Schriftenreihe der ASI –
Arbeitsgemeinschaft Sozialwissenschaftlicher Institute

Natalja Menold
Tobias Wolbring *Hrsg.*

Qualitätssicherung sozialwissenschaftlicher Erhebungsinstrumente

Springer VS

Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute

Reihe herausgegeben von
Frank Faulbaum
Universität Duisburg-Essen

Stefanie Kley
Universität Hamburg

Birgit Pfau-Effinger
Universität Hamburg

Jürgen Schupp
DIW Berlin

Jette Schröder
GESIS – Leibniz-Institut für Sozialwissenschaften

Christof Wolf
GESIS – Leibniz-Institut für Sozialwissenschaften

Weitere Bände in der Reihe http://www.springer.com/series/11434

Natalja Menold · Tobias Wolbring
(Hrsg.)

Qualitätssicherung sozialwissenschaftlicher Erhebungsinstrumente

Springer VS

Hrsg.
Natalja Menold
GESIS – Leibniz-Institut
für Sozialwissenschaften
Mannheim, Deutschland

Tobias Wolbring
FAU Erlangen-Nürnberg
Nürnberg, Deutschland

ISSN 2625-9427 ISSN 2625-9435 (electronic)
Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute
ISBN 978-3-658-24516-0 ISBN 978-3-658-24517-7 (eBook)
https://doi.org/10.1007/978-3-658-24517-7

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer VS

Springer VS ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Inhalt

Ansätze zur Antwortvalidität

Qualitätsmangement in der Praxis

Vorwort

Daten aus allgemeinen Bevölkerungsumfragen, aber auch aus Befragungen von Spezialpopulationen werden in den Sozialwissenschaften und darüber hinaus weithin genutzt, um gesellschaftliche Zustände und deren Wandel zu beschreiben, theoretische Erklärungsansätze zu testen und praktische Empfehlungen für Politik, Wirtschaft und Gesellschaft abzuleiten. Sozialwissenschaftliche Erhebungsinstrumente sind die verwendeten Messinstrumente, also bei Umfragen die Fragen und Items im Fragebogen. Die Qualität sozialwissenschaftlicher Erhebungsinstrumente hat wesentlichen Einfluss auf die Belastbarkeit entsprechender empirischer Schlussfolgerungen, die auf Grundlage von Umfragedaten gezogen werden und die vielfältige gesellschaftliche Themen wie etwa soziale Ungleichheiten, politische Stimmungen und die Integration von Migranten betreffen. Es handelt sich dabei um essentielle Fragen der Nutzung der Ergebnisse von Umfragedaten, wie beispielsweise: Was kann man über den untersuchten Gegenstand aussagen? Stellen die Ergebnisse die wahren Unterschiede, Zusammenhänge oder Veränderungen dar, oder handelt es sich um methodische Artefakte und Verzerrungen?

Fragen nach der Qualitätssicherung sozialwissenschaftlicher Erhebungsinstrumente, die im Mittelpunkt dieses Sammelbandes stehen, der im Rahmen der Schriftenreihe der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute (ASI e.V.) erscheint, sind daher zentral für die Nutzung der Umfrageergebnisse in Wissenschaft und Praxis. Betroffen sind davon die Güte sozialwissenschaftlicher Befunde und der kumulative Erkenntnisfortschritt in der Soziologie und anderen sozialwissenschaftlichen Disziplinen. Leidet die Qualität der erhobenen Daten nämlich unter methodischen Defiziten, so lässt sich der Schaden im Nachhinein oft nur schwer oder gar nicht beheben, sofern er den Nutzerinnen und Nutzern der Daten denn überhaupt bekannt ist. Hingegen kann die Datenqualität in den vorbereitenden Phasen von Umfrageprojekten, d.h. bereits im Vorfeld sozialwissenschaftlicher

Datenerhebungen mit entsprechenden Maßnahmen der Qualitätssicherung maßgebend beeinflusst werden.

Qualitätseinschränkungen können sich aus unterschiedlichen Quellen ergeben, wie z.B. Frageformulierung, Gestaltung von Antwortskalen sowie Gliederung und Layout von Fragebögen. Andere unerwünschte Effekte wie Interviewereffekte, Antworttendenzen oder sozialer Erwünschtheit auf der Seite der Befragungsteilnehmer können die mit den Umfragedaten erzielten Ergebnisse verzerren. Kulturvergleichende Umfragen oder Befragungen spezieller Populationen werfen häufig zusätzliche Herausforderungen auf. So stellen sich etwa in diesem Zusammenhang Fragen nach Unterschieden und Problemen im Frageverständnis, aber auch nach kulturellen Unterschieden und zeitlichen Veränderungen im Antwortverhalten. Die Aussagekraft komparativ angelegter Untersuchungen hängt somit also auch wesentlich von der Vergleichbarkeit in räumlicher und zeitlicher Hinsicht und einer gelingenden Harmonisierung von Erhebungsinstrumenten ab.

Diese kurze Darstellung deutet nicht nur die zentrale Bedeutung, sondern auch den Facettenreichtum des Themas „Qualitätssicherung sozialwissenschaftlicher Erhebungsinstrumente“ an. Der vorliegende Band stellt daher die Qualität von Erhebungsinstrumenten, Verfahren zur Bestimmung ihrer Güte und Methoden der Qualitätssicherung in den Mittelpunkt. Trotz des nicht unerheblichen Umfangs des Bandes wird dabei keineswegs der Anspruch erhoben, alle Aspekte des breiten Themenkomplexes umfassenden abzudecken und alle offenen Fragen zu klären. Jedoch hoffen wir mit den vorliegenden Beiträgen wichtige Impulse zur weiteren Steigerung der Güte sozialwissenschaftlicher Daten zu geben und damit indirekt auch positive Wirkungen auf die Belastbarkeit umfragebasierter Forschungsergebnisse und darauf basierender Praxisempfehlungen zu entfalten.

