Comparing chatbots and online surveys for (longitudinal) data collection:

An investigation of response patterns, data quality and user evaluation

Brahim Zarouali [a]
+32 32655049
Brahim.zarouali@kuleuven.be

Theo Araujo [b]
T.B.Araujo@uva.nl

Jakob Ohme [c]
j.ohme@fu-berlin.de

Claes de Vreese [b]
C.H.deVreese@uva.nl

Affiliations:

[a] Institute for Media Studies, KU Leuven, Parkstraat 45, 3000 Leuven, Belgium

[b] Amsterdam School of Communication Research, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands.

[c] Weizenbaum Institute for the Networked Society, Freie Universität Berlin, Hardenbergstraße 32, 10623 Berlin, Germany

Funding: This work was funded by the Digital Communication Methods Lab (RPA Communication) of the University of Amsterdam

Disclosure statement: No potential conflict of interest was reported by the authors.

Comparing chatbots and online surveys for (longitudinal) data collection:

An investigation of response characteristics, data quality and user evaluation

Abstract

As chatbots are gaining more popularity than ever, they have recently been considered as interesting tools for survey administration in social science research. To explore this idea, we investigated the extent to which there are differences in response characteristics and data quality between a traditional, web-based survey and a conversational, chatbot-based survey (which we integrated in an instant messaging app). In addition, we zoomed into how respondents evaluate both survey modes. By using a longitudinal design, we also explored how response characteristics evolved over a period of two-weeks. Overall, we did not find evidence that chatbots might be better survey administration tools than web surveys. On the contrary, the web survey often seemed to generate more favorable response characteristics and data quality. Finally, when it comes to user perceptions, we found that the chatbot survey was evaluated less favorably in terms of perceived enjoyment, usefulness, and security. Based on these results, we draw conclusions about whether chatbots can be considered as valid alternatives for traditional web survey methods.

Introduction

Chatbots are gaining increased importance in communication research and practice, enabled by recent advances in computational techniques. Chatbots can be defined as "software application that accepts natural language as input and generates natural language as output, engaging in a conversation with the user" (Griol, Carbó, & Molina, 2013, p. 355). These conversational agents are usually operating via popular instant messaging platforms (e.g., Facebook Messenger, WhatsApp, Skype, LinkedIn, etc.). They interact automatically with users about a wide variety of subjects, including health (Bell, Wood, & Sarkar, 2019; Bickmore & Gruber, 2010), marketing and customer service (Araujo, 2018; Verhagen, van Nes, Feldberg, & van Dolen, 2014; Zarouali et al., 2018) and news (Jones & Jones, 2019; Zarouali et al., 2020). While most studies about chatbots investigated the antecedents and consequences of communicating with these agents, communication scholars are yet to fully explore the implications of these chatbots as *data collection engines*.

Using such conversational interfaces for survey design might be interesting, based on the idea that more natural and conversational interactions can occur between the participants and the survey tool, hereby increasing the response quality and user satisfaction (Celino & Calegari, 2021). Moreover, earlier research indicates that users perceive chatbots as a more anonymous way of getting personal (health-related) information than other modes of communication such as telephone services or search engines (Crutzen et al., 2011), and that people may be willing to engage in high levels of personal self-disclosure with chatbots as interaction partners (Lee et al., 2020). At times, this has to do with chatbots triggering machine-like perceptions - thus being perceived as less personal than disclosing to a human (Sundar & Kim, 2019), and at times, due to them triggering human-like perceptions when compared to other technology (Ischen et al., 2020).

Against this background, chatbots may become an alternative to web surveys (Kim, Lee, & Gweon, 2019), and could be used to collect numeric and text data in an automated, large-scale manner (Araujo, 2020). However, little empirical knowledge is available about the effectiveness of using chatbots in existing messaging platforms for collecting quantitative and qualitative social science data. The goal of this study is to provide evidence of the methodological promises and pitfalls of chatbots as a data collection tools by directly comparing them to "traditional" web surveys. Web surveys were chosen because they are the most popular mode for collecting survey data (compared to other modes, such as paper-and-pencil surveys, or telephone surveys), as well as the most efficient approach -in terms of time, costs, etc.- to recruit large amounts of people (Dillman et al., 2014). Therefore, it would be most relevant to compare the chatbot survey mode to this web-based survey mode.

