another noteworthy advantage of ELS. In order to interpret the determined scales appropriately, there is no need for such a thing as factor rotation. This being the case, ELS avoids unnecessary model complications and all the specific issues involved therein. Consequently, there is also no need for an always somewhat arbitrary oblique rotation to map any given correlations between latent dimensions. Since the statistical identification of multiple dimensions using ELS does not demand a predefined space of orthogonal dimensions, the constructs can correlate with each other (or not) in a natural way from the outset.

References

Bißantz, S. (2021). elisr: Exploratory Likert Scaling. R package version 0.1.1.

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276. doi:10.1207/s15327906mbr0102_10

Cattell, R.B. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences.* New York: Plenum Press.

Cheng, Y.-Y., Wang W.-C., & Ho, Y.-H. (2009). Multidimensional Rasch Analysis of a Psychological Test With Multiple Subtests: A Statistical Solution for the Bandwidth-Fidelity Dilemma. *Educational and Psychological Measurement, 69*(3), 369–388. doi:10.1177/0013164408323241

DeVellis, R. F. (2006). Classical Test Theory. *Medical Care*, *44*(11), 50–59. doi:10.1097/01.mlr.0000245426.10853.30

DeVellis, R. F. (2012). *Scale Development: Theory and Applications* (3rd ed.). Thousand Oaks, CA: Sage Publications.

De Winter, J. C. F., & Dodou, D. (2016). Common Factor Analysis versus Principal Component Analysis: A Comparison of Loadings by Means of Simulations.

Communications in Statistics – Simulation and Computation, *45*(1), 299–321. doi:10.1080/03610918.2013.862274

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychological Methods*, *4*(3), 272–299. doi:10.1037/1082-989X.4.3.272

Fava, J. L., & Velicer, W. F. (1992). The Effects of Overextraction on Factor and Component Analysis. *Multivariate Behavioral Research*, *27*(3), 387–415. doi:10.1207/s15327906mbr2703_5

Fava, J. L., & Velicer, W. F. (1996). The Effects of Underextraction in Factor and Component Analyses. *Educational and Psychological Measurement*, *56*(6), 907–929. doi:10.1177/0013164496056006001

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests: Grundlagen und Anwendungen*. Bern, Stuttgart, Vienna: Huber.

Fromm, S. (2012). *Datenanalyse mit SPSS für Fortgeschrittene 2: Multivariate Verfahren für Querschnittsdaten* (2. Auflage). Wiesbaden: VS Verlag für Sozialwissenschaften.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of Coefficient Alpha as an Index of Test Unidimensionality. *Educational and Psychological Measurement*, *37*(4), 827–838. doi:10.1177/001316447703700403

Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement*, *9*(2), 139–164. doi:10.1177/014662168500900204

Heidenreich, K. (1984). Grundbegriffe der Meß- und Testtheorie. In E. Roth (Eds.), *Sozialwissenschaftliche Methoden: Lehr- und Handbuch für Forschung und Praxis* (pp. 352–384). Munich, Vienna: Oldenbourg.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of Unidimensional Scales From a Multidimensional Item Bank in the Polytomous Mokken IRT Model. *Applied Psychological Measurement*, *19*(4), 337–352. doi:10.1177/014662169501900404

Hoyle, R., & Duvall, J. (2004). Determining the Number of Factors in Exploratory and Confirmatory Factor Analysis. In D. Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences* (pp. 302–317). Thousand Oaks, CA: Sage Publications.

Jacoby, W. G. (1991). *Data Theory and Dimensional Analysis*. Newbury Park, London, New Delhi: Sage Publications.

Kaplan, D. (Ed.) (2004). *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: Sage Publications.

Kopp, J., & Lois, D. (2014). *Sozialwissenschaftliche Datenanalyse: Eine Einführung* (2. überarbeitete und aktualisierte Auflage). Wiesbaden: Springer VS.

Lawrence, F. R., & Hancock, G. R. (1999). Conditions Affecting Integrity of a Factor Solution under Varying Degrees of Overextraction. *Educational and Psychological Measurement*, *59*(4), 549–579. doi:10.1177/00131649921970026

Ledesma, R. D., Valero-Mora, P., & Macbeth, G. (2015). The Scree Test and the Number of Factors: A Dynamic Graphics Approach. *The Spanish Journal of Psychology*, *18*. doi:10.1017/sjp.2015.13

Levonian, E., & Comrey, A. L. (1966). Factorial Stability as a Function of the Number of Orthogonally-Rotated Factors. *Behavioral Science*, *11*(5), 400–404. doi:10.1002/bs.3830110511

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

McIver, J. P., & Carmines, E. G. (1981). *Unidimensional Scaling.* Beverly Hills: Sage Publications.

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis: With Applications in Political Research*. The Hague: Mouton.

Müller-Schneider, T. (2001). Multiple Skalierung nach dem Kristallisationsprinzip: Eine Alternative zur explorativen Faktorenanalyse. *Zeitschrift für Soziologie, 30*(4), 305–315. doi:10.1515/zfsoz-2001-0404

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How Many Principal Components? Stopping Rules for Determining the Number of Non-Trivial Axes Revisited. *Computational Statistics & Data Analysis*, *49*(4), 974–997. doi:10.1016/j.csda.2004.06.015

Rose, J. (2014). *The Public Understanding of Political Integrity: The Case for Probity Perceptions*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Sakaluk, J. K., & Short, S. D. (2017). A Methodological Review of Exploratory Factor Analysis in Sexuality Research: Used Practices, Best Practices, and Data Analysis Resources. *Journal of Sex Research*, *54*(1), 1–9. doi:10.1080/00224499.2015.1137538

Schulze, G. (1992). *Die Erlebnisgesellschaft: Kultursoziologie der Gegenwart*. Frankfurt am Main: Campus-Verlag. [(2005). *The Experience Society.* London: Sage Publications.]

Sijtsma, K., Debets, P., & Molenaar, I. W. (1990). Mokken Scale Analysis for Polychotomous Items: Theory, a Computer Program and an Empirical Application. *Quality and Quantity*, *24*(2), 173-188 doi:10.1007/BF00209550

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to Nonparametric Item Response Theory* (Vol. 5). Thousand Oaks, CA: Sage Publications.

Stelzl, I. (1979): Ist der Modelltest des Rasch-Modells geeignet, Homogenitätshypothesen zu überprüfen? Ein Bericht über Simulationsstudien mit inhomogenen Daten. *Zeitschrift für experimentelle und angewandte Psychologie, 26*(4)*,* 652–672.

Stokman, F., & van Schuur, W. (1980). Basic Scaling. *Quality and Quantity*, *14*(1), 5–30. doi:10.1007/BF00154792

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Boston, MA: Pearson.

TenVergert, E., Gillespie, M., & Kingma, J. (1993). Testing the Assumptions and Interpreting the Results of the Rasch Model Using Log-Linear Procedures in SPSS. *Behavior Research Methods, Instruments, & Computers, 25*(3), 350–359. doi:10.3758/BF03204525

Van Abswoude, A. A. H., Vermunt, J. K., Hemker, B. T., & van der Ark, L. A. (2004). Mokken Scale Analysis Using Hierarchical Clustering Procedures. *Applied Psychological Measurement*, *28*(5), 332–354. doi:10.1177/0146621604265510

Van Schuur, W. H. (2003). Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory. *Political Analysis*, *11*(2), 139–163. doi.org/10.1093/pan/mpg002

Velicer, W. F., & Jackson, D. N. (1990). Component Analysis versus Common Factor Analysis: Some Further Observations. *Multivariate Behavioral Research*, *25*(1), 97–114. doi:10.1207/s15327906mbr2501_12

Wolff, H.-G., & Bacher, J. (2010). Hauptkomponentenanalyse und explorative Faktorenanalyse. In C. Wolf & H. Best (Eds.), *Handbuch der sozialwissenschaftlichen Da-*

tenanalyse (pp. 333–365). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-92038-2_15

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of Under- and Overextraction on Principal Axis Factor Analysis with Varimax Rotation. *Psychological Methods, 1*(4), 354–365. doi:10.1037/1082-989X.1.4.354

Zwick, W. R., & Velicer, W. F. (1986). Comparison of Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin*, *99*(3), 432–442. doi:10.1037//0033-2909.99.3.432

Appendix

Table A1 Eigenvalues of extracted components (greater than 1)

Component	Eigenvalue	Component	Eigenvalue	Component	Eigenvalue
1	15.18	11	1.77	21	1.11
2	7.88	12	1.65	22	1.08
3	7.11	13	1.50	23	1.06
4	5.12	14	1.43	24	1.05
5	3.08	15	1.35	25	1.02
6	2.52	16	1.32		
7	2.21	17	1.27		
8	2.08	18	1.22		
9	2.01	19	1.19		
10	1.88	20	1.16		

Table A2 Results of the extension procedure: extended scales 1 and 4

Scales and steps	Items	r_{itm}	$\bar{r}$	α
Scale 1				
1	Brass music ‖ Bavarian folk music	0.87	0.87	0.93
2	German folk songs	0.78	0.79	0.92
3	Popular theater (TV)	0.66	0.70	0.91
4	*Heimat* films[1]	0.69	0.67	0.91
5	German hits	0.64	0.63	0.91
6	Shows/quizzes (TV)	0.61	0.59	0.91
7	Light music	0.56	0.56	0.91
8	Local broadcasts	0.48	0.52	0.91
9	Nature broadcasts	0.46	0.49	0.90
10	Comedy movies	0.44	0.46	0.90
11	Light fiction	0.41	0.43	0.90
Scale 4				
1	Pop music ‖ Rock and pop (TV)	0.76	0.76	0.86
2	Rock music	0.78	0.74	0.90
3	Soul music	0.63	0.66	0.89
4	Reggae music	0.66	0.62	0.89
5	Going to the movies	0.58	0.58	0.89
6	Attending concerts (rock, pop, jazz)	0.60	0.55	0.90
7	Going to a discotheque	0.59	0.53	0.90
8	Folk music	0.55	0.50	0.90
9	Blues music	0.57	0.49	0.91
10	Oldies (e.g., The Beatles)	0.55	0.47	0.91
11	Going to a pub	0.48	0.45	0.91
12	Visiting a night club	0.43	0.43	0.91
13	Meeting in the city	0.43	0.41	0.91
14	Science Fiction, fantasy (TV)	0.40	0.39	0.91

[1] See footnote 1, Table 1.

Note: Minimal item–total correlation for the extension procedure = 0.4; r_{itm} = marginal item–total correlation; $\bar{r}$ = average inter-item correlation; α = Cronbach's alpha.

methods, data, analyses | Vol. 16(1), 2022, pp. 77-106 DOI: 10.12758/mda.2021.06

The Market Value of Corporate Social Performance in BRICS Countries. Differential Results Based on Panel Data Methods

Sinem Ates
Department of Business Administration, Yalova University

Abstract

Although the causal effect of social performance on financial performance is a critical issue for companies and their stakeholders, there has been no consistent econometric approach in the relevant literature to examine this relationship yet. From this point of view, the main motivation of this study is twofold: first, it aims to reveal the differential results of static and dynamic panel data methods used to estimate the impact of corporate social performance (CSP) on corporate financial performance (CFP). Second, in order to take the initiative for a consistent and reliable estimation method of the causal relationship between CSP and CFP, this study aims at drawing attention to the challenges of system generalized method of moments, which is suggested as an efficient method to solve the endogeneity problem in dynamic models. To this end, the impact of CSP on CFP for a sample of BRICS countries was analyzed through both static and dynamic panel data specifications. The main results reveal that static panel data models estimated with pooled OLS, random and fixed effects result in inconsistent and biased parameter estimates. This study discusses that although the two-step system GMM is suggested as a reliable method to deal with the endogeneity issue, some critical specifications should be considered while utilizing this method to achieve robust and efficient results.

Keywords: Corporate Social Performance, Two-Step System GMM, Static Panel Data, Dynamic Panel Data, BRICS Countries

Businesses, as open systems, interact with the environment in which they operate by utilizing resources from and producing outputs into their environment. Corporate social performance (henceforth, CSP) deals with the positive and negative outcomes of this interaction in terms of not only economic but also other dimensions such as environmental, social, and governance (Wood, 2010). Achieving a satisfying CSP in the eyes of its stakeholders may bring a business several benefits such as easy access to resources, increased employee loyalty, improved brand reputation (Haanaes et al., 2011). On the other hand, it is also argued that the investment in CSP activities means the misallocation of resources since it is not in investors' best interest (Aupperle et al., 1985). These contradicting views on the CSP activities of businesses have stimulated the researchers to investigate the impact of CSP on corporate financial performance (henceforth, CFP).

The causal link between CSP and CFP has been examined through many academic studies without a uniform conclusion. Different proxy variables used to measure CSP and CFP, diversity in sample and time frame of the studies, ignoring endogeneity are some of the factors which have been cited as the reasons behind the inconsistencies in the inferences of the researches on this issue (Brooks & Oikonomou, 2018). However, aside from the studies dealing with different samples, it is possible to obtain different results even within a single study. The main cause of this inconsistency is different methods applied to estimate the model developed to reveal the link between CSP and CFP.

Although the causal effect of social performance on financial performance is a critical issue for companies and their stakeholders, there has been no consistent econometric approach to examine this relationship yet. While most of the studies conducted static panel data methods with pooled OLS, random or fixed effects estimators (e.g. Buallay, 2019a, 2019b, 2019c; Minutolo et al., 2019; Miralles-Quirós et al., 2019; Park et al., 2018) fewer researches utilized dynamic panel data methods (Deng & Cheng, 2019; Nekhili et al., 2019).