Der Band gliedert sich in vier Abschnitte mit einer unterschiedlichen Anzahl an Beiträgen. Der erste Abschnitt „Messqualität und Messprobleme in der Fragebogenkonstruktion“ umfasst Beiträge zur Identifikation von Messfehlern mit Paradaten, zu Problemen bei der Harmonisierung internationaler Umfragedaten und zur Gestaltung von Antwortskalen. Der zweite Abschnitt „Qualitätssicherung durch qualitative Techniken“ behandelt anhand von Mixed-Method-Studien, qualitativen Interviews und kognitiven Pretests mit Eye-Tracking-Verfahren verschiedene Ansätze zur Entwicklung hochqualitativer Erhebungsinstrumente. Der dritte Abschnitt „Ansätze zur Antwortvalidität“ stellt einerseits auf Probleme sozialer Er-

wünschtheit, insbesondere bei heiklen Themen und fiktiven Sachverhalten, ab und behandelt anderseits Fragen nach der Korrespondenz von Einstellungsmessungen mit tatsächlichem Verhalten. Der letzte Abschnitt „Qualitätsmanagement in der Praxis“ stellt Ansätze der Qualitätssicherung der Erhebungsinstrumente anhand der Befragung der speziellen Population geistig behinderter Jugendlicher und anhand von Umfragen im Gesundheitsbereich vor.

Wir wünschen viel Freude beim Lesen und möchten uns abschließend bei Bettina Zacharias ganz herzlich bedanken, die uns bei der Formatierung der Beiträge für diesen Sammelband tatkräftig unterstützt hat.

Mannheim und Nürnberg im September 2018

Natalja Menold und Tobias Wolbring

Messqualität und Messprobleme in der Fragebogenkonstruktion

Identification of Measurement Problems of Survey Items and Scales Using Paradata

Jochen Mayerl [1], *Henrik Andersen* [1] *& Christoph Giehl* [2]

1 Department of Sociology, Empirical Social Research Unit, Chemnitz University of Technology

2 Department of Social Sciences, Empirical Social Research Unit, Technische Universität Kaiserslautern

Abstract

This article discusses some various applications of paradata in the form of response latencies in identifying survey measurement error. Specifically, it presents empirical analyses regarding response latencies as they pertain to such problems as acquiescence bias, question order effects (contrast and assimilation effects) and social desirability bias. It demonstrates that response latencies can provide helpful insight into cognitive processes that would be otherwise unobservable. Finally, we briefly touch on the challenges involved with the collection and use of paradata.

Keywords: Paradata, measurement error, response effects, response latencies, social desirability, acquiescence, question order, dual-process theory

1 Analytical Framework: Paradata and Measurement Problems

Computer-assisted survey modes like computer assisted self-interviews (CASI, e.g. web surveys), computer assisted telephone interviews (CATI) or computer assisted personal interviews (CAPI) enable the (half-) automatic collection of

N. Menold und T. Wolbring (Hrsg.), *Qualitätssicherung sozialwissenschaftlicher Erhebungsinstrumente*, Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute, https://doi.org/10.1007/978-3-658-24517-7_1

additional context information of surveys. Survey researchers call this additional information paradata (Couper 1998; Couper and Kreuter 2013). In a *narrow sense*, paradata cover all computer-assisted, automatically collected information about the survey and response process. These data are byproducts of surveys that are not directly influenced by interviewers. At the item- and respondent-level this includes data on such things as mouse clicks, time stamps and response latencies. Survey-level paradata (including metadata) cover information on the duration of the surveys, contact information, levels of effort to recruit respondents and, in the case of web surveys, server- and client-side data. In a *broader sense*, paradata cover all kinds of additional context information regarding a survey, including automated data collection, audio recordings (which can be analyzed and coded afterwards) and even observational data such as interviewer protocols.

Paradata have the benefit of being unobtrusive and non-reactive and are thus less susceptible to manipulation by the respondent. In general, paradata can be implemented easily and cost-effectively in computer-assisted surveys. In this paper, we focus on survey response latencies as a special form of paradata. In survey research, response latencies are a promising tool that can be used to identify and explain measurement error and to gain a better understanding of response behaviour, including the cognitive information processes which ultimately lead to responses in surveys (e.g. Mayerl and Urban 2008).

Survey Response Latencies and Mode of Information Processing

In psychological research, reaction times have been used for decades as a common method of measuring cognitive processes (e.g. Fazio 1990b). In survey research, the development of computer assisted survey technology (CATI, CAPI, CASI) made it possible to include such measurements even in large-scale survey projects (e.g. Bassili and Fletcher 1991). Survey response latencies can be used as proxy variables for a wide range of mental processes. One of the most prominent applications involves their use as a proxy measure for cognitive processing modes (e.g. Fazio 1990a; Mayerl 2009) with faster responses suggesting a more automatic-spontaneous mode; slower responses a deliberate-controlled one.[1] A second common application is to interpret

1 For more on dual-process models see Belli et al. 2001; Chaiken 1980; Esser 2010; Fazio 1990b; Krosnick 1991; Mayerl 2009, 2013; and Petty and Cacioppo 1986.

response latencies as a measure of the chronic cognitive accessibility of a social judgments, foremost attitudes (e.g. Fazio 1986), with fast responses indicating a strong cognitive association between an object and its evaluation.[2]

Identification and Explanation of Response Effects

In general, response effects are defined systematic measurement error (i.e. bias) that results from either temporary defined as systematic response *sets* (due to instrument-specific or situational influences) or stable, respondent-related characteristics, known as response *styles* (Mayerl 2013, Paulhus 2002 and Tourangeau and Yan 2007). While the former are content-specific (e.g. question order effects rely on contextual information), the latter are independent of the content of the items being asked (e.g. an individual's tendency to acquiescence, straight-lining, etc.).

A lot of predictor variables have been identified in survey research to explain or predict the occurrence of response effects including, amongst others, lack of motivation or time, lack of knowledge about or interest in the issue, negative attitudes towards surveys in general, personal psychological tendencies (e.g. towards acquiescence or need for social approval), and situational cues or characteristics (see Mayerl 2013). Many of these predictors can be linked to dual-process theory, which states that (situational or individual) motivation and (situational or individual) opportunity are the main mechanisms that govern whether information is processed automatically or deliberately. In this article, we will therefore apply the dual-process theoretical framework to attempt to explain different types of response effects. More precisely, we expect specific response effects to appear in the automatic-spontaneous response mode (i.e. fast responses) and others in the deliberative-controlled mode (i.e. slow responses). This is the theoretical background of our work on how paradata may be used to identify specific measurement problems in surveys.

After shortly discussing data treatment issues involving response latencies (section 2), we present three empirical applications for understanding and

2 Both applications – degree of mental effort and cognitive accessibility – are not contradictory since highly accessible social judgments are mainly predictive within the automatic-spontaneous mode (Fazio 1990a; Mayerl 2013; see section 3.1).

identifying response effects using paradata: acquiescence bias (3.1), question order effects (3.2) and social desirability bias (3.3).

2 Data Collection and Preparation of Response Latencies

The presence or absence of an interviewer is of central importance in the collection and application of response latency data. Computer assisted surveys with interviewers like CATI or CAPI allow relatively precise *active* measures of response times (see Bassili 1996). A 4-screen-per-item-technique (Mayerl and Urban 2008) is an example of an active response latency measurement. From the point of view of the interviewer, it works as follows: the interviewer reads the question (first screen) and immediately starts the reaction time measurement by pressing a key (second screen appears). As soon as the respondent answers, the interviewer stops the timer and the third screen appears where the interviewer saves the given answer. On the fourth and last screen, the interviewer validates the timer measurement (e.g., whether the interviewer made a mistake or the respondent was distracted). The duration of the appearance of the second screen represents the reaction time which will be used later on in the analyses (for more detailed information on the 4-screen-per-item-technique, see Mayerl and Urban 2008).

CASI surveys like web-based or tablet surveys typically only allow the collection of *passive* reaction times, e.g. time stamps measuring the screen time. Such passive measures are less precise since they include reading time and often cannot differentiate between latencies associated with multiple questions that may appear on the same screen simultaneously. There are, however, techniques available to gain more precise latency measurements in CASI surveys. For example, in tablet surveys, client-side Java scripts can be installed to measure latencies based on actions by the respondent (such as mouse clicks), thereby allowing measurements at item-level (see our example in section 3.3).

Response latencies, in order to be interpreted properly, must be treated extensively after data collection (see Fazio 1990a; Mayerl 2013; Mayerl and Urban 2008). To name a few steps, it is necessary to establish the respondents' baseline response latency in order to control for respondents that just generally tend to respond faster or slower than others do. Furthermore, outliers play an important role as they may signify that the measurement was an

invalid indicator of the cognitive process: very long response latencies often mean the respondent was distracted or interrupted while very fast responses may be the result of an error on the part of the respondent or interviewer (accidentally tapping on the screen, for example). Finally, to name just one other issue, response latencies tend to be highly skewed to the right similar to typical income distributions. This can be a problem for ordinary least squares (OLS) multivariate regression analyses and thus researchers often transform the distribution by taking the natural log (for more on the treatment on response latencies, see Couper and Kreuter 2013; Mayerl and Urban 2008).

3 Empirical Applications

3.1 Acquiescence Effects and the Strength of Generalized Attitudes as Predictors

Acquiescence bias is a very prominent response effect and an important source of systematic measurement error in surveys. It simply describes the respondent's personal tendency to generally agree to questions irrespective of the content (e.g. Knowles and Condon 1999; Tourangeau and Rasinski 1988). In the context of a generic dual process model of response behaviour, Mayerl (2009, 2013) describes acquiescence as a simple decision heuristic that results from a lack of motivation and/or opportunity for thoughtful thinking. Thus, acquiescence bias should more often occur in the case of automatic-spontaneous information processing (i.e. short response latencies) than in deliberative information processing (i.e. long response latencies; see Mayerl 2013, p. 6). In addition, in line with Fazio (1990b), spontaneous information processing can be divided into two types of response behaviour depending on the chronic accessibility of the social judgment triggered by the survey question: in the case of high chronic cognitive accessibility, spontaneous information processing is led by this highly accessible judgment (e.g. attitudes). Conversely, in the case of less accessible judgments, simple decision heuristics, as well as situational or contextual cues lead the information processing. Such simple heuristics include response sets like acquiescence in surveys.