Panel data have been widely used to derive causal inferences in social science research, however, it has been argued to confront a range of problems such as specification problems (Kittel & Winner, 2005), endogeneity especially in static panel data models (Semykina & Wooldridge, 2010), lack of robustness across different panel data models (Kittel, 2006) and, so on. When these technical issues are not handled in a reliable manner, they may affect the conclusions based on analyses with panel data (Kittel, 2008).

Leszczensky and Wolbring (2019) reviewed several panel data estimation methods in terms of their exogeneity assumptions and discussed the ways of relax-

Direct correspondence to
Dr. Sinem Ates, Department of Business Administration, Yalova University, Yalova, Turkey
E-mail: sinem.ates@yalova.edu.tr

ing exogeneity assumption which does not hold in much of social science research. The authors concluded that pooled OLS and random effects estimators will be biased if the exogeneity assumption is violated due to time-invariant unobserved heterogeneity and reverse causality between independent and dependent variables. Although unobserved heterogeneity does not constitute a problem for fixed effects and first-difference models, reverse causality remains a factor leading to biased estimates since it violates exogeneity assumption of the mentioned models. The authors demonstrate that although lagged first difference model prevents biases caused by both unobserved heterogeneity and reverse causality, it suffers from bias if the effect of independent variables on the dependent variable is fully lagged. Finally, they reviewed dynamic panel data models including the generalized method of moments (GMM) and cross-lagged panel model with fixed effects as the more reliable methods to prevent bias due to reverse causality. Based on the Monte-Carlo simulations they conducted, the authors suggested that researchers utilize a cross-lagged panel model with fixed effects in the case of reverse causality since it enables to overcome the problems caused by the misspecification of temporal lags. Like Leszczensky and Wolbring (2019), Allison et al. (2017) revealed that the cross-lagged panel model with fixed effects is less biased than the GMM model. However, they also pointed out that a cross-lagged panel model with fixed effects may be problematic in the cases of serial correlation and unbalanced panel. In this study having an unbalanced panel dataset, we tried to achieve a more reliable GMM estimation utilizing the sequential model selection process of Kripfganz (2019) and using the Stata command "xtdpdgmm" instead of "xtabond2" which has been claimed to have inaccurate aspects and some bugs (Kiviet, 2020; Kripfganz, 2019).

To our knowledge, there is a limited number of studies investigating the impact of the panel data estimation method on the inference regarding the nexus between CSP and CFP. Garcia-Castro et al. (2010) especially focused on the issue of endogeneity. Using the KLD index as the proxy for CSP and four measures of CFP, namely ROA, ROE, Tobin's Q, and MVA, the authors compared the results of pooled OLS, fixed effect, and instrumental variables (IV) estimation methods and suggested IV to deal with endogeneity. Elsayed and Paton (2005), more similar to this study, revealed the differential results of static and dynamic panel data methods applied to estimate the models investigating the impact of environmental performance on financial performance. Both studies have a sample of firms from developed countries, the US and the UK, respectively. Using a sample of 28 air carriers from various countries, Lahouel et al. (2019) emphasized the convenience of the dynamic system generalized method of moments (GMM) estimator comparing it with other estimators such as fixed effects, GLS, fixed effects instrumental variables, and two-stage least squares methods. Lin et al. (2019) compared the results

of pooled OLS, fixed effect, and system GMM while examining the relationship between CSP and CFP.

Although these studies highlight the GMM as a more efficient method to estimate the effect of CSP on CFP, the GMM estimator has its challenges which have not been addressed in the mentioned studies but can bias the results significantly unless handled correctly. None of the mentioned studies include a model selection process to find the most efficient and consistent model specification for GMM estimation. They simply add the one-year lagged dependent variable in the GMM model, however, a model selection process would result in a more efficient and consistent model specification including further lags of the dependent variable and also explanatory variables. Additionally, the classification of regressors as endogenous, predetermined, or exogeneous has not been discussed in the mentioned studies although this classification would have significant effects on the results of GMM estimation. Finally, the Stata command (xtabond2) for GMM estimations used in these studies (Lahouel et al., 2019; Lin et al., 2019) has been proven to have some bugs when dummies with factor notation are included in the model and forward orthogonal deviations are used (Kripfganz, 2019; Kiviet, 2020). To sum up, the existing studies investigating the impact of the panel data estimation method on the inference regarding the nexus between CSP and CFP have just compared the results of GMM with other estimation methods without addressing the challenges of GMM estimator which may bias the results significantly unless handled correctly. The main motivation of this study is to fill in this gap and raise awareness of these challenges for the empirical studies testing the impact of CSP on CFP and take the initiative for a consistent and reliable estimation method to be applied in the studies on this specific issue. In accordance with this motivation, this research investigates the effect of CSP on CFP for a sample of BRICS countries representing a group of emerging markets with a strong prospect of economic growth. Utilizing both static and dynamic panel data models, pooled OLS, fixed effects, random effects, and two-step system GMM methods were applied and differential results of these methods were revealed. Finally, the two-step system GMM was suggested as the most reliable method along with some critical specifications to be considered while utilizing this method.

The contribution of this study to the literature is fourfold: First, this study reveals the differential results based on the estimation method used even in the same dataset. Second, using dynamic panel data estimation methods clarifies the dynamic and long-run relationship between CSP and CFP. Third, this study clarifies the critical factors researchers should consider while applying system GMM as a dynamic panel data estimation method. Finally, having a sample of BRICS countries, this study enriches the extant literature for emerging countries.

The remaining part of the paper proceeds as follows: The next section discusses the relevant literature. While *Research Methodology* is concerned with the

research design, *Results* presents the findings of the research. A summary of the research, implications of the findings, and limitations of the study are given in *Conclusion*.

Literature Review

The literature review of this study aims to focus attention on the different estimation methods employed in academic studies investigating the link between CSP and CFP. Towards this purpose, the framework of the literate review has been determined with some limitations. The mentioned framework covers the articles indexed in the Web of Science over the last two years (2018-2019) and which used, in at least one of its research models, TOBIN'S Q and ESG SCORES&DISCLOSURE as the proxies for CSP and CFP, respectively.

Table 1, which summarizes the reviewed literature, displays the diversity of panel data estimation methods applied to estimate the link between TOBIN'S Q and ESG SCORES&DISCLOSURE. It should be noted that in some studies, ESG scores were used as a measure of sustainability performance (Aboud & Diab, 2018; Ionescu et al., 2019; Miralles-Quirós et al., 2019; Nekhili et al., 2019; Park et al., 2018) while others use ESG disclosure level as a proxy for transparency or CSR activities (Atan et al., 2018; Buallay, 2019a; 2019b; 2019c; Chauhan & Kumar, 2018; Kim et al., 2018; Li et al., 2018; Minutolo et al., 2019; Yu et al., 2018) Some studies used both types of measurements in a single study (Fatemi et al., 2018).

All the studies listed in Table 1 investigate the relationship between CSP and CFP which has been suggested to be endogenous. This endogeneity is mainly due to the fact that managers' decisions about corporate social responsibility activities just like other strategic decisions are not independent of their anticipation of the financial effect of those decisions (Garcia-Castro et al., 2010; Hamilton & Nickerson, 2003). While a solution for the endogeneity problem in the models with Tobin's Q as dependent and ESG scores/disclosures as the independent variable was not mentioned in some studies (Aboud & Diab, 2018; Atan et al., 2018; Ionescu et al., 2019; Minutolo et al., 2019; Miralles-Quirós et al., 2019; Park et al., 2018; Yu et al., 2018) on Table 1, some claimed that country-level control variables were used to deal with the endogeneity issue (Buallay, 2019a; 2019b; 2019c). More reliable estimation methods to solve the endogeneity problem such as the two-stage least squares method (Chauhan & Kumar, 2018; Fatemi et al., 2018; Li et al., 2018) and two-step GMM (Kim et al., 2018; Nekhili et al., 2019) were used in just a few studies listed on Table 1. However even in most of the studies applying more reliable methods for endogeneity, lagged dependent variable (i.e. TOBIN'S Q value of previous year) was not included as an independent variable in the research model (Fatemi et al., 2018; Li et al., 2018; Kim et al., 2018).

Table 1 Summary of Literature Review

Study	Sample Data	Sample Year	Estimation Method	Endogeneity Solution	CSP-CFP Relationship
Aboud & Diab (2018)	1.507 observations of the listed firms in the Egyptian stock market	2007-2016	OLS	Not mentioned	Positive
Atan et al. (2018)	162 observations of 54 Malaysian companies	2010-2013	OLS, fixed effects, random effects	Not mentioned	Insignificant
Buallay (2019a)	2,350 observations of 235 listed banks on the European Union countries	2007-2016	Random - effects	Usage of country-level control variables	Positive
Buallay (2019b)	3,420 observations of 342 listed financial institutions from 20 countries	2007-2016	Fixed - effects	Usage of country-level control variables	Positive (only for the model including Tobin's Q as the proxy for CFP)
Buallay (2019c)	7,248 observations of 392 manufacturing companies and 4,457 observations of 530 banks from 80 countries	2008-2017	Random - effects	Usage of country-level control variables	Positive (negative) for manufacturing (banking) sector
Chauhan & Kumar (2018)	3,837 observations of 630 Indian non-financial firms	2007-2016	OLS with industry and year fixed effects	Usage of 2SLS in robustness tests	Positive
Fatemi et al. (2018)	1,640 observations of 403 U.S. listed companies	2006-2011	Two-stage least squares (2SLS)	Usage of 2SLS	Negative (positive) for ESG disclosure & ESG concerns (ESG strengths)
Ionescu et al. (2019)	434 observations of 73 travel and leisure companies listed in S&P Global Broad Market Index Universe	2010-2015	Ordinary Least Squares (OLS)	Not mentioned	Mixed

Study	Sample Data	Sample Year	Estimation Method	Endogeneity Solution	CSP-CFP Relationship
Kim et al. (2018)	250 observations of 48 listed Korean firms	2010-2014	Generalized Method of Moments (GMM)	Usage of two-step GMM	Positive
Li et al. (2018)	2,415 observations of FTSE 350 firms in the UK	2004-2013	OLS, 2SLS, Heckman	Usage of 2SLS and Heckman Methods	Positive
Minutolo et al. (2019)	2,960 observations of 467 firms in the S&P 500	2009-2015	Fixed - effects	Not mentioned	Positive (greatest for large firms as measured by sales)
Miralles-Quirós et al. (2019)	996 observations of 166 banks from 31 countries	2010–2015	OLS	Not mentioned	Positive (negative) for environmental & governance (social) performance
Nekhili et al. (2019)	91 French firms	2007-2017	Two-step system GMM	Usage of two-step system GMM	Mixed
Park et al. (2018)	3,390 observations of firms listed in the Korea Stock Exchange	2012-2017	OLS with industry and year fixed effects	Not mentioned (for the model including Tobin's Q as the dependent variable)	Positive
Yu et al. (2018)	1996 non-financial firms from MSCI All Country World Index	2012-2016	Generalized Least Squares	Not mentioned (for the model including Tobin's Q as the dependent variable)	Positive non-linear

In the models investigating the effect of CSP on CFP, omitting the lagged dependent variable within the independent variables requires an assumption of no correlation between the current and historical values of CFP which is not well-reasoned (Garcia-Castro et al., 2010). Current financial performance, which is the dependent variable of these models, cannot be explained disregarding the feedback from the past realizations of financial performance (Lahouel et al., 2019) since strategic management decisions are highly affected by past financial performance (Garcia-Castro et al., 2010). Past financial performance was also empirically found to explain the variation in current financial performance (e.g. Capkun et al, 2009; Nguyen et al., 2014; Thrikawala, 2017). This correlation between past and present financial performance is just one of the sources of endogeneity problem existing in the static panel data models investigating the causal effect of CSP on CFP.

The reverse causality between CSP and CFP is another factor causing endogeneity bias in the models estimated with pooled OLS, random or fixed effects which are based on an exogeneity assumption (Leszczensky & Wolbring, 2019). CSP has been argued to be "*both a predictor and consequence of firm financial performance*" since it could be that companies achieving a satisfying financial performance have slack resources to invest in social responsibility activities or a better CSP leads to better financial performance due to accompanying results such as enhanced stakeholder relations or increased employee productivity (Waddock & Graves, 1997).

The other sources of endogeneity such as unobserved heterogeneity or inadequate measurements of variables are also valid for the models developed for the causal link between CFP and CSP. Recognizing the endogeneity issue for the studies on the CFP-CSP link, some researchers (Garcia-Castro et al., 2010; Lahouel et al., 2019) have started to utilize econometric models which provide more reliable estimates in the case of endogeneity. These studies showed that the positive and significant relationship between CSP and CFP turns to an insignificant relationship when estimated by a model that addresses the endogeneity issue. Although the number of studies highlighting the endogeneity issue in the research on the CSP-CFP relationship has been increasing recently, the studies emphasizing and providing guidance for the challenges of panel data methods used to solve endogeneity problems are not common. This study aims to guide researchers to handle the challenges of the GMM estimator which has been recently advised to use in the research on the CSP-CFP link (Lahouel et al., 2019).

Research Methodology

Sample

The sample of this study is based on firms from BRICS countries. The initial sample consists of firm-year observations from BRICS countries available at the Datastream database for the period 2009-2018. After eliminating the observations with missing data, the final sample covers an unbalanced panel of 3,687 firm-years. Table 2 presents the classification of this sample by both industry and country. Most of the firm-years in the final sample belong to the firms from South Africa (25.3%) and China (28.4%) and the most observed industries are financials (19.1%) and basic materials (13.7%).