Mayerl (2009, 2013) tested both propositions: first, that acquiescence bias is stronger in spontaneous information processing (i.e. in the case of short response latencies) and second, that acquiescence bias is stronger in the case

of spontaneous information processing when chronic attitude accessibility is low (i.e. when response latencies are short and attitudes are cognitively not accessible). Two test strategies were applied: first, a split ballot survey experiment was conducted to test the first proposition at the aggregate level. Second, a multiple group structural equation model was estimated to test both propositions at the individual level.

The data used for this analysis was gathered in a German nation-wide random sample CATI-survey in 2005 (n=2002).[3] Reaction times were measured actively by interviewers in hundredths of seconds. The 'raw' reaction times were regressed on a baseline speed measure and the residuals were used as a measure of response latency (so called 'residual index', see Mayerl and Urban 2008 for more details). In addition, latencies were controlled for timer validation by interviewers and statistical outliers were removed (cut at ±2 standard deviations above/below the mean).

Split ballot experiment: Analysis on aggregate level

A split ballot experiment was conducted in which the wording of a three-indicator need for cognition scale (items according to Keller et al. 2000) was varied, differentiating between two conditions: positive/original wording (condition A) versus negative/opposite wording (condition B; n=250 for each condition). The analysis focused on examining the proportion of respondents in condition A agreeing to the positively worded items versus those in condition B disagreeing with the negatively worded ones. For the analysis, the negatively formulated items from condition B were recoded to match the response scale from condition A (see Mayerl 2013 for more details on this study).

Table 1 reports the percentage of respondents agreeing to each item, divided into two groups: fast vs. slow response latencies (median split).

3 DFG-funded project "Response latency measurement in survey research. Analyzing the cognitive basis of attitudes and information processing", principal investigator Prof. Dr. Dieter Urban, University of Stuttgart

Table 1 Mode of information processing and acquiescence effect

	automatic-spontaneous mode (response latency ≤ median)			deliberative-controlled mode (response latency > median)			moderator effect of response mode (%)
	A: „original“ (in %)	B: „negative“ (recoded) (in %)	difference (in %)	A: „original“ (in %)	B: „negative“ (recoded) (in %)	difference (in %)	
Item1)							
original: I like abstract thinking.							
negative: I do *not* like abstract thinking.							
1-2 agreement	59.8	31.4	28.4	30.2	46.8	-16.6	45
3 undecided	17.4	25.6	-8.2	33.3	19.4	13.9	22.1
4-5 disagreement	22.7	43.0	-20.3	36.5	33.9	2.6	22.9
*ANOVA: Condition („C“; 1: A; 0: B): p=.470; Response latency („RL“; 1: fast, 0: slow): p=.243; Interaction C*RL: p=.000; N=438*							
F=5.940; df=3; p=.001							
Item2)							
original: I like to think for hours about something.							
negative: I do *not* like hours of thinking about something.							
1-2 agreement	56.4	42.5	13.9	34.0	48.8	-14.8	28.7
3 undecided	18.8	28.7	-9.9	39.2	22.0	17.2	27.1
4-5 disagreement	24.8	28.7	-3.9	26.8	29.3	-2.5	1.4
*ANOVA: Condition („C“; 1: A; 0: B): p=.553; Response latency („RL“; 1: fast, 0: slow): p=.268; Interaction C*RL: p=0.041; N=440*							
F=1.874; df=3; p=.133							

Table 1 continued

	automatic-spontaneous mode (response latency ≤ median)			deliberative-controlled mode (response latency > median)			moderator effect of response mode (%)
	A: „original“ (in %)	B: „negative“ (recoded) (in %)	difference (in %)	A: „original“ (in %)	B: „negative“ (recoded) (in %)	difference (in %)	
Item3)							
original: I think it is exciting to learn new ways of thinking.							
negative: I think it is *less* exciting to learn new ways of thinking.							
1-2 agreement	70.4	57.9	15.8	48.3	51.9	-4.8	20.6
3 undecided	23.2	14.5	7.0	37.9	18.6	20.7	13.7
4-5 disagreement	6.3	27.6	-22.3	12.6	29.5	-15.9	6.4
*ANOVA: Condition („C“; 1: A; 0: B): p=.203; Response latency („RL“; 1: fast, 0: slow): p=.000; Interaction C*RL: p=.088; N=433*							
F=7.495; df=3; p=.000							

Source: Mayerl 2013, p. 12

As shown in Table 1, the mode of information processing acts as a significant moderator of acquiescence bias for all three items. As the results for item 1 show, in the case of fast responders, 59.8% agreed to the positively worded item, but only 31.4% disagreed to the opposite, leading to a difference of 28.4% – this is the acquiescence effect at the aggregate level for item 1. In contrast, slow responders in condition A agreed less to the positive wording (30.2%) than slow responders in condition B disagreed to the negative wording (46.8%) – this means there was no evidence of acquiescence amongst slow responders (on the contrary, slow responders tend more towards disagreement). The same pattern is true for the other two items.

In terms of statistical significance, the interaction between the randomized condition (positive vs. negative wording) and response latency was statistically significant for items 1 and 2 ($p<.05$) and marginally significant in the case of item 3 ($p<.10$). The finding that the main effect of the randomized questionnaire version is non-significant in all three cases underlines that acquiescence bias is true for fast responders only. In sum, the results presented in Table 1 support the first proposition that acquiescence bias occurs in the spontaneous response mode only.[4]

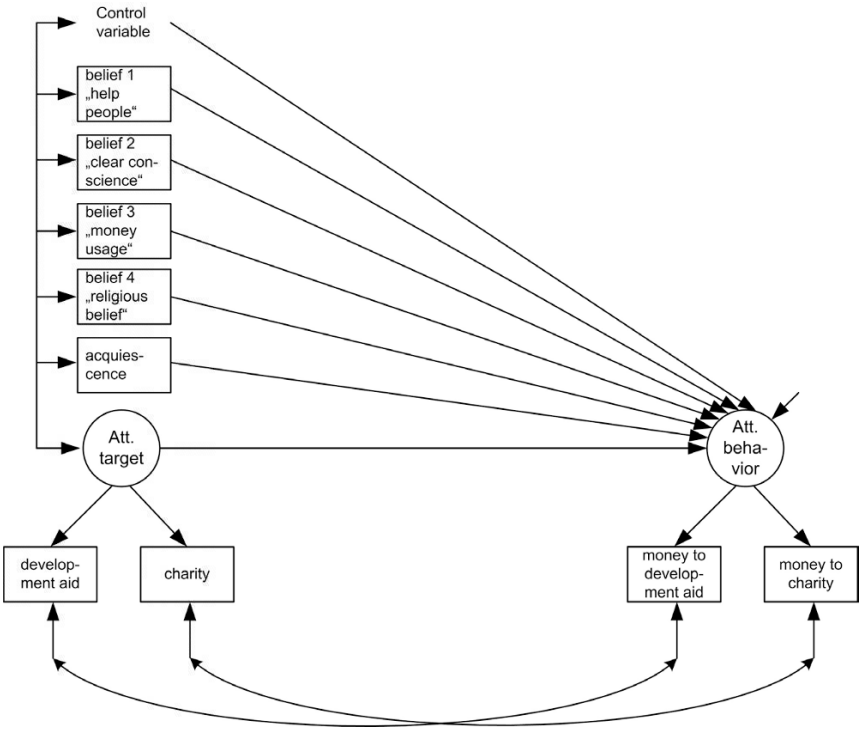
Structural Equation Models: Analysis at the individual level

Another way of analyzing acquiescence bias is to use a score of the individual's general tendency towards acquiescence. Mayerl (2009, 2013) operationalized such a scale of tendency to agree ("acquiescence") by computing the sum of "totally agree"-answers on 100 items (all of which used a 5-point rating scale) covering different item contents (resulting in a scale with values ranging from 0 to 100).

Figure 1 shows the structural equation model with the predictor variable acquiescence score and, in line with Aizen and Fishbein (1980), generalized attitudes and behavioural beliefs influencing specific behavioural attitudes

4 Interestingly, when analyzing acquiescence bias without distinguishing between fast and slow responders, we found significantly more agreement to the original wording as opposed to the reversed wording for items 1 and 3. In contrast to the findings reported in Table 1, however, there was no significant difference in the case of item 2 ($p=0.608$, see Mayerl 2013, p.10). Thus, taking response latencies into account may help to identify response effects that would have been overlooked when using classical techniques of analyzing survey data.

towards donating money to charity (see Mayerl 2013 for more details on specification, operationalization and estimation of the shown multiple group structural equation model). In order to properly interpret the results, it is important to note that the items of both attitude constructs were 5-point rating scales with 1=agree and 5=disagree. As such, the acquiescence scale was expected to have a negative effect on attitudes towards behaviour.



Source: Mayerl 2013, p. 15

Figure 1 Structural equation model explaining attitude towards donating money

In a first step, a two-group structural equation model was estimated with fast versus slow respondents (see Table 2).[5] In line with the first proposition, the acquiescence score only had a significant effect on behavioural attitudes in the case of spontaneous information processing. Interestingly, the effect of the generalized attitude towards the target (charity organization) was significantly stronger in the case of spontaneous processing (χ^2 difference test: $p<.05$).

Table 2 Selected output of a 2-group model (two response modes; metric invariance)

	automatic-spontaneous mode (resp. latency ≤ median; N = 557)			deliberative-controlled mode (resp. latency > median; N = 557)		
	b	SE	t	b	SE	t
Structural effects on Attitude towards behavior:						
acquiescence	-.061**	.021	-2.889	-.025	.022	-1.123
attitude towards target	.799**	.111	7.218	.530**	.108	4.883

Y = attitude towards donating money for charity organizations; ** $p \leq .01$; * $p \leq .05$; + $p \leq .10$; without marker: n.s. with $p > .10$; χ^2=39.054; df = 28; p = .080; CFI = .993; RMSEA = .027 ($CI_{0.90}$: .000 to .045); SRMR = .020; Control variable: measure of altruistic motivation; Source: Mayerl 2013: 16

In the next step, attitude accessibility was introduced as a second moderator variable, leading to a four-group structural equation model (2 latency groups * 2 accessibility groups; see Mayerl 2013, p. 18f.). The chronic attitude accessibility of the attitude towards the target (i.e. towards charity organizations) was operationalized by membership in a charity organization, assuming that members have a higher direct experience with that attitude object.[6] As a result, and in line with the second proposition, the acquiescence score indeed had a significant effect only under a very specific condition: when a respondent answers in a spontaneous way *and* she or he has *no* chronically acces-

5 Fast vs. slow respondents were operationalized using a median split of the mean response latencies for both indicators of the behavioural attitude construct.