Table 2 Sample Classification by Industry & Country

							Total	
Industry		Brazil	China	India	Russia	S.Africa	N	%
Basic Materials		95	98	60	69	182	504	13.7
Consumer Discretionary		106	145	64	3	107	425	11.5
Consumer Staples		82	49	79	11	99	320	8.7
Energy		34	97	43	82	10	266	7.2
Financials		96	232	159	32	186	705	19.1
Health Care		28	68	81	7	39	223	6.1
Industrials		75	169	49	0	140	433	11.7
Real Estate		48	59	32	8	91	238	6.5
Technology		8	37	44	0	34	123	3.3
Telecommunications		33	42	39	36	44	194	5.3
Utilities		110	52	48	46	0	256	6.9
Total	N	715	1,048	698	294	932	3,687	
	%	19.4	28.4	18.9	8.0	25.3		

Data and Variable Description

The dependent variable of the research models in this study was the corporate financial performance which was measured by Tobin's Q ratio. Tobin's Q, which is a market-based measure of CFP, was calculated by the following equation: (Market capitalization + Total Liabilities) / Total Assets. It is defined as "*the ratio between the market value of the firm over the reproduction cost of its assets*" (Lindenberg & Ross, 1981). CFP can also be measured by accounting-based measures such as return on assets, return on equity, return on sales, net profit. However, Tobin's Q, as a market-based measure of CFP, was preferred in this study since unlike accounting-based measures, market-based measures have the ability to capture the value of long-term investments, are less vulnerable to managerial manipulations, are not influenced by firm-specific accounting procedures, and reflect investors' expectation about companies' future economic benefits. Accounting-based measures reflect only the historical performance of companies, are subject to managerial manipulation, and depend on accounting policies adopted by the company (McGuire et al., 1988). Based on these arguments, Tobin's Q as a proxy for the market value of the company was used as the dependent variable of the research models and an accounting-based measure of CFP representing asset profitability of the company (return on assets) was included in the control variables as in many similar studies due to the fact that profitability has known to be a significant determinant of the market value of the company (Hirschey, 1982; Hsu & Jang, 2009; Kim et al., 2018; Minutolo et al., 2019; Miralles-Quirós et al., 2019; Park et al., 2018).

Corporate social performance is the independent variable of main interest in this study and was measured by companies' environmental, social, and governance pillar scores and additionally overall ESG score derived from the ASSET4 database of Datastream. Processing over 400 firm-specific ESG measures gathered from publicly available information, ESG scores measure a company's performance based on 10 main categories such as product responsibility, emissions, human rights, and so on. Among these category scores, resource use, emissions, and environmental innovation scores constitute environmental pillar score; workforce, human rights, community, and product responsibility scores are weighted with specific rates to calculate social pillar score and governance pillar score calculation is based on management, shareholders and CSR strategy scores. Overall ESG Score is a weighted average calculation of all category scores (Thomson Reuters, 2019).

Following the relevant literature, some firm-specific data were included in the regression models as control variables. "ROA" is the return on assets, directly derived from Datastream to measure the profitability of the company. The variable "SIZE" is a proxy for firm size and was calculated as the natural logarithm of assets. "LEV", which was calculated as the ratio of liabilities to assets, was deter-

mined as a proxy for financial risk. Finally, "CAPEX" represents the percentage ratio of capital expenditures to sales. Table 3 provides the descriptions and sources of all variables used.

Regression Models and Estimation Methods

Panel data, which include cross-sectional units observed at different periods, have been largely used in the researches investigating the impact of CSP on CFP or vice-versa. Panel data are known to provide several advantages over cross-sectional and time-series data such as allowing to control for unobserved characteristics of cross-sectional units, improvement in accuracy of estimations, reduction of multicollinearity problem, and so on (Baltagi, 2005; Hsiao, 1985). However, panel data have several estimation methods that may or may not be appropriate for the dataset and models handled. In this paper, to explore the effect of CSP on CFP, both static and dynamic regression models were developed and estimated with different estimation methods. In this way, differential results based on the selected regression model specification and estimation method have been revealed.

Static Panel Data Models

The static panel data regression model developed to express the CFP as a function of CSP is as follows:

$$CFP_{it} = \beta_0 + \beta_1 CSP_{it} + \beta_2 X_{it} + a_i + u_{it} \quad (1)$$

where CFP_{it} represents TOBIN'S Q; CSP_{it} is environmental (*ENV*) or social (*SOC*) or governance (*GOV*) or overall (*ESG*) score of the firm; X_{it} represents control variables (ROA, SIZE, LEV, CAPEX); a_i is the unobserved time-invariant factors affecting CFP_{it}; u_{it} is the unobserved time-varying factors affecting CFP_{it}; β_0 is the constant term; i and t stand for the firm and the time, respectively.

Using pooled OLS to estimate Equation (1) requires an assumption that the composite error term ($a_i + u_{it}$) is uncorrelated with the explanatory variables (CSP_{it} and X_{it}). This assumption holds only if the model includes all the variables simultaneously affecting CSP and CFP which is not realistic for empirical studies (Leszczensky & Wolbring, 2019). When this assumption does not hold, pooled OLS results in heterogeneity bias (also called unobserved heterogeneity) which is one of the sources of endogeneity problem (Wooldridge, 2012).

Equation (1) can also be estimated by random or fixed effects estimators. The main distinction between random and fixed effects estimators is the assumption regarding the correlation of a_i with explanatory variables. While the random

effects estimator assumes that a_i is uncorrelated with explanatory variables, the fixed effects estimator allows correlation between the a_i and explanatory variables. Unlike pooled OLS or random effects estimators, fixed effect estimator is free from bias due to time-invariant unobserved heterogeneity since a_i is allowed to be correlated with explanatory variables, that is, it captures all time-invariant unobserved heterogeneity.

However, endogeneity may be also due to reverse causality between CFP and CSP and the dynamic characteristic of CFP. Reverse causality remains as a factor leading to biased estimates in both random and fixed effects estimators due to their strict exogeneity assumption which requires the unobserved time-varying error term is uncorrelated with explanatory variables. The reverse causality between CFP and CSP violates this assumption, thereby lead to biased estimates in both models (Leszczensky & Wolbring, 2019).

Dynamic endogeneity, as another problem that should be taken into consideration in CSP-CFP models, means the existence of a correlation between past and present values of the dependent variable. If this is the case, a regression model without a lagged dependent variable among explanatory variables, just as Equation (1), would produce inconsistent parameter estimates when those lagged dependent variables are correlated with other explanatory variables. Due to the dynamic nature of economic relationships (Baltagi, 2005), a dynamic panel data model should be developed and estimated with appropriate estimation methods.

Dynamic Panel Data Models

Dynamic panel data models capture the temporal dependency of the dependent variable by the inclusion of a lagged dependent variable among explanatory variables. Expression of Equation (1) with a dynamic panel data model specification is as follows:

$$CFP_{it} = \sum_{l=1}^{p_0} \gamma_l \, CFP_{it-1} + \sum_{m=1}^{M} \sum_{l=0}^{p_m} \beta_1^{(m)} X_{it-1}^{(m)} + \sum_{s=2}^{T} T_s d_{it}^{(s)} + a_i + u_{it} \tag{2}$$

where CFP_{it} is explained by; $p0 \geq 1$ lags of CFP, $pm \geq 0$ lags of M explanatory variables $X_{it}^{(m)}$, T-1 time dummies where $d_{it}^{(s)}$ for $t = s$ and zero otherwise, random or fixed individual effects $a_{i,}$ and idiosyncratic disturbances $u_{it.}$ Equation 2 was adopted from Kiviet (2020) who formulated the model specification for GMM estimator in software programs.

The addition of lagged dependent variable among explanatory variables brings with some basic problems which cannot be solved by pooled OLS, random or fixed effects estimators. Applying pooled OLS to Equation (2) produces biased and inconsistent parameter estimates due to the fact that the lagged dependent vari-

able ($CFP_{it\text{-}n}$) is correlated with a_i. Since this correlation does not disappear as the number of observations in the dataset gets larger, pooled OLS results in biased estimates (Bond, 2002). Similarly, the random effects estimator cannot solve this correlation problem. One possible way to draw out a_i from Equation (2) is using the fixed effects estimator. However, after the within-groups transformation under the fixed effects estimator, the within transformed lagged dependent variable will be still correlated within the transformed error term when T is fixed (Baltagi, 2005; Bond, 2002).

Instrumental variables (IV) and generalized method of moments (GMM) are suggested as the most efficient methods to estimate the models with lagged dependent variables among the explanatory variables, when the time dimension of panel data is short (Kripfganz, 2019). There have been several IV and GMM estimators suggested and compared with each other since the early 1980s. (Anderson & Hsiao, 1981, 1982; Arellano, 1989; Arellano & Bond, 1991; Ahn & Schmidt 1995; Arellano & Bover 1995; Blundell & Bond, 1998…).

IV estimator developed by Anderson and Hsiao (1981) produces consistent but not efficient estimates due to not utilizing all available moment conditions (Ahn & Schmidt 1995). As a more efficient method compared to the IV estimator, the GMM estimation of Arellano and Bond (1991) transforms the data by differencing, thereby called difference GMM. Differencing means subtracting the previous observation of a variable from the current one. Instead of this transformation, Arellano and Bover (1995) introduce forward orthogonal deviations which transform the data by subtracting the average of all future available observations of a variable. This method prevents data loss caused by the differencing method in unbalanced panels. Arellano and Bover (1995) / Blundell and Bond (1998) proposed system GMM which improves efficiency by introducing more instruments than the difference GMM. System GMM uses not only lagged levels as instruments for equations in first-differences but also lagged differences as instruments for equations in levels (Roodman, 2009).

System GMM requires some assumptions to produce consistent estimates. The existence of negative first-order serial correlation and the absence of second-order serial correlation in the residuals should be satisfied for a consistent system GMM estimation. Additionally, the validity of instruments depends on the absence of correlation between the instrumental variables and error term. This exogeneity assumption of the instruments can be empirically tested by the overall overidentification and the incremental overidentification tests for each subset of instruments (Kripfganz, 2019). GMM has also some moment conditions and exclusion restrictions which cannot be tested. GMM estimation of a model including some endogenous regressors requires some exclusion restrictions on the model since these endogenous regressors cannot be used as instrumental variables because of their correlation with the error term. However, the number of instrumental vari-

ables should be higher than or at least equal to the number of regressors in the model. Based on this requirement of GMM, some lagged variables cannot be kept in the model since they are used as instruments. The resulting exclusion of regressors from the model constitutes an exclusion restriction on the model which cannot be tested (Kiviet, 2020). The moment conditions based on the classification of the variables in the model are as follows (Kiviet, 2020; Kripfganz, 2019):

$E(y_{is}\Delta\varepsilon_{it}) = 0$ *for* $s \leq t-2$,

$E(x_{is}^{m}\Delta\varepsilon_{it}) = 0$ *for* $s \leq t-2$ *if* x_{it}^{m} *is endogenous,*

$E(x_{is}^{m}\Delta\varepsilon_{it}) = 0$ *for* $s \leq t-1$ *if* x_{it}^{m} *is predetermined,*

$E(x_{is}^{m}\Delta\varepsilon_{it}) = 0$ *for* $\forall s$ *if* x_{it}^{m} *is exogenous*

Taking into consideration its efficiency in unbalanced panels, system GMM was used to estimate Equation (2) in this study. However, the GMM estimator should be applied rigorously because it has some challenges which may cause biased results unless handled correctly. It is not advised for panels having a long time dimension. In the cases of many instruments, the results of GMM may be biased. Since the GMM estimator is complicated, it may produce biased estimates due to the wrong use by researchers. (Roodman, 2009).

The commands used in statistical software programs to apply the GMM estimator should be clearly understood by the user to be able to find the best reliable specification. In this study, the Stata command "xtdpdgmm" was used for the GMM estimation of Equation (2). Kripfganz (2019) has introduced "xtdpdgmm" command by asserting that "xtabond2", which is another Stata command for GMM estimation, has some bugs when dummies with factor notation are included in the model and forward orthogonal deviations are used. In a recent paper, Kiviet (2020) discussed all the inaccurate aspects of "xtabond2" in detail and cited "xtdpdgmm" as a "*promising improved alternative*".

A model specification search, which was suggested by Kiviet (2020) and Kripfganz (2019), has been conducted to find the most efficient and consistent model specification for the estimation of Equation (2). The followed process of model specification search was explained through the subsection of "Results of Dynamic Panel Data Model" in depth.

Results

Descriptives

Table 4 provides mean, standard deviation (S.D.), minimum and maximum values of variables used in regression models in this study. All the financial variables were winsorized at the bottom and top 5% to mitigate the effect of outliers.

Table 4 Descriptive Statistics

	Mean	S.D.	Min	Max
TOBIN'S Q	1.66	1.05	.75	4.72
ROA	7.30	5.98	-.84	21.56
SIZE	15.58	1.63	12.85	18.97
LEV	.58	.21	.21	.93
CAPEX	9.84	11.26	.40	42.67
ESG	50.12	16.68	7.77	95.43
ENV	49.25	21.38	4.56	98.38
SOC	50.47	21.59	4.73	98.54
GOV	50.69	20.54	2.28	98.37

Notes: All financial variables (*TOBIN'S Q, ROA, SIZE, LEV*) are winsorized at 5%. Variables are defined in Table 3.

Pairwise correlations between the variables of regression models are presented in Table 5. The variables ESG, ENV, SOC, and GOV were not included in the same regression model. Except for these variables, all the correlation coefficients in Table 5 are below 80% which means that there is no multicollinearity problem in the models of this study. Calculated variance inflation factors of these variables also confirm that multicollinearity is within acceptable limits.