6 Direct experience is a well-known measure of chronic attitude accessibility in social psychology, see Fazio 1986, 1990a; Fazio et al. 1982.

sible overall judgment (i.e. attitude). In the other conditions, the acquiescence score does not show a significant effect on attitude statements. Interestingly, and again in line with dual process theory, a second finding was that the effect of generalized attitudes is significantly strongest in the case of spontaneous responses with highly accessible attitudes (χ^2 difference test: $p<.05$).

Such analyses demonstrate the usefulness of paradata in gaining a deeper understanding of the conditions under which measurement problems like acquiescence bias occur. In addition, it was shown that paradata help to identify subsamples of respondents for whom attitudes are strong predictors as opposed to subsamples with less predictive attitudes.

3.2 Question Order Effects for Fast and Slow Responses

It is a well-known phenomenon that answering survey questions is not just dependent on individuals' attitudes and beliefs, for example, but also on the context of a survey situation. Such context effects can arise because of the specific content of a question, the response scale and/or the preceding question. It is called a question order effect if such a preceding question influences the response behaviour on the following question (Strack and Martin 1987).

To examine question order effects, it is helpful to consider the process in which respondents answer survey questions. In a first step, respondents read and understand the question. Next, respondents either search their memory for a pre-existing opinion or retrieve the relevant information in order to form an opinion on the spot (Cannel et al. 1981; Strack and Martin 1987; Sudmann et al. 1996; Tourangeau and Rasinski 1988). If such opinions are formed on the spot, respondents usually will not retrieve all the necessary information but only those that are easily accessible, especially if a respondent's motivation to answer in a deliberative way is low and/or time pressure is high (Tourangeau and Rasinski 1988). Regarding information accessibility, Strack et al. (1987) showed that a previous question (the so-called context question) can activate a specific set of information which is then easily accessible when answering a following question (the target question), leading to a question order effect.

In addition, Strack et al. (1987) distinguish between assimilation effects and contrast effects of question order. An assimilation effect is given if the judgment of the target question shifts in the direction of the context question due to the accessibility of the prior activated information and a lack of

motivation and/or opportunity to retrieve new or additional information to form a better suiting answer. On the other hand, a contrast effect is given if the judgement of the target question shifts in the opposite direction of the context question due to the so-called "given-new-contract" (Clark 1985): in a natural conversation, people try to add new information instead of repeating what is already known. For example, if one is asked about the well-being of his wife and then about his family, the person would most likely subtract the information about his wife within the answer to the question about the well-being of his family, given that the person with whom he is speaking already knows how his wife is. Furthermore, Strack et al. (1987) examined under which circumstances assimilation and contrast effects appear. In conclusion, they found that contrast effects come out stronger if successive items are perceived as belonging together (e.g. if only the context and the target question share the same headline) due to the aforementioned "given new contract". Assimilation effects appear if the bond between successive items is perceived as rather week (e.g. if successive items are only part of the same question battery alongside some other items).

Therefore, it is obvious that assimilation effects require very low mental effort. A previously activated piece of information can simply be reused, no new information needs to be retrieved, no new opinion needs to be formed (Tourangeau et al. 1989). Contrast effects, on the other hand, require more mental effort because a respondent first needs to form an opinion and then subtract the already given information from that opinion (Schwarz and Strack 1999). The dual-process model provides the framework for incorporating response latencies as a measure of mental effort (see section 1). Within this framework, we can assume that assimilation effects, which require low levels of mental effort, occur quickly within an automatic-spontaneous response mode while contrast effects, which require higher levels of mental effort, occur more slowly within a deliberative-controlled response mode.

To examine this assumption, correlations between successive item pairs (i.e. the context item and the target item) are analyzed within an experimental split ballot design. The experimental group received a set of four questions on a 7-point rating scale with a fixed question order, while the control group received the same set of questions but with a random question order. This results in three potential question order effects (between the first and the second, the second and the third and the third and the fourth item). The four

questions[7] are part of an item battery on "pro environmental attitudes". The first and the third item (within the fixed question order design) are positively worded, the second and the fourth item are negatively worded (agreement indicates a less positive environmental attitude). Those two negative items were recoded before all analyses (with a scale ranging from 1: 'weak pro environmental attitude' to 7: 'strong pro-environmental attitude' for all four items after the recoding). The data stems from a 2012 random web access sample with a total of 883 participants.

To test the assumption that assimilation effects mainly occur for fast responses while contrast effects mainly occur for slow responses, we examined the correlations within a 2x2 (fixed vs. random question order * fast vs. slow responses)[8] research design.

If assimilation effects appear, the covariance between pairs of sequentially ordered items should be higher under the condition of fixed question order compared to a random question order (i.e. $Cov(a \Leftrightarrow b)_{fixed} > Cov(a \Leftrightarrow b)_{random}$). On the other hand, if contrast effects appear, the covariance between sequential items should be lower compared to randomized question order (i.e. $Cov(a \Leftrightarrow b)_{fixed} < Cov(a \Leftrightarrow b)_{random}$). If our assumption is true that these question order effects are moderated by response latency, the correlation between the context question and the target question for fast responses should be stronger within the fixed-order design group than in the random-order design group (assimilation effect). Further, the correlation between the context and the target question for slow responses should be weaker within the fixed-order design group than in the random-order design group (contrast effect).