Table 3 Variables Definition

	Description	Source
Dependent Variables		
TOBIN'S Q	the ratio of (market capitalization + total liabilities) to total assets	Datastream
Control Variables		
ROA	return on assets	Datastream
SIZE	the logarithm of total assets	Datastream
LEV	the ratio of liabilities to assets	Datastream
CAPEX	capital expenditure % sales	Datastream
Independent Variables		
ESG	overall ESG score	Datastream
ENV	environmental pillar score	Datastream
SOC	social pillar score	Datastream
GOV	governance pillar score	Datastream

Table 5 Pairwise Correlations

Variable	TOBIN'S Q	ROA	SIZE	LEV	CAPEX	ESG	ENV	SOC	GOV
TOBIN'S Q	1								
ROA	.7037***	1							
SIZE	-.4142***	-.4255***	1						
LEV	-.3089***	-.4817***	.4790***	1					
CAPEX	-.1324***	-.0311*	.0069	-.2001***	1				
ESG	-.0141	-.0026	.1821***	.0881***	-.0439***	1			
ENV	-.0442***	-.0269	.2487***	.0751***	-.0231	.8469***	1		
SOC	.0041	.0478***	.0709***	.0611***	-.0347**	.8569***	.6633***	1	
GOV	.0088	-.0342**	.1096***	.0727***	-.0478***	.6314***	.2830***	.2883***	1

Notes: All financial variables (*TOBIN'S Q, ROA, SIZE, LEV*) are winsorized at 5% level. Variables are defined in Table 3. *, **, *** stand for significance levels of <.10, <.05, <.01, respectively.

Regression Results

Results of Static Panel Data Model

Table 6 provides the pooled OLS, random, and fixed effects estimation results of the static panel data model expressed with Equation (1). In order to choose the most consistent and efficient estimator between pooled OLS, random, and fixed effects estimators, we carried out Breusch-Pagan LM and Hausman tests, respectively. The significant p-value of the test statistic of the Breusch-Pagan LM test indicates that random individual effects are significant, and their variances are not "0" (Baltagi, 2005). This means that the estimation of Equation (1) with the pooled OLS estimator results in biased estimates. As a second step, we employed the robust Hausman test to decide between random and fixed-effects estimators. The null hypothesis under the Hausman test, which is also an assumption of random effects, is that unobserved effect a_i is not correlated with explanatory variables. The rejection of the robust Hausman test due to the significant test statistic means that the assumption of random effects estimator is violated, therefore fixed effects estimator should be preferred.

Fixed effects estimation results in Table 6 indicate that environmental, social, and overall EGS performance of the companies have a small but positive impact on the corporate financial performance which was proxied by Tobin's Q ratio. Among the control variables, ROA was also found to be positively correlated with Tobin's Q ratio. SIZE has the biggest significant effect on Tobin's Q with a negative sign. In line with these findings, the firms with higher environmental and social performances, higher profitability, and smaller size can be said to have a higher market value.

However, for the fixed effects estimator to be consistent, the explanatory variables should be strictly exogenous. The exogeneity of the explanatory variables in Equation (1) was tested by the Wooldridge test for strict exogeneity. This test is based on the comparison of the models below:

$$CFP_{it} = \beta_0 + \beta_1 CSP_{it} + \beta_2 X_{it} + a_i + u_{it} \quad (1)$$

$$CFP_{it} = \beta_0 + \beta_1 CSP_{it} + \beta_2 X_{it} + \beta_1 CSP_{i(t+1)} + \beta_2 X_{i(t+1)} + a_i + u_{it} \quad (3)$$

The first model is the standard model which was estimated by fixed effects. In addition to the variables in the first model, the second model also includes the future values of all explanatory variables. The main idea behind the Wooldridge test for strict exogeneity is to test whether the future values in the second model are significant or not.

Table 6 Pooled OLS & Random Effects & Fixed Effects Estimation Results of Static Model – Equation 1

	Pooled OLS				RE				FE			
ROA	.102***	.102***	.103***	.102***	.058***	.058***	.058***	.058***	.045***	.045***	.045***	.046***
	(.003)	(.003)	(.003)	(.003)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)
SIZE	-.052***	-.054***	-.052***	-.054***	-.147***	-.147***	-.145***	-.146***	-.233***	-.235***	-.229***	-.223***
	(.009)	(.009)	(.009)	(.009)	(.023)	(.022)	(.022)	(.022)	(.049)	(.048)	(.048)	(.047)
LEV	.334***	.330***	.348***	.328***	.113	.116	.113	.112	.276	.285	.268	.258
	(.073)	(.074)	(.074)	(.073)	(.143)	(.143)	(.143)	(.143)	(.178)	(.178)	(.178)	(.177)
CAPEX	-.000	.000	-.000	.000	.001	.001	.001	.001	.002	.002	.002	.002
	(.001)	(.001)	(.001)	(.001)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)
ESG	**-.000**				**.000**				**.003****			
	(.001)				**(.001)**				**(.001)**			
ENV		**.000**				**.000**				**.002****		
		(.001)				**(.001)**				**(.001)**		
SOC			**-.001****				**.000**				**.002***	
			(.001)				**(.001)**				**(.001)**	
GOV				**.000**				**.000**				**.001**
				(.001)				**(.001)**				**(.001)**
Constant	2.342***	2.344***	2.375***	2.326***	4.357***	4.370***	4.357***	4.351***	4.798***	4.874***	4.785***	4.736***
	(.184)	(.181)	(.185)	(.183)	(.379)	(.380)	(.378)	(.377)	(.720)	(.725)	(.724)	(.713)
YEAR	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
IND	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

	Pooled OLS				RE				FE			
N	3,687	3,687	3,687	3,687	3,687	3,687	3,687	3,687	3,687	3,687	3,687	3,687
R2	.626	.626	.626	.626	.578	.578	.578	.578	.401	.400	.400	.406
B&P LM					.000	.000	.000					
R.Hausman									.000	.000	.000	.000

Notes: OLS, RE, and FE represent ordinary least squares, random effects, and fixed effects estimators, respectively. Standard errors which are robust to heteroscedasticity and autocorrelation are in parenthesis. All models include time (YEAR) and industry (IND) dummy variables. N denotes the number of observations. R2: square of overall correlation. B&P LM is the p-value of the test statistic of the Breusch-Pagan LM test for random effects. R. Hausman is the p-value of the test statistic of the Cluster-Robust Hausman Test. Variables are defined in Table 3. *, **, *** stand for significance levels of <.10, <.05, <.01, respectively.

After estimating the two models by fixed effect estimators and robust standard errors, the F test for joint significance of future values of explanatory variables resulted in a significant F statistic (56.99, $p<0.01$). This means that the future values of explanatory variables are correlated with the error term, thereby violates the strict exogeneity assumption of fixed effects. Therefore, we can argue that the parameter estimates in Table 6 are inconsistent and biased.

Results of Dynamic Panel Data Model

Kiviet (2020) and Kripfganz (2019) suggested a model specification search as the first step to obtaining consistent, efficient, and accurate parameter estimates as the result of the GMM estimator. After including all relevant regressors to the model based on the economic theory, the model specification process requires the classification of regressors as endogenous, predetermined, or exogeneous. A variable is strictly exogenous if it is uncorrelated with the time-varying error term at all leads and lags. On the contrary, endogenous variables are correlated with the time-varying error term at all leads and lags. Finally, predetermined variables are uncorrelated with present and future values of time-varying error term but potentially correlated with historical values (Arellano, 2003).

This study tries to follow the steps of the "sequential model selection process" of Kripfganz (2019) who adapted it from Kiviet (2019) with some revisions. Kiviet (2020) suggested treating all unlagged explanatory variables as endogenous unless proven otherwise. The first step of the model selection process is specifying an initial candidate "maintained statistical model (MSM)" including all relevant explanatory variables with sufficient lags. This initial MSM with collapsed and/or curtailed instruments for forward orthogonal deviations transformation, should include time dummies and assume all regressors as endogenous. The second step tells to estimate the initial MSM by two-step GMM estimator with Windmeijer standard errors robust to finite-sample bias.

Following the instructions in the first and second steps, an initial candidate MSM based on Equation (2) was developed. This initial model included 3 lags for all variables assuming all the unlagged regressors as endogenous. In order to prevent the biases caused by too many instruments, this initial model included the collapse option which is one of the approaches to reduce the number of instruments. Finally, since the forward orthogonal deviations (FOD) transformation minimizes data loss in unbalanced panels, the initial candidate MSM was specified as a FOD-transformed model (Kripfganz, 2019). Then two-step GMM estimator with Windmeijer standard errors robust to finite-sample bias was used to estimate this initial candidate MSM. The two-step GMM estimator is more efficient than the one-step GMM estimator when the time-varying error term is heteroskedastic and Windmei-

jer-corrected standard errors are used to correct the finite-sample bias of two-step standard errors.

After developing this initial candidate MSM, 28 more candidates were developed by; a) removing lagged variables with the highest p-value, b) treating explanatory variables that were classified as endogenous in the initial model as predetermined one by one, and c) adding industry dummies. The consistency of all these candidate models was checked by the specification tests used after the GMM estimation. More precisely, the Arellano-Bond test was used to check the autocorrelation of the first-differenced residuals. The existence of negative first-order serial correlation and the absence of second-order serial correlation was confirmed for all the candidate models. To test the overall validity of instruments, the Hansen overidentification test was utilized and finally, the incremental overidentification test was carried out to check the validity of each subset of instruments. Specification test results were satisfactory for all candidate models. As suggested by Kripfganz (2019), the model and moment selection criteria (MMSC) of Andrews and Lu (2001) was utilized to decide the most efficient one among the candidate models. The models with the lowest values of Akaike (AIC), Bayesian (BIC), and Hannan-Quinn (HQIC) criteria were selected and reported in Table 7.

The model specification with the lowest values of MMSC-AIC, MMSC-BIC, and MMSC-HQIC criteria was the one including TOBIN'S Q variables lagged by one, two, and three periods, and also time and industry dummies. This model treated the variables SIZE, LEV, and CAPEX as predetermined, but ROA as endogenous. It should be noted that the models treating ROA as predetermined could not pass the specification tests. This model was estimated by the two-step system GMM estimators with collapsed instruments and Windmeijer standard errors robust to finite-sample bias for the FOD-transformed model. Table 7 provides the parameter estimates of this model specification with overall ESG, ENV, SOC, and GOV as the main interest of variables, respectively.

The fixed effects results in Table 6 and system GMM results in Table 7 differ considerably with regards to the relationship between CSP and CFP. More precisely, whereas fixed effect results reveal that environmental, social, and overall EGS performance have a significant positive effect on the CFP, two-step system GMM results reveal the opposite, i.e. a significant negative impact. The insignificant relationship between governance performance and CFP is valid in both fixed effects and system GMM estimations. When it comes to control variables, whereas SIZE has a negative and significant coefficient estimate in fixed-effects results, the coefficient estimate of SIZE is not significant for all the models estimated with system GMM. Additionally, based on the fixed effects results it is possible to say that there is not a significant relationship between CAPEX and TOBIN'S Q. However, according to system GMM results, CAPEX has a significant negative effect on TOBIN'S Q except for the SOC model.

The negative causal effect of CSP on CFP can be explained by the trade-off hypothesis of Preston and O'Bannon (1997). The trade-off hypothesis, which is based on Friedman's (1970) argument indicating that "*the social responsibility of business is to increase its profits",* claims that social responsibility activities such as environmental protection, charity work consume company resources in a way that is not for the best interest of investors. Accordingly, the companies which are bearing financial costs due to their social responsibility activities fall into a disadvantaged position in comparison to their counterparts which use less or no resources for these types of activities. Ultimately, higher levels of CSP can lead to lower levels of financial performance. It is highly probable that this hypothesis is valid for a sample of developing countries as analyzed in this study since it is not an unexpected case that awareness of social responsibility activities in developing countries is less than that of developed countries.

As seen in Table 7, the first lag of TOBIN'S Q has the biggest coefficient estimate which means that the current value of TOBIN'S Q is highly dependent on the lagged value of it. Omitting this variable will result in biased parameter estimates for the other variables in the regression model. Equation (1), as a static model, does not incorporate this temporal dependency of TOBIN'S Q, thereby produces biased and inconsistent parameter estimates even it is estimated with the fixed effects estimator.

In order to verify the robustness of the system GMM results reported in Table 7, financial firms were excluded from the sample, and Equation (2) was re-estimated. The coefficient estimates of the main interest variables (ESG, ENV, SOC, GOV) were quantitatively similar to the reported parameter estimates in Table 7.

GMM results reported in Table 7 are based on a lower number of observations (1,966) than the original number of observations (3,687) in the sample because of the lagged dependent variables in the dynamic model. In order to see if the different results between FE and GMM are purely based on the omission of the dynamic terms in FE, FE results for the GMM subset of observations were also provided in the Appendix. When the results reported in the Appendix are compared with the GMM results in Table 7, it is seen that while FE estimations of the models result in positive and insignificant coefficients for ESG, ENV, and SOC variables, GMM estimation produces negative and significant coefficients for the same variables. Accordingly, we can conclude that the different results between FE and GMM are not based on the lower number of observations in GMM estimation but the omission of the dynamic terms in FE.