---

7 Question a: "When reading newspaper articles on environmental problems, or when watching corresponding telecasts, I often become indignant and angry."
Question b: "I am not worried when I think about the environmental conditions under which our children and grandchildren will have to live." (rating scale recoded for analysis)
Question c: "In favor of the environment, all of us should be willing to cut down on our standard of living."
Question d: "Politicians are doing enough to protect the environment." (rating scale recoded for analysis)

8 Using a median-split of the response latency residual-index as described in chapter 3.1

Table 3 Correlation coefficients for question order effects

	automatic-spontaneous mode (response latency ≤ median) (n=141)			deliberative-controlled mode (response latency > median) (n=146)			
correlated items	A: fixed (n=54)	B: random (n=87)	Response effect: Δ=A-B	A: fixed (n=84)	B: random (n=62)	Response effect: Δ=A-B	Moderation by response latency?
a ⇔ b	r=.137$^{n.s.}$ (b=.194)	r=.117$^{n.s.}$ (b=.169)	Δr=.020 (Δb=.025)	r=.163$^{n.s.}$ (b=.211)	r=.372*** (b=.442)	Δr=-.209 (Δb=-.231)	yes
b ⇔ c	r=.584*** (b=.419)	r=.180* (b=.129)	Δr=.404 (Δb=.290)	r=.261** (b=.199)	r=.696*** (b=.625)	Δr=-.435 (Δb=-.426)	yes
c ⇔ d	r=.252* (b=.279)	r=.265** (b=.268)	r=.013 (Δb=.011)	r=.067$^{n.s}$ (b=.063)	r=.370*** (b=.390)	Δr=-.303 (Δb=-.327)	yes

r=Pearson's correlation coefficient; b=unstandardized regression coefficient;
*** p ≤ .001; ** p ≤ .01; * p ≤ .05; n.s. p > .05

As seen in Table 3, contrast effects of question order (a negative correlation-difference between the fixed and random design groups) appear only for slow responses. This is true for all successive item pairs (a-b, b-c, and c-d). An assimilation effect of question order (a positive correlation-difference between the fixed and random design groups) is given for the correlation between items b and c, while the correlation-difference between items a and b as well as between items c and d is close to zero. This means that contrast effects are indeed observable for slow responses, while assimilation effects tend to be associated with fast responses.[9]

In conclusion, the findings presented here suggest assimilation is indeed a satisficing strategy within an automatic-spontaneous response mode, while contrast effects, requiring more mental effort, occur in a deliberate-controlled one. The findings highlight the usefulness of response latencies in uncovering question order effects. Further research is needed, however, to gain a deeper

9 Interestingly, Mayerl and Giehl (2018) and Mayerl and Urban (2008) found comparable results in similarly structured studies based on CATI data. The results of those studies, along with the findings presented here may suggest, however, that CATI surveys reinforce assimilation, CASI surveys contrast effects. For a discussion on possible explanations for this, see Strack et al. (1987).

understanding of the processes and the mechanisms underlying these kinds of effects, especially when taking mode effects into account.

3.3 Social Desirability Bias

Social desirability (SD) bias describes self-reported survey responses that present the respondent in a more favourable light than is actually accurate. It is seen as a major source of systematic measurement error and affects both prevalence estimates and observed relations between constructs.

Some researchers see SD bias as the result of a stable personality trait in that some people just generally feel a stronger 'need for social approval' than others. To this end, many researchers have attempted to construct innocuous scales meant to measure this personality trait and use it to adjust estimates in multivariate models (i.e. the Edwards Social Desirability Scale by Edwards 1957, the Marlowe-Crowne Social Desirability Scale by Crowne and Marlowe 1960, the Other- and Self-Deceptions Questionnaires by Sackeim and Gur 1978 and the Balanced Inventory of Desirable Responding (BIDR) by Paulhus 1984 to name just a few, see Uziel 2010 for a concise overview). However, while some interesting work has been done to show how these scales may actually be tapping into different types of mostly pathological personality traits (neuroticism, self-consciousness etc., see McCrae and Costa 1983) they often fail at their original task of correcting biased estimates. Some have argued that the early scales in particular did not accurately reflect the underlying factorial structure (i.e. Wiggens 1964; Paulhus 1984, 1991, 2002) and have suggested more elaborate scales sometimes differentiating between overt lies (impression management) and more subtle misreporting based on honest but inaccurate beliefs (self-deception).[10]

A more fundamental criticism of these SD scales revolves the idea that SD bias should not even be conceptualized as a personality trait and that it is rather the result of more situational factors. As Tourangeau and Yan (2007) summarize, SD bias could be the result of either or both a response style (consistent with the personality trait argument outlined above) or a short-lived and situationally-motivated response set. Researchers such as Stocké

10 Although recently even more elaborate factorial structures have been suggesting which further differentiate between egoistically and 'socially' motivated misreporting, see Uziel 2010 and Paulhus 2002 for overviews.

(2004) have put forth the idea that SD bias is the result of a complex interaction involving the respondent's underlying tendency to report in a socially desirable fashion ('need for social approval'), the subjective sensitivity of the survey question and thus the desirability of the individual response possibilities ('desirability belief' or 'trait desirability')[11] as well as the interview situation itself (i.e. in anonymous surveys, from whom can one expect to gain approval for their response?). Furthermore, as Tourangeau and Yan (2007) point out, to the extent to which the likelihood of receiving a biased response is increased by the perceived sensitivity of the survey question, the potential 'true' answers of the respondents are both decisive and elusive: a question about voting is only sensitive to a respondent that did not vote (p. 860).