Table 7 Two-Step System GMM Estimation Results of Dynamic Model – Equation 2

	ESG MODEL	ENV MODEL	SOC MODEL	GOV MODEL
L1.TOBIN'S Q	.762*** (.087)	.745*** (.089)	.742*** (.087)	.756*** (.084)
L2.TOBIN'S Q	-.067 (.054)	-.055 (.055)	-.081 (.057)	-.048 (.053)
L3.TOBIN'S Q	-.048 (.040)	-.046 (.039)	-.044 (.040)	-.063 (.040)
ROA	.011 (.008)	.010 (.007)	.014* (.007)	.011 (.007)
SIZE	.019 (.028)	.021 (.030)	-.006 (.025)	.001 (.026)
LEV	-.009 (.381)	.081 (.331)	.185 (.374)	-.051 (.348)
CAPEX	-.006* (.003)	-.006* (.003)	-.006 (.004)	-.008** (.004)
ESG	**-.006*** **(.002)**			
ENV		**-.003**** **(.001)**		
SOC			**-.003**** **(.002)**	
GOV				**-.002** **(.001)**
Constant	.853 (.532)	.647 (.560)	1.080** (.520)	.845 (.519)
YEAR	YES	YES	YES	YES
IND	YES	YES	YES	YES
N	1,966	1,966	1,966	1,966
AR2	.812	.8382	.716	.997
Hansen	.621	.5593	.477	.386
Inc. Hansen (p values)	all>.10	all>.10	all>.10	all>.10

Notes: This table represents the parameter estimates of the two-step GMM estimation of Equation (2) with time (YEAR) and industry (IND) dummies, collapsed instruments, and Windmeijer-corrected standard errors for the FOD-transformed model treating all the unlagged explanatory variables as predetermined except ROA which is assumed to be endogenous. L1 & L2 & L3. TOBIN'S Q stand for TOBIN'S Q variables lagged by one, two, and three periods, respectively. Windmeijer-corrected standard errors are presented in parenthesis. N denotes the number of observations. AR2 is the p value of the test statistic of the Arellano-Bond test for second-order serial correlation. Hansen is the p value of the test

statistic of the Hansen overidentification test. Inc. Hansen represents the p values of test statistics of incremental overidentification tests. Variables are defined in Table 3. *, **, *** stand for significance levels of <.10, <.05, <.01, respectively.

Conclusion

Using a sample including 3,687 observations of listed firms in BRICS countries for the period 2009-2018, this study examined the impact of CSP on CFP utilizing both static and dynamic panel data models and also various estimators including pooled OLS, random & fixed effects, and system GMM. The main motivation behind the empirical analyses of this study was to expose the inconsistent results between the static and dynamic panel data models. It was also aimed to draw attention to the challenges of the two-step system GMM which may result in biased parameter estimates unless taken into account properly.

The results of static and dynamic panel data specifications and estimations differ considerably on the main conclusion regarding the effect of CSP on CFP. Whereas the static model specification estimated with fixed effects indicates a positive and significant relationship between CSP (except for governance performance) and CFP, the results of dynamic panel data specification estimated by system GMM suggests the opposite. More precisely, there is a negative and significant relationship between CSP (except for governance performance) and CFP according to the dynamic panel data analyses. This inconsistency between the results of static and dynamic panel data analyses mainly stems from the fact that static panel data models miss the temporal dependency of the dependent variable. Accordingly, dynamic endogeneity remains a problem and result in biased parameter estimates under static panel data specifications.

The findings of this research should prompt the researchers to test the robustness of the results of static panel data analyses as it reveals the insufficiency of static panel data models while examining the nexus between CSP and CFP. However, this study also wants to draw attention to the challenges of system GMM as a dynamic panel data estimation method. System GMM is suggested as a more efficient estimator under dynamic endogeneity, however, researchers should apply system GMM rigorously to handle its challenges properly. Otherwise, system GMM may lead to wrong inferences just as static panel data methods.

This study has also some crucial findings for the authorities of capital markets and listed companies in BRICS countries. The finding indicating a negative impact of CSP on CFP should prompt capital markets to develop policies to increase the market value of corporate social responsibility activities of companies by raising awareness of the listed companies and their investors on the significance of sustainable development.

The limitations of this study may open the way to new ideas for further research. This study utilized only a market-based performance measure, further research should consider also accounting-based performance measures as a proxy for CFP. Governance indicators such as board composition, board size can be included in the models to mitigate the effect of omitted variable bias on the results.

References

Aboud, A., & Diab, A. (2018). The impact of social, environmental and corporate governance disclosures on firm value. *Journal of Accounting in Emerging Economies.* https://doi.org/10.1108/JAEE-08-2017-0079.

Ahn, S. C., & Schmidt, P. (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics*, *68*(1), 5-28.

Anderson, T. W., & Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association*, *76*(375), 598-606.

Anderson, T. W., & Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, *18*(1), 47-82.

Andrews, D. W., & Lu, B. (2001). Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics*, *101*(1), 123-164.

Arellano, M. (1989). A note on the Anderson-Hsiao estimator for panel data. *Economics Letters*, *31*(4), 337-341.

Arellano, M. (2003). *Panel Data Econometrics.* Oxford University Press.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, *58*(2), 277-297.

Arellano, M., & Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, *68*(1), 29-51.

Atan, R., Alam, M. M., Said, J., & Zamri, M. (2018). The impacts of environmental, social, and governance factors on firm performance. *Management of Environmental Quality: An International Journal*. https://doi.org/10.1108/MEQ-03-2017-0033.

Aupperle, K. E., Carroll, A. B., & Hatfield, J. D. (1985). An empirical examination of the relationship between corporate social responsibility and profitability. *Academy of Management Journal, 28*(2), 446-463.

Baum, C. F., Schaffer, M. E., & Stillman, S. (2003). Instrumental variables and GMM: Estimation and testing. *The Stata Journal, 3*(1), 1-31.

Baltagi, B. H. (2005). *Econometric analysis of data panel.* England: John Wiley & Sons Ltd.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, *87*(1), 115-143.

Brooks, C., Oikonomou, I. (2018). The effects of environmental, social and governance disclosures and performance on firm value: a review of the literature in accounting and finance. *British Accounting Review, 50*(1), 1-15. https://doi.org/10.1016/j.bar.2017.11.005.

Buallay, A. (2019a). Is sustainability reporting (ESG) associated with performance? Evidence from the European banking sector. *Management of Environmental Quality: An International Journal*. https://doi.org/10.1108/MEQ-12-2017-0149.

Buallay, A. (2019b). Between cost and value: investigating the effects of sustainability reporting on a firm's performance. *Journal of Applied Accounting Research*, *20*(4), 481-496. https://doi.org/10.1108/JAAR-12-2017-0137.

Buallay, A. (2019c). Sustainability reporting and firm's performance. *International Journal of Productivity and Performance Management*. https://doi.org/10.1108/IJPPM-10-2018-0371.

Capkun, V., Hameri, A. P., & Weiss, L. A. (2009). On the relationship between inventory and financial performance in manufacturing companies. *International Journal of Operations & Production Management, 29*(8), 789-806. http://dx.doi.org/10.1108/02635570210423235.

Chauhan, Y., & Kumar, S. B. (2018). Do investors value the nonfinancial disclosure in emerging markets?. *Emerging Markets Review*, *37*, 32-46. https://doi.org/10.1016/j.ememar.2018.05.001.

Deng, X., & Cheng, X. (2019). Can ESG Indices Improve the Enterprises' Stock Market Performance? - An Empirical Study from China. *Sustainability*, *11*(17), 4765. https://doi.org/10.3390/su11174765.

Elsayed, K., & Paton, D. (2005). The impact of environmental performance on firm performance: static and dynamic panel data evidence. *Structural Change and Economic Dynamics*, *16*(3), 395-412. https://doi.org/10.1016/j.strueco.2004.04.004.

Fatemi, A., Glaum, M., & Kaiser, S. (2018). ESG performance and firm value: The moderating role of disclosure. *Global Finance Journal*, *38*, 45-64. http://dx.doi.org/10.1016/j.gfj.2017.03.001.

Friedman, M. (1970). A Friedman doctrine: The social responsibility of business is to increase its profits. *The New York Times Magazine, 13*(1970), 32-33.

Garcia-Castro, R., Ariño, M. A., & Canela, M. A. (2010). Does social performance really lead to financial performance? Accounting for endogeneity. *Journal of Business Ethics*, *92*(1), 107-126. https://doi.org/10.1007/s10551-009-0143-8.

Haanaes, K., Balagopal, B., Arthur, D., Kong, M. T., Velken, I., Kruschwitz, N., & Hopkins, M. S. (2011). First look: The second annual sustainability & innovation survey. *Sloan Management Review, 52*(2), 76-84.

Hamilton, B. H., & Nickerson, J. A. (2003). Correcting for endogeneity in strategic management research. *Strategic Organization*, *1*(1), 51-78.

Hirschey, M. (1982). Intangible capital aspects of advertising and R & D expenditures. *The Journal of Industrial Economics*, *30*(4), 375-390.

Hsiao, C. (1985). Benefits and limitations of panel data. *Econometric Reviews*, *4*(1), 121-174. https://doi.org/10.1080/07474938508800078.

Hsu, L. T. J., & Jang, S. S. (2009). Effects of restaurant franchising: does an optimal franchise proportion exist?. International Journal of Hospitality Management, *28*(2), 204-211. https://doi.org/10.1016/j.ijhm.2008.07.002.

Ionescu, G. H., Firoiu, D., Pirvu, R., & Vilag, R. D. (2019). The impact of ESG factors on market value of companies from travel and tourism industry. *Technological and Economic Development of Economy*, *25*(5), 820-849. https://doi.org/10.3846/tede.2019.10294.

Kim, W. S., Park, K., & Lee, S. H. (2018). Corporate social responsibility, ownership structure, and firm value: Evidence from Korea. *Sustainability*, *10*(7), 2497. https://doi:10.3390/su10072497.

Kittel, B. (2006). A crazy methodology? On the limits of macro-quantitative social science research. *International Sociology, 21*(5), 647-677.

Kittel, B. (2008). Statistical narratives and the properties of macro-level variables: labor market institutions and employment performance in macrocomparative research. In L. Kenworthy & A. Hicks (Eds.), *Method and Substance in Macrocomparative Analysis* (pp. 29-66). London: Palgrave Macmillan.

Kittel, B., & Winner, H. (2005). How reliable is pooled analysis in political economy? The globalization-welfare state nexus revisited. *European Journal of Political Research, 44*(2), 269-293. https://doi.org/10.1177/0268580906067835.

Kiviet, J. F. (2019). Microeconometric dynamic panel data methods: Model specification and selection issues. MPRA Paper 95159, Munich Personal RePEc Archive

Kiviet, J. F. (2020). Microeconometric dynamic panel data methods: Model specification and selection issues. *Econometrics and Statistics*, *13*, 16-45. https://doi.org/10.1016/j.ecosta.2019.08.003.

Kripfganz S. (2019). Generalized method of moments estimation of linear dynamic panel-data models. London Stata Conference 2019 (No. 17), Stata Users Group.

Lahouel, B. B., Gaies, B., Zaied, Y. B., & Jahmane, A. (2019). Accounting for endogeneity and the dynamics of corporate social–corporate financial performance relationship. *Journal of Cleaner Production, 230*, 352-364. https://doi.org/10.1016/j.jclepro.2019.04.377.

Leszczensky, L., & Wolbring, T. (2019). How to deal with reverse causality using panel data? Recommendations for researchers based on a simulation study. *Sociological Methods & Research*, 1-29. https://doi.org/10.1177/0049124119882473.

Li, Y., Gong, M., Zhang, X. Y., & Koh, L. (2018). The impact of environmental, social, and governance disclosure on firm value: The role of CEO power. *The British Accounting Review*, *50*(1), 60-75. https://doi.org/10.1016/j.bar.2017.09.007.

Lin, W. L., Ho, J. A., & Sambasivan, M. (2019). Impact of corporate political activity on the relationship between corporate social responsibility and financial performance: A dynamic panel data approach. *Sustainability, 11*(1), 60. https://doi.org/10.3390/su11010060.

Lindenberg, E. B., & Ross, S. A. (1981). Tobin's q ratio and industrial organization. *Journal of Business*, *54*(1), 1-32.

McGuire, J. B., Sundgren, A., & Schneeweis, T. (1988). Corporate social responsibility and firm financial performance. *Academy of Management Journal, 31*(4), 854-872.

Minutolo, M. C., Kristjanpoller, W. D., & Stakeley, J. (2019). Exploring environmental, social, and governance disclosure effects on the S&P 500 financial performance. *Business Strategy and the Environment*, *28*(6), 1083-1095. https://doi.org/10.1002/bse.2303.

Miralles-Quirós, M. M., Miralles-Quirós, J. L., & Redondo Hernández, J. (2019). ESG Performance and shareholder value creation in the banking industry: International differences. *Sustainability*, *11*(5), 1404. https://doi:10.3390/su11051404.

Nekhili, M., Boukadhaba, A., Nagati, H., & Chtioui, T. (2019). ESG performance and market value: the moderating role of employee board representation. *The International Journal of Human Resource Management*, 1-27. https://doi.org/10.1080/09585192.2019.1629989.

Nguyen, T., Locke, S., & Reddy, K. (2014). A dynamic estimation of governance structures and financial performance for Singaporean companies. *Economic Modelling, 40*, 1-11. https://doi.org/10.1016/j.econmod.2014.03.013.

Park, J. H., Park, H. Y., & Lee, H. Y. (2018). The effect of social ties between outside and inside directors on the association between corporate social responsibility and firm value. *Sustainability*, *10*(11), 3840. https://doi.org/10.3390/su10113840.

Preston, L. E., & O'bannon, D. P. (1997). The corporate social-financial performance relationship: A typology and analysis. *Business & Society, 36*(4), 419-429.

Roodman, D. (2009). How to do xtabond2: An introduction to difference and system GMM in Stata. *The Stata Journal*, *9*(1), 86-136.

Semykina, A., & Wooldridge, J. M. (2010). Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics, 157*(2), 375-380. https://doi.org/10.1016/j.jeconom.2010.03.039.

Thomson Reuters (2019). Thomson Reuters ESG Scores. https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/esg-scores-methodology.pdf (accessed on 18.03.2020).

Thrikawala, S., Locke, S., & Reddy, K. (2017). Dynamic endogeneity and corporate governance-performance relationship. *Journal of Economic Studies, 44*(5), 727-744. https://doi.org/10.1108/JES-12-2015-0220.

Waddock, S. A., & Graves, S. B. (1997). The corporate social performance–financial performance link. *Strategic Management Journal, 18*(4), 303-319.

Wood, D. J. (2010). Measuring corporate social performance: A review. *International Journal of Management Reviews, 12*(1), 50-84. https://doi.org/10.1111/j.1468-2370.2009.00274.x.

Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach* (5th ed.). Cengage Learning.

Yu, E. P. Y., Guo, C. Q., & Luu, B. V. (2018). Environmental, social and governance transparency and firm value. *Business Strategy and the Environment*, *27*(7), 987-1004. https://doi.org/10.1002/bse.2047.

Appendix

	ESG MODEL	ENV MODEL	SOC MODEL	GOV MODEL
ROA	0.041*** (0.005)	0.041*** (0.005)	0.041*** (0.005)	0.041*** (0.005)
SIZE	-0.218*** (0.061)	-0.225*** (0.061)	-0.219*** (0.061)	-0.213*** (0.059)
LEV	0.264 (0.209)	0.273 (0.209)	0.263 (0.207)	0.261 (0.207)
CAPEX	-0.002 (0.002)	-0.002 (0.002)	-0.002 (0.002)	-0.002 (0.002)
ESG	**0.001** **(0.001)**			
ENV		**0.001** **(0.001)**		
SOC			**0.001** **(0.001)**	
GOV				**-0.001** **(0.001)**
Constant	4.546*** (0.912)	4.630*** (0.918)	4.558*** (0.912)	4.568*** (0.900)
YEAR	YES	YES	YES	YES
IND	YES	YES	YES	YES
N	1,966	1,966	1,966	1,966
R2	0.388	0.386	0.387	0.381

Notes: This table represents the parameter estimates of fixed effect estimation of Equation (1) for GMM set of observations. Standard errors which are robust to heteroscedasticity and autocorrelation are in parenthesis. All models include time (YEAR) and industry (IND) dummy variables. N denotes for the number of observations. R2: square of overall correlation. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

methods, data, analyses | Vol. 16(1), 2022, pp. 107-128 DOI: 10.12758/mda.2021.10

Testing the Effects of Automated Navigation in a General Population Web Survey

Jeldrik Bakker[1,2], *Marieke Haan*[3], *Barry Schouten*[1,2], *Bella Struminskaya*[2], *Peter Lugtig*[2], *Vera Toepoel*[2], *Deirdre Giesen*[1] & *Vivian Meertens*[1]

[1] *Statistics Netherlands*

[2] *Utrecht University*

[3] *University of Groningen*

Abstract

This study investigates how an auto-forward design, where respondents navigate through a web survey automatically, affects response times and navigation behavior in a long mixed-device web survey. We embedded an experiment in a health survey administered to the general population in The Netherlands to test the auto-forward design against a manual-forward design. Analyses are based on detailed paradata that keep track of the respondents' behavior in navigating the survey. We find that an auto-forward design decreases completion times and that questions on pages with automated navigation are answered significantly faster compared to questions on pages with manual navigation. However, we also find that respondents use the navigation buttons more in the auto-forward condition compared to the manual-forward condition, largely canceling out the reduction in survey duration. Furthermore, we also find that the answer options 'I don't know' and 'I rather not say' are used just as often in the auto-forward condition as in the manual-forward condition, indicating no differences in satisficing behavior. We conclude that auto-forwarding can be used to reduce completing times, but we also advice to carefully consider mixing manual and auto-forwarding within a survey.

Keywords: mixed-device surveys, web surveys, auto-forward, paradata, usability

Web surveys are completed on a range of different devices: PCs, laptops, tablets, and smartphones. Since mobile devices vary in screen size and type of navigation, surveys designed for PCs and laptops tend to be more difficult to navigate on mobile devices. Survey designers have recognized this challenge and have adapted to the smaller screens and different mode of data entry used on smartphones. Nonetheless, even when surveys are "mobile-friendly", web surveys still take longer on smartphones compared to tablets and PCs (Couper, Antoun, & Mavletova, 2017; Couper & Peterson, 2017). Survey duration is an important factor to take into account, because it is a proxy for respondent burden (Zhang & Conrad, 2014). It is conjectured that the maximal duration of a survey that a respondent is willing to complete depends on the type of the device: respondents are less willing to complete longer surveys on smartphones (Hintze, Findling, Scholz, & Mayrhofer, 2014). Therefore, not accounting for survey duration when designing surveys for mixed-mode surveys can result in coverage errors, higher nonresponse, and lower data quality (Cook, 2014; Wells, Bailey, & Link, 2014; Struminskaya, Weynandt & Bosnjak, 2015).

Prior research shows that survey duration can be shortened by using an auto-forward design (Giroux, Tharp, & Wietelman, 2019; Selkälä & Couper, 2018; de Bruijne, 2016; Lugtig, Toepoel, Haan, Zandvliet, & Klein Kranenburg, 2019). In an auto-forward design, respondents automatically advance to the next question after an answer is given. This design feature can improve the survey experience in two ways. First, the required cognitive effort by respondents is reduced by adding smart navigation (i.e., to not have to decide whether the question was the last on the page and to not have to search for the 'next' button). Second, as auto-forward can increase the speed of the survey's advancement, the time spent on the survey is reduced. Respondents find surveys with auto-forward more enjoyable, more interesting, less difficult, and less lengthy compared to designs where manual-forwarding is the standard (Roberts, de Leeuw, Hox, Klausch, & de Jongh, 2012). Furthermore, auto-forwarding seems to decrease satisficing behavior (Selkälä, Callegaro, & Couper, 2020).

There are also potential disadvantages to using auto-forwarding (for an overview, see Giroux et al. 2019). Respondents may get confused because they are used to a page-by-page design in which they use navigation buttons which are often provided in web surveys (Bergstrom, Lakhe, & Erdman, 2016). This confusion may

Acknowledgments

Jeldrik Bakker and Marieke Haan contributed equally to this work and share the co-first authorship

Direct correspondence to

Jeldrik Bakker, Statistics Netherlands & Utrecht University
E-mail: j.bakker@cbs.nl

lead to the accidental skipping of questions resulting in higher item nonresponse (de Bruijne, 2016). Furthermore, the automated pace of the survey may discourage respondents to change answers by using navigation buttons which can lead to more suboptimal responses. Finally, many surveys include questions that are not fit for auto-forwarding, such as open answer questions or "select all that apply" questions. If some questions are auto-forwarded and others not, this may also confuse respondents. In this paper, we use paradata, more specifically we analyze the clicking and answering behavior and response timings between the manual- and auto-forward versions to better understand how auto-forwarding affects both response times and data quality. For this, we use an experimental design that was embedded in a health survey conducted among the general population in The Netherlands.

The remainder of this paper is organized as follows: In section 2, we introduce our research questions and hypotheses. In section 3, we describe the data and methods. We discuss results in section 4. We end with conclusions and discussion in the last two sections.

Study Design and Research Questions

We build on earlier studies that used auto-forwarding design (for an overview see: Giroux et al., 2019). Most of these studies show that response times are generally shortened because of auto-forwarding (Hays et al., 2010; Roberts et al., 2012; Selkäla & Couper, 2018 – for PCs only; Lugtig et al., 2019), but some researchers also find no effects on completion times between auto-forward and manual-forward surveys (Arn et al., 2015; de Bruijne, 2015; Selkäla & Couper, 2018 – for smartphones only), or even longer completion times for auto-forward surveys (Roberts et al., 2013). In this paper, we focus on response times, respondent navigation behavior (i.e., mouse clicks or taps with a finger) and how often respondents answer 'I don't know' and 'I rather not say'. We answer four research questions: 1) Does auto-forwarding reduce response times?, 2) Does auto-forwarding lead to more efficient navigation through the survey?, 3) If so, is more efficient navigation independent of screen size?, and 4) Does auto-forwarding affect how often the answer options 'I don't know' and 'I rather not say' are used?

Our first research question comes from the hypothesis (H1) that auto-forwarding reduces the amount of time needed per survey question. We answer this question in the context of official general population surveys that often are, or were, interviewer-assisted and traditionally have a survey duration of 30 minutes and longer. Our second research question is, however, the most important: it concerns the actual effort needed by respondents to navigate through the survey. To investigate efficient navigation, we compare the number of clicks between an auto-forward version and a manual-forward version of a survey. A respondent is not efficiently

navigating through the survey when navigation buttons are used unnecessarily. We expect that auto-forwarding results in more efficient navigation (H2). The third research question is a follow-up question, which differentiates among smartphones, tablets and PCs. We expect to find more efficient completion on smaller screens (H3). The fourth research question is a first exploration into the impact of an auto-forward interface on item-nonresponse. Because almost all questions that are used in our survey are mandatory, the alternatives for item-nonresponse are: 'I rather not say', 'I don't know,' or selecting a random answer option. In line with the research of Selkälä et al. (2020) we expect that auto-forwarding decreases satisficing behavior, which we define in less 'I rather not say' and 'I don't know' responses (H4).

In order to investigate the four questions, we collected and analyzed audit trail paradata at the survey page-level (see Kreuter, 2013). The paradata we collected provide information about each page of the web survey and about each action requiring server contact (e.g., navigating to the next or the previous page, or start/quit the survey), including page-level response times. Our study will help to determine whether auto-forwarding should be used more widely in web surveys.

Method

Data Collection

Our experiment was linked to the Health Survey (HS) of Statistics Netherlands (SN), which is a repeated cross-sectional survey employing monthly simple random samples from the Dutch population register. The HS is a relatively long survey, with a median completion time of 29.2 minutes. It consists of 409 questions divided over 220 web pages, covering 48 topics, ranging from general health, visits to general practitioners and dentists, hospitalization, medicine use, to health-related behaviors such as smoking, food intake, and physical activity. Respondents have to go through all modules, but the number of questions per module varies based on their medical history and lifestyle. The survey had a predefined order and questions about the same topic were grouped together. The location of the auto-forward questions and manual-forward questions was almost randomly distributed over the survey, except for a block of questions about activities. This block primarily asked questions about either frequencies or duration of activities, and consisted almost solely of questions where auto-forwarding was not possible. The HS uses a sequential mixed-mode design with web followed by face-to-face interviewing. In this paper, we only use the web-administered part of the survey.

The HS auto-forwarding experiment employed a separate sample that ran parallel to the regular HS. The sampling frame was composed of earlier respondents to SN surveys of individuals aged 16 years and older that responded to at least

one of the surveys on a mobile device in the period of September 2016 to June 2017. The stratified simple random sample design with six strata was used: three age groups (16–29, 30–49, and 50 years and older) crossed with a type of device (smartphone, tablet). From each stratum the same number of sampling units was selected, leading to unequal sample inclusion probabilities. Thus, older respondents and respondents who previously used a tablet for survey completion have larger inclusion probabilities. We chose this sampling design in order to be reach higher statistical efficiency in testing the impact of device and age on response times and survey navigation. Sampled respondents were randomly allocated to one of the interface conditions: manual-forward and auto-forward (see section 3.2). Fieldwork took place in August–September 2017. Paradata on response times and navigation were collected using version 5.0.5 of the BLAISE computer-assisted interviewing system (Blaise, 2018).

Overall, 2098 individuals were sent an invitation letter by post and a maximum of two reminders in case they did not participate after one and two weeks. All sample members received a 5€ unconditional cash incentive. In total, 1535 sample units started the survey and 1461 sample units completed the survey with a response rate of 69.6% (AAPOR 2016, RR1). The high response rate can be partly explained by the sample composition of former respondents that completed at least one survey of SN on a smartphone or tablet. In total 74 respondents (4.8%) broke off the survey, 45.9% under the auto-forward condition and 54.1% under the manual-forward condition.

Table 1 shows the choice of device of respondents by age group and highest-attained educational level. The break-off rates per device varied very little and are not shown.

Table 1

Device use by age

	Smartphone		Tablet		PC		Unknown		Total		RR1
Age	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)	(%)
16-29	211	(48.8)	104	(24.1)	117	(27.1)	0	(0)	432	(100)	61.9
30-49	179	(37.1)	204	(42.2)	99	(20.5)	1	(0.2)	483	(100)	69.0
50+	126	(23.1)	276	(50.5)	142	(26.0)	2	(0.4)	546	(100)	78.0
Total	516	(35.3)	584	(40.0)	358	(24.5)	3	(0.2)	1461	(100)	69.6

Device use by education level

	Smartphone		Tablet		PC		Unknown		Total	
Education	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)
Low	48	(30.4)	84	(53.2)	26	(16.5)	0	(0)	158	(100)
Middle	198	(39.4)	186	(37.0)	119	(23.7)	0	(0)	503	(100)
High	241	(35.2)	252	(36.8)	189	(27.6)	2	(0.3)	684	(100)
Other	29	(25.0)	62	(53.4)	24	(20.7)	1	(0.9)	116	(100)
Total	516	(35.3)	584	(40.0)	358	(24.5)	3	(0.2)	1461	(100)

Design of the Survey Interface

At the start of the survey, respondents were randomized into one of two interface conditions:

1) In the manual-forward version, respondents had to navigate between survey web pages using 'previous' and 'next' buttons (the default design in surveys fielded by SN.
2) In the auto-forward version, respondents were auto-forwarded to the subsequent survey web page when they answered the last question, unless that last question was a 'check all that apply' question or an open-ended question.

Within the auto-forward condition, auto-forwarding was applied for 75.5% of the pages. For 24.5% of the pages which contained 'check all that apply' questions or open questions, manual-forwarding was applied. Respondents were required to answer every question within the survey except for questions about sexuality.

The auto-forward interface included 'previous' and 'next' buttons and was completely similar in the visual design to the manual-forward interface (see Figure A1 in the Appendix). Respondents could thus navigate backward and forward in the auto-forward condition when they, for example, wanted to correct an answer provided earlier in the survey or review a previous question. We decided to include the 'next' button in the auto-forward condition to avoid confusion between pages where auto-forward was possible and those where it was not. Respondents were not informed about the auto-forward design prior to the survey start.