For the aforementioned reasons, identifying and correcting SD bias often involves a) the inclusion of SD scales, b) estimating the sensitivity of individual survey items and c) either incorporating the anonymity (objective or perceived) as a control variable in multivariate analyses or actually manipulating the anonymity in experimental designs. This last point has been the focus of a large body of research on SD responding. Experimental techniques such as the randomized response technique (RRT), and item or person count technique (ICT and PCT) have been developed to investigate the effect of full anonymity on SD bias. The premise that links these experimental techniques is to provide an experimental group with full anonymity to report on a sensitive question. Research on the effectiveness of these techniques is, however, ongoing, and in order to achieve full anonymity, the examination of SD bias is often limited to the aggregate level (for more on these techniques, see Holbrook and Krosnick 2010a and b; Wolter 2012; Wolter and Preisendörfer 2013).

As such, there is continued interest in developing new ways of examining SD bias. One novel approach, and the focus of the remainder of this section, is the use of non-reactive paradata in the form of response latencies to identify socially desirable responding.

With regards to SD bias, researchers have formulated a variety of hypotheses. Strack and Martin (1987) Tourangeau and Rasinski (1988), Bassili (2003) and Holtgraves (2004) have discussed SD bias in the form of an 'editing' process. Respondents go through all the normal steps of answering a survey question, i.e. reading and understanding the question, either a) searching

11 See Wolter 2012 for a discussion on the operationalization of desirability beliefs and trait desirability.

their memory for pre-generated opinions, or b) retrieving relevant information and forming an opinion on the spot, and then translating their response according to the available scale values. Once the respondent has gone through the process of coming to their 'true' answer, they may then decide to edit their response to gain or avoid social (dis)approval. This final editing step is hypothesized to take time; so, responses biased by SD should take longer according to this line of reasoning. This hypothesis describes the idea of deliberately misleading the interviewer (or any other potential audience) and is closely related to the impression management conception of SD.

Other researchers have argued more explicitly within the dual-process framework and highlight the idea that SD responses may be able to happen very quickly. Researchers such as Kohler and Schneider (1995), Amelang and Müller (2001), and Holtgraves (2004) have discussed the idea of SD responding in terms of a 'biased retrieval' of relevant information. This idea is closely linked with Krosnick's weak satisficing strategy: the respondent, for various reasons, performs only a quick, perfunctory information retrieval. Because people generally have positive view of themselves, if their retrieval is confirmatory, they will quickly find several pieces of information to support their potentially self-deceiving positive outlook. An even more extreme scenario is plausible in which strong satisficing leads to not just a biased retrieval stage but a so-called 'retrieval skip' in which the socially desirable response is obvious to the extent that the respondent has so little motivation (or opportunity) to consider their answer that an SD response becomes as easy and quick as acquiescence (DePaulo et al. 2003; Holtgraves 2004).

Stocké and Hunkler (2004, 2007) suggest that the desirability of the question may play a role in determining which of the aforementioned scenarios play out. They discuss the non-linear nature of the subjective desirability of survey questions: with increasing perceived desirability or undesirability, it becomes clearer to the respondent which response is the most desirable or most undesirable.[12] To a respondent with little motivation or opportunity to respond in a deliberate fashion, it should be possible for them to answer very

12 The question as to whether the hypothesis applies symmetrically to both positively and negatively keyed items (i.e. does most undesirable equal least desirable and does least undesirable equal most desirable?) is still the focus of research (see Paulhus 1984 for an overview of the discussion). The results of the analysis presented here also touch upon this question.

quickly in a socially desirable way. Where the desirability is less clear, the motivation to respond deliberately may increase for fear of answering in a way that is 'unacceptable'. Thus, their hypothesis focuses on the trait desirability of the survey questions and the non-linear fashion in which they affect response latencies.

In 2017 we attempted to empirically examine the discussions outlined above. In a research project conducted from 2014-2016, which focused substantively on explaining pre-service biology and chemistry teachers' behaviour with regards to experimenting in the classroom, we included an item battery of 30 questions asking the respondents to report the extent to which they were suited for their chosen profession (Andersen and Mayerl 2017)[13]. 550 tablet-computer-based CASI surveys were conducted. Response latencies were recorded actively for every tap on the screen along with extensive information regarding the nature of the 'event' (entering a response, selecting to continue onto the next page, going back to the previous page, changing a response, etc., for more detailed information, see Andersen and Mayerl 2017). Furthermore, we measured the desirability of the individual items using an external survey and included a shortened version of the Crowne and Marlowe SD scale. We were primarily focused on examining the question of whether socially desirable responses were linked to faster or slower response latencies as the dependent variable.

The results of the study can be found below in Table 4. The most relevant results for this discussion involve the item's trait desirability and the respondent's need for social approval. Contrary to the hypothesis by Stocké and Hunkler (2004, 2007), we do not observe a non-linear relationship between the item desirability and response latencies (squared term non-significant). Instead, we see that the main effect for trait desirability (measured on a bipolar scale from -4: very undesirable to +4: very desirable) is significant and negative (-.071, $p<.01$). For desirably rated items (on the positive side of the scale, > 0), responses become faster as the desirability increased. For the undesirably rated items (on the negative side of the scale < 0), responses actually became slower as the undesirability increased. Figure 2 displays the effect of trait desirability on response latencies (based on Andersen and Mayerl 2017).

13 "EVA3PLUS", funded by the state Ministry of Education, Science, Further Education and Culture Rhineland-Palatinate (now Ministry of Education and Ministry of Science, Further Education and Culture).