Data Preparation

Before we move to the analysis methods, we first describe the data preparation. The data preparation consisted of three steps: selection of complete responses, processing of paradata, and omission of outliers.

As a first step, we selected only those cases with complete data. We removed the 74 sample units who broke off as they provided only partial information on response times. Given the small size of this group, we decided not to complicate our analyses by including censored data. After the selection, we had 713 respondents in the auto-forward condition and 748 respondents in the manual-forward condition.

As a second step, we translated the web survey paradata to meaningful features and variables. We coded the device that respondents used to complete the survey using user agent strings. Whenever a person accesses any website, the website receives information. This information is referred to as the user agent string and contains characteristics of the device in order for the website to be able to adapt to the device. These strings have a known format and allow one to derive the type of

device. For three respondents, the user agent string showed that a mobile device was used, but it was unclear whether it was a smartphone or a tablet. We excluded these three respondents from the analysis. Some respondents (n=53) switched between devices during the survey. In the analysis, these respondents are allocated to the device in which they answered the majority of the questions. Next, we processed the survey web page response times. The page-level response time was calculated as the difference between the time stamp of entering a page and the time stamp of leaving the page. The total response time (i.e., respondent-level) was calculated by summing up the page-level response times for a respondent. Since both respondent-level and page-level response times are right-skewed, we applied a log transformation to the response times.

In the third step, we removed outliers at the respondent level and at the page level. We applied the interquartile rule for outliers for both respondent-level and page-level outliers. We calculated the interquartile range (IQR) for the data, multiplied the IQR by 1.5, and added this to the third quartile (Upton & Cook, 1996). A log-transformed response time was marked as an outlier if it was larger than the third quartile plus 1.5 times the IQR. At the respondent level, 14 respondents were removed based on the interquartile rule, leading to 1,444 respondents (705 in the auto-forward design and 739 in the manual-forward design). At the page level, about three percent of the log-transformed response times were removed (i.e., 4,589 out of 152,423 log-transformed response times).

In the following sections, all response times are transformed back from the log scale to aid interpretation.

Analysis

We answer the four research questions through three analyses. We use multi-level analysis to answer the first research question on response times. We use standard regression analysis explaining the numbers of navigational actions to answer the second and third research questions. We use Chi-square tests to answer the final research question on the choice of 'I don't know' and 'I rather not say' responses. All analyses were conducted in R version 3.6.2 (R Core Development Team, 2019).

Multi-level analysis of log response times. Similar to Antoun and Cernat (2019), the page-level log-transformed response times form the dependent variable in the analysis which are clustered by adding a level for the respondent and a level for the page. The respondent-specific influence and the page-specific influence are entered as a random effect. We include experimental condition, age, education and type of device as explanatory variables at the respondent-level and include respondent random effects that vary across age and device groups.

Regression analysis of navigation behavior by clicks and taps (from here on called clicks). We first investigated the clicks between conditions with descriptive

statistics using normalized data, meaning the number of clicks was divided by the number of respondents in each group.

Secondly, we conducted a regression analysis where we included all the previous button clicks as well as the unnecessary use of the 'next' button (i.e., failed attempts to proceed to the next page). To minimize item non-response, all survey questions - except the questions about sexuality - were mandatory. Clicking the 'next' button without answering the question thus resulted in a warning message that a question was left unanswered preventing moving forward to the next page. The unnecessary clicks were all caused by manually clicking the 'next' button while not having answered all of the questions on a page.

For more insight, we followed-up with an investigation of the 10 pages where differences in clicks between the two conditions were the largest. The difference in clicks was calculated by taking the absolute difference between the number of clicks per page per type of navigation button in the manual-forward condition and the auto-forward condition.

Chi-square tests for the answer options 'I don't know' and 'I rather not say.' For each answer option, a Chi-square test is conducted to test in which condition this type of answer is used the most. To simplify the analysis, we compared respondents that never chose such answers to respondents that chose such answers at least once.

Results

Does auto-forwarding reduce response times?

Table 2 shows several models to explain the variance in the log-transformed response time. In the empty model (i.e., the model with no predictors), 60% of the variance in the log response time was explained by the page and 10% by the respondent. The full model only included variables related to the respondent and this model explained 24% of the respondent variance.

These results confirm our first hypothesis (H1) that auto-forwarding reduces the total response times. When correcting for education, age, device, and including the interaction of device and age, respondents in de auto-forward condition required on average 0.65 seconds less per page than respondents in the manual-forward condition (10.97 vs. 11.61 seconds). The survey consisted of an average of 106.6 pages, which, thus, translates to an average 68.9 seconds reduction of the total completion time.

Table 2 Log-transformed response time per page predicted by type of navigation, age, education, device for the total data and split between pages with single choice & matrix questions (only pages with auto-forward) and open ended & check all that apply questions (only pages with manual-forward)

Model	Empty model		+ navigation		+ age and education		+ device		+ device * age		Only pages with auto- forward		Only pages with manual- forward	
	Coeff.	(s.e.)	Coeff.	(s.e.)	Coeff.	(s.e.)	Coeff.	(s.e.)	Coeff.	(s.e.)	Coeff.	(s.e.)	Coeff.	(s.e.)
Fixed part														
Intercept	2.45	(.04) ***	2.48	(.04) ***	2.46	(.04) ***	2.45	(.04) ***	2.45	(.04) ***	2.32	(.04) ***	2.82	(.07) ***
Navigation (Ref. = Manual-forward)														
Auto-forward			-0.05	(.01) ***	-0.06	(.01) ***	-0.06	(.01) ***	-0.06	(.01) ***	-0.07	(.01) ***	0.02	(.01)
Education (Ref. = Low)														
Middle					-0.05	(.02) **	-0.04	(.02) *	-0.04	(.02) *	-0.05	(.02) ***	-0.03	(.02)
High					-0.13	(.02) ***	-0.12	(.02) ***	-0.12	(.02) ***	-0.12	(.02) *	-0.09	(.02) ***
Other					0.01	(.03)	0.01	(.03)	0.01	(.03)	0.01	(.03) ***	0.00	(.03)
Age (Ref. = 16-29)														
30-49					0.08	(.01) ***	0.07	(.01) ***	0.06	(.02) **	0.05	(.02) **	0.12	(.02) ***
50+					0.21	(.01) ***	0.20	(.01) ***	0.20	(.02) ***	0.18	(.02) ***	0.25	(.03) ***
Device (Ref. = Smartphone)														
Tablet							0.05	(.01) ***	0.05	(.02) *	0.05	(.02) *	0.09	(.03) **
PC							-0.04	(.01) ***	-0.06	(.02) **	-0.06	(.02) *	-0.03	(.03)
*Device * Age*														
Tablet * 30-49									-0.00	(.03)	-0.01	(.03)	-0.05	(.04)
PC * 30-49									0.04	(.03)	0.04	(.03)	-0.08	(.04)
Tablet * 50+									-0.01	(.03)	-0.01	(.03)	-0.01	(.04)
PC * 50+									0.02	(.03)	0.02	(.03)	-0.02	(.04)

Model	Empty model		+ navigation		+ age and education		+ device		+ device * age		Only pages with auto- forward		Only pages with manual- forward	
	Coeff.	(s.e.)	Coeff.	(s.e.)	Coeff.	(s.e.)	Coeff.	(s.e.)	Coeff.	(s.e.)	Coeff.	(s.e.)	Coeff.	(s.e.)
Random part														
$\sigma^2_{respondent}$	0.06	(.24)	0.06	(.24)	0.04	(.21)	0.04	(.21)	0.04	(.21)	0.04	(.21)	0.05	(.22)
σ^2_{page}	0.28	(.53)	0.28	(.53)	0.28	(.53)	0.28	(.53)	0.28	(.53)	0.21	(.46)	0.23	(.48)
$\sigma^2_{residuals}$	0.14	(.37)	0.14	(.37)	0.14	(.37)	0.14	(.37)	0.14	(.37)	0.13	(.36)	0.17	(.42)
Marginal R^2	0.00		0.00		0.03		0.03		0.03		0.04		0.04	
Conditional R^2	0.71		0.71		0.71		0.71		0.71		0.68		0.63	
Deviance	131773		131757		131393		131337		131335					
AIC	131781		131767		131419		131367		131373					
BIC	131821		131817		131547		131516		131562					

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

The reduction in response time was only observed for pages where an auto-forward functionality could be applied (i.e., pages with only single choice or matrix questions). On these pages, the auto-forward functionality resulted in a 0.72 second or 7.1% reduction in response time (10.16 vs. 9.44 seconds); $t(1,417) = -6.64$, $p < .001$. On the other pages (i.e., pages with open-ended and check-all-that-apply questions), we observed a 0.28 seconds increase in response time (16.85 vs. 17.13 seconds). The latter difference is not significant; $t(1,420) = 1.27$, $p = .20$.

As for education, higher-educated respondents completed the survey faster than lower-educated respondents; $t(1,426) = -5.98$, $p < .001$. Furthermore, older respondents needed more time to complete the survey than the other age groups, with the youngest respondents being the fastest; $t(1,776) = 8.41$, $p < .001$. Tablet users needed more time to complete the survey than smartphone users: $t(1,672) = 2.04$, $p = .04$, while PC users needed less time: $t(2,055) = -2.86$, $p = .004$. Finally, we did not find interaction effects between age and device type.

Does auto-forwarding lead to more efficient navigation through the survey, and, if so, is any improvement related to type of device?

Contrary to our hypothesis (H2), auto-forwarding led to less efficient navigation through the survey. When looking at all navigations (i.e., automated navigations and the manual clicks), auto-forwarding increased the average number of clicks to the previous page by 1.0 (auto-forward: M = 2.9, SD = 6.4; manual-forward: M = 1.9, SD = 2.9) and the (attempted) navigations to proceed to the next page increased by 16.0 (auto-forward: M = 137.3, SD = 23.2; manual-forward: M = 121.3, SD = 9.6). The unnecessary clicks, which are all caused by manual clicking, account for 16.0% of the total next-page navigations and are also more frequent in the auto-forward condition (auto-forward: M = 28.1, SD = 20.8; manual-forward: M = 13.5, SD = 5.3).

The results presented in Table 3 confirm that both buttons (i.e., all 'previous' button clicks and unnecessary 'next' button clicks) are used significantly more often in the auto-forward condition. An effect for device was only apparent for respondents aged 50 and older, who used the navigation buttons less when using a PC than when using a mobile device (i.e., a tablet or a smartphone). This finding is in the opposite direction of hypothesis (H3). Furthermore, we found fewer clicks for the higher-educated respondents.

To understand these results better, we examined pages where differences in clicks between the two conditions were the strongest. Tables A1 and A2 (see Appendix) provide an overview of the pages with the largest difference in clicks per type of navigation, including the difference in the number of clicks between the conditions.

As Table A1 shows (see the Appendix), the 'previous' button is used most in the auto-forward condition when questions are cognitively demanding, when a new

Table 3 Regression analyses with number of clicks per person as a dependent variable

	Estimate (B)	SE	t
Intercept	13.51 ***	0.93	14.48
Button (Ref. = Next)			
Previous	-11.51 ***	0.58	-20.00
Condition (Ref. = Manual-forward)			
Auto-forward	14.47 ***	0.58	24.82
Device (Ref. = Smartphone)			
Tablet	1.43	0.94	1.52
PC	-0.19	0.91	-0.21
Age (Ref. = 16-29)			
30-49	0.02	0.80	0.03
50+	1.81 *	0.91	2.00
Education (Ref. = Low)			
Middle	-0.36	0.74	-0.49
High	-1.52 *	0.73	-2.09
Other	-1.05	0.97	-1.08
Button * Condition			
Previous * Auto-forward	-13.62 ***	0.82	-16.54
Device * Age			
Tablet * 30-49	0.72	1.23	0.58
PC * 30-49	-0.97	1.33	-0.73
Tablet * 50+	0.01	1.27	0.01
PC * 50+	-4.01 **	1.33	-3.02
$R^2 = 0.48$			

*$p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

topic is introduced, or when respondents think they might have answered a question already (i.e., respondents check the previous question because of similarities in question wordings). Within the auto-forward condition, we do not find increased use of the previous button between pages with automated navigation and pages with manual navigation, indicating respondents are not confused by this transition; $t(356) = 0.43$, $p = .67$.

Respondents unnecessarily use the 'next' button most in the auto-forward condition. This finding is most apparent on pages with multiple questions (see Table 4, Table A2 in the Appendix). On those pages, multiple single-choice questions were presented.

Table 4 Regression analyses with the frequency of using the 'next' button unnecessarily per page as a dependent variable

	Estimate (B)	SE	t
Intercept	3.10 ***	0.29	10.64
Condition (Ref. = Manual-forward)			
Auto-forward	0.93 *	0.40	2.36
Number of questions (Ref. = 1)			
2	1.96 ***	0.25	7.68
> 2	2.02 ***	0.32	6.30
Question type (Ref. = open/check-all that apply)			
Single-choice or matrix	-0.86 **	0.28	-3.12
Number of questions * condition			
2 questions * Auto-forward	-1.17 ***	0.35	-3.37
> 2 questions * Auto-forward	-0.74	0.44	-1.68
Question type * Condition			
Single-choice or matrix * auto-forward	0.83 *	0.38	2.17
$R^2 = 0.27$			

*$p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Does auto-forwarding affect how often the answer options 'I don't know' and 'I rather not say' are used?

An auto-forward functionality had no effect on how often respondents gave either an 'I rather not say' or an 'I don't know' answer. Contrary to our expectations (H4), these two answer options were used just as often in the auto-forward condition as in the manual-forward condition. The answer 'I rather not say' was given at least once by 79.0% of the respondents in the manual forward condition and by 80.1% in the auto-forward condition; $\chi 2(1, N=1,444) = 0.28$, $p=.59$. The answer 'I don't know' was given at least once by 30.6% of the respondents in the manual-forward condition and by 33.3% in the auto-forward condition; $\chi 2(1, N=1444) = 1.29$, $p=.26$.

Conclusion

In this study, we randomly assigned respondents to an auto-forward design or a manual-forward design in a long mixed-device web survey on health. We compare these two conditions across devices used for survey completion (PC, tablet, and smartphone). We find slightly shorter completion times for all devices in the

auto-forward design compared to the manual-forward design. Results also show that questions on pages with automated navigation are answered significantly faster than the questions on pages with manual navigation (i.e., where respondents needed to use the navigation buttons).

However, the difference in completion times between the conditions is relatively small. Therefore, we used paradata to investigate how respondents navigated the survey. Analyses of clicks on the 'previous' button show that it is used more often in the auto-forward condition compared to the manual-forward condition. Such increased use might be explained by the novelty of the design and its pace: respondents may not be used to automated navigation within a survey. Within the auto-forward condition, we do not find more use of the 'previous' button between pages with automated navigation and pages with manual navigation, indicating that respondents are not confused by this transition.

We also find that respondents in the auto-forward condition unnecessarily use the 'next' button (i.e., failed attempts to proceed to the next page). This finding may be explained by the following reasons: 1) respondents wish to navigate faster than the pace of the automated navigation of the survey, 2) respondents are not used to an auto-forward design and use the 'next' button as a common habit, 3) respondents did not notice that a new page with a new question has finished loading, or 4) respondents who mistakenly missed a question might think they should click the next button because they are not taken to the next page. However, in reality, these respondents forgot to fill in a question and for that reason they do not automatically go to the next page. Only after filling in the overlooked question, they will automatically be forwarded to the next page (i.e., the next button should not be used in this situation).

Contrary to our expectation, we found no significant difference for the use of 'I don't know' and 'I rather not say' answers between the auto-forward and manual-forward condition.

This result deviates from the outcomes of Selkälä et al.'s study (2020). The difference between their study and ours is that we used a long survey with different types of questions with a mix of auto-forward and manual-forward which may have affected answering behavior differently.

Discussion

Overall, we conclude that auto-forwarding can be used to reduce completion times. Since it is difficult to include auto-forwarding with check-all-that-apply, open and numerical questions we advise to carefully consider mixing manual and auto-forwarding within one survey. Ideally, survey layout and navigation should be predictable within a survey and across devices (Antoun, Katz, Argueta, & Wang, 2018).

In line with the recommendations of Giroux et al. (2019), we advise to include clear instructions to inform respondents about their navigation possibilities within the survey. A particular challenge for future research is how to implement auto-forwarding in surveys that include different types of questions.

Our study has some limitations. The main limitation, as mentioned above, is that our survey contained questions in which auto-forward cannot be applied. Future research should replicate our design in a long survey where auto-forward can be applied to all questions. A second limitation is the self-selection of respondents to complete the survey on a mobile device. Random assignment of respondents to a certain device leads to issues of respondent noncompliance (de Bruijne & Wijnant, 2013; Mavletova, 2013; Wells, Bailey, & Link, 2014). Therefore, our sample was composed of earlier respondents to SN individual surveys that responded at least once with a mobile device. Those respondents are likely to be more motivated than a freshly recruited cross-section.

Another further step would be to examine the quality of answers provided to different auto-forward interface conditions in more detail. We only explored the impact of auto-forwarding on item nonresponse. Furthermore, we advise to evaluate users' experience of the auto-forward interface in more detail pre- or post-survey, for example, by conducting semi-structured open interviews and adding open-ended evaluation questions.

Data Availability

The data are available on site or by means of remote access. This can be requested by contacting the corresponding author at j.bakker@cbs.nl .

Software Information

We used R version 3.6.2 (R Core Development Team, 2019). The R-script can be requested by contacting the corresponding author at j.bakker@cbs.nl . Paradata were collected using Version 5.0.5 of the BLAISE computer-assisted interviewing system (Blaise, 2018).

References

American Association for Public Opinion Research (2016). Standard definitions: Final dispositions of case codes and outcome rates for surveys. Ann Arbor, MI: AAPOR.

Antoun, C., & Cernat, A. (2019). Factors affecting completion times: A comparative analysis of smartphone and PC web surveys. *Social Science Computer Review*, *38*, 477-489. https://doi.org/10.1177/0894439318823703

Antoun, C., Katz, J., Argueta, J., & Wang, L. (2018). Design heuristics for effective smartphone surveys. *Social Science Computer Review*, *36*, 557-574. https://doi.org/10.1177/0894439317727072

Arn, B., Klug, S., & Kolodziejski, J. (2015). Evaluation of an adapted design in a multi-device online panel: A DemoSCOPE case study. *Methods, data, analyses*, *9*, 185-212. https://doi.org/10.12758/mda.2015.011

Bergstrom, J. C., Lakhe, S., & Erdman, C. (2016). Navigation buttons in web-based surveys: Respondents' preferences revisited in the laboratory. *Survey Practice*, *9*. https://doi.org/10.29115/SP-2016-0005

Blaise (2018). https://www.blaise.com (accessed: August 2018).

Cook, W. A. (2014). Is mobile a reliable platform for survey taking? Defining quality in online surveys from mobile respondents. *Journal of Advertising Research, 54*, 141–148. https://doi.org/10.2501/JAR-54-2-141-148

Couper, M. P., Antoun, C., & Mavletova, A. (2017). Mobile web surveys: A total survey error perspective. In P. Biemer, S. Eckman, B. Edwards, E. de Leeuw, F. Kreuter, L. Lyberg, C. Tucker, & B. West (Eds.), *Total Survey Error in Practice*, (pp. 133-154). New York, NY: Wiley.

Couper, M. P., & Peterson, G. J. (2017). Why do web surveys take longer on smartphones? *Social Science Computer Review*, *35*, 357-377. https://doi.org/10.1177/0894439316629932

De Bruijne, M.A. (2015). Designing web surveys for the multi-device internet. PhD thesis., The Netherlands: Tilburg University: Center for Economic Research.

De Bruijne, M.A. (2016). Online vragenlijsten en mobiele devices (Online questionnaires and mobile devices). *Jaarboek van de Marktonderzoekassociatie*, 137-151. https://adoc.pub/9-online-vragenlijsten-en-mobiele-devices.html

De Bruijne, M.A., & Wijnant, A. (2013). Comparing survey results obtained via mobile devices and computers: An experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey. *Social Science Computer Review*, 31(4), 482-504. https://doi.org/10.1177/0894439313483976

Giroux, S., Tharp, K., & Wietleman, D. (2019). Impacts of implementing an automatic advancement feature in mobile and web surveys. *Survey Practice*, *12*. https://doi.org/10.29115/SP-2018-0034

Hays, R. D., Bode, R., Rothrock, N., Riley, W., Cella, D., & Gershon, R. (2010). The impact of next and back buttons on time to complete and measurement reliability in computer-based surveys. *Quality of Life Research*, *19*, 1181-1184. https://doi.org/10.1007/s11136-010-9682-9

Hintze, D., Findling, R, D., Scholz, S., & Mayrhofer, R. (2014, December). Mobile device usage characteristics: The effect of context and form factor on locked and unlocked usage. In *Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia* (pp. 105-114). ACM. https://doi.org/10.1145/2684103.2684156

Kreuter, F. (2013) Improving surveys with paradata: Introduction. In F. Kreuter (Ed.)*, Improving surveys with paradata: Analytic uses of process information*, (pp. 1-9). New York, NY: Wiley.

Lugtig, P., Toepoel, V., Haan, M., Zandvliet, R., & Klein Kranenburg, L. (2019). Recruiting hard-to-reach groups into a probability-based online panel by promoting smartphone use. M*ethods, data, analyses*, *13*, 291-306. https://doi.org/10.12758/mda.2019.04

Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review*, *31*, 725-743. https://doi.org/10.1177/0894439313485201

R Core Development Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Roberts, A., de Leeuw, E. D., Hox, J., Klausch, T., & de Jongh, A. (2012). Leuker kunnen het wel maken. Online vragenlijst design standaard matrix of scrollmatrix. *In het 38e jaarboek van de MOA : Developments in Market Research*, (pp. 133-148). http://dspace.library.uu.nl/handle/1874/291084

Selkälä A., Callegaro M., & Couper M.P. (2020) Automatic Versus Manual Forwarding in Web Surveys - A Cognitive Load Perspective on Satisficing Responding. In: Meiselwitz G. (eds). *Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis. HCII 2020. Lecture Notes in Computer Science*, vol 12194. Springer, Cham. https://doi.org/10.1007/978-3-030-49570-1_10

Selkälä, A., & Couper, M.P. (2018). Automatic versus manual forwarding in web surveys S*ocial Science Computer Review*, *36*, 669-689. https://doi.org/10.1177/0894439317736831

Struminskaya, B., Weyandt, K., & Bosnjak, M. (2015). The effects of survey completion using mobile devices on data quality - Evidence from a probability-based general population panel. *Methods, data, analyses*, *9*, 261–292. https://doi.org/10.12758/mda.2015.014

Upton, G. & Cook, I. (1996). Understanding Statistics. Oxford: Oxford University Press.

Wells, T., Bailey, J. T., & Link, M. W. (2014). Comparison of smartphone and online computer survey administration. *Social Science Computer Review, 32*, 238-255. https://doi.org/10.1177/0894439313505829

Zhang, C. & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*, 127-135. https://doi.org/10.18148/srm/2014.v8i2.5453

Appendix

Table A1 Top 10 pages with the largest difference in click behavior of the 'previous' button between the auto-forward and manual-forward condition.

	Auto-forward		Manual-forward		Difference			
Page nr*	n	%	n	%	n	%	Topic	Remarks/difficulties
220	38	1.9	7	0.5	31	3.6	Occupational accident	Intro page + long text
214	55	2.7	27	1.9	28	3.2	Eating vegetables	Similar question block to previous** + specifications
60	35	1.7	10	0.7	25	2.9	Education	Similar question block to previous** + specifications
81	30	1.5	10	0.7	20	2.3	Education finished	-
184	31	1.5	11	0.8	20	2.3	Blood sugar	-
218	23	1.1	4	0.3	19	2.2	Eating fruit	Relates to the previous page
255	33	1.6	14	1.0	19	2.2	Alcohol consumption	Very similar to the previous question (4 vs. 6 glasses)
49	28	1.4	10	0.7	18	2.1	Paid work	Numeric question + conditional 2nd quest.
219	25	1.2	8	0.6	17	1.9	Eating fish	Similar questions to previous[1] + specifications
225	23	1.1	6	0.4	17	1.9	Accidents	Intro page + long text + possible sensitive subject
Total	321	15.7	107	7.6	214	24.6		

*The page number is the raw numbering according to the programming of the survey. Many pages are not shown to respondents due to routing.
**Similar question block to the previous question block refers to the almost exact same wording of the blocks.

Table A2 Top 10 pages with the largest difference in unnecessary click behavior of the 'next' button* between the auto-forward and manual-forward condition

	Auto-forward		Manual-forward		Difference				
Page nr**	n	%	n	%	n	%	Topic	Remarks/difficulties	# Questions
156	321	1.6	36	0.4	285	2.8	Chronic disease	Matrix questions	10
158	288	1.5	24	0.2	264	2.6	Psychological health	Matrix questions	5
155	321	1.6	79	0.8	242	2.4	Chronic disease	Matrix questions	11
159	266	1.3	30	0.3	236	2.3	Acute illness	Matrix questions	6
257	256	1.3	44	0.4	212	2.1	Narcotic use	Matrix questions	11
59	359	1.8	150	1.5	209	2.0	Education	Intro page + long introduction	1
11	842	4.3	634	6.4	208	2.0	Household info	Multiple questions on screen	3
157	263	1.3	102	1.0	161	1.6	Chronic disease	2nd question on same page is conditional	2
148	170	0.9	25	0.3	145	1.4	Diabetes	2nd and 3rd questions are conditional	3
160	154	0.8	9	0.1	145	1.4	Pain	2nd question on same page is conditional	2
Total	3,240	16.4	1,133	11.4	2,107	20.5			

* Failed attempts to proceed to the next page.
**The page number is the raw numbering according to the programming of the survey. Many pages are not shown to respondents due to routing.

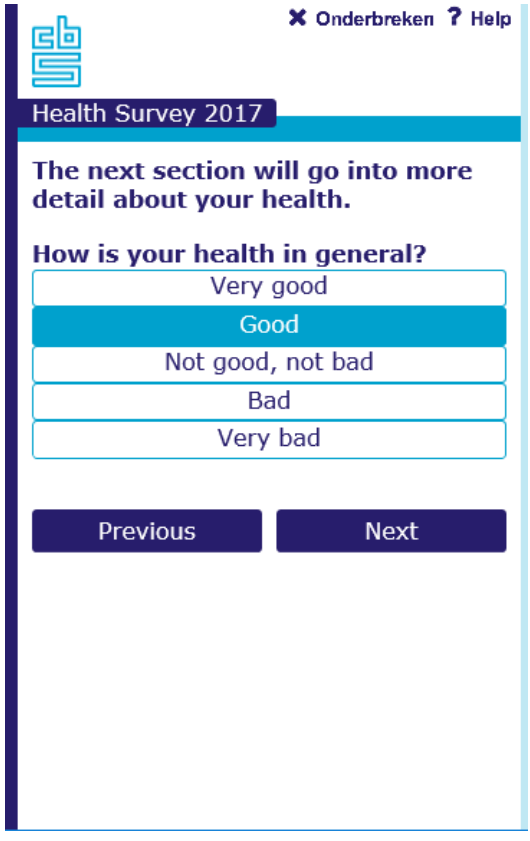

Figure A1 Screenshot of the survey layout

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)