This was done by conducting a longitudinal, pre-registered study about news exposure among Dutch users (N = 304). Respondents answered daily questions about their news consumption during a period of 14 days either via an online survey or a chatbot embedded in a messaging platform. Based on these data, we explored differences in *response characteristics*, *data quality* and *user evaluations* of chatbots vs. online surveys. The analyses in this study involve data on two different levels: i) cross-sectional data to look into the methodological differences between chatbots and online surveys at one point in time (day 1); ii) longitudinal data (data from the daily surveys from day 1 - day 14) to zoom into how some methodological differences between the two survey modes (chatbot vs. web survey) change over time.

Altogether, this study aims to compare chatbots to online surveys as data collection tools by looking at response characteristics and data quality variables. This will be done at one point-in-time (cross-sectional design), as well as over a period of 14 days (longitudinal design). In addition, we zoom into the differences in respondents' evaluation of both survey

modes in terms of utilitarian and hedonic features, perceived security, and cognitive load after the 14-day period. With this study, we contribute to the methodological discussion of whether chatbots can serve as reliable survey administration tools that might lead to higher data quality and lower response biases, and as such, can be used as an effective alternative to more traditional methods such as web surveys.

Literature review

Background: using chatbots as data collection engines

In the 1960s, Joseph Weizenbaum (1966) developed ELIZA, which can be considered to be the first chatbot in the history of computer science. This chatbot was designed in a way that it mimics human conversation based on simple scripted responses. Since this early prototype, chatbots have come a long way and are now used by millions of businesses all around the globe (Parnham, 2021). They have gained a lot of interest since many big tech platforms (e.g., Facebook, Microsoft) allowed chatbots to be integrated in their instant messaging services (e.g., Facebook Messenger, Skype). To understand how chatbots work via these messaging platforms, it is important to stress that chatbots need to perform three different activities: (1) apply natural language understanding techniques to process what the user has typed (e.g., analyze and categorize what the user has said or typed, extract relevant entities, etc.), (2) use a dialogue management function to identify how the user input relates to a general dialogue and context (e.g., to situate what the user has said in a larger dialogue context), and (3) use natural language generation techniques to provide an appropriate response (e.g., identify and generate a suitable response to what the user said) (Griol et al., 2013).

Over the past few years, scholars have showed a growing interest in using chatbots as survey administration tools (e.g., Celino & Re Calegari, 2020; Kim et al., 2019; Xiao et al., 2020). These chatbots allow participants to have a more natural and conversational interaction

with the survey tool. In addition, they are usually being deployed on widely used instant messaging platforms, such as Facebook Messenger (e.g., Van den Broeck et al., 2019; Zarouali et al., 2018), which can increase convenience for participants because they can take surveys in familiar environments where they already spent a lot of time (Kim et al., 2019; Araujo, 2020). A chatbot survey is based on the idea that the machine asks survey questions, and the user replies directly via chat. During this process, the chatbot interprets all the user answers and dynamically responds to them (usually by going to the next survey question). Although conversational surveys can be developed based on natural language understanding and artificial intelligence (e.g., Xiao et al., 2020), most chatbot surveys (including the one used in this study) will simply be designed based on a pre-programmed conversation flow through a chat interface containing all the survey questions.

In the next section, we zoom into the methodological promises (and pitfalls) of using chatbots in administering questionnaires and discuss how these conversational agents emerged as a response to the limitations of using online web surveys.

Online web surveys vs. chatbots in survey administration

Online web surveys are widely used as a survey mode to collect quantitative (closed, numeric answers) and qualitative (open answers) in social sciences (Ha & Zhang, 2019). The many advantages of collecting these data via online surveys have been well-documented. For instance, compared to more traditional survey modes (e.g., telephone survey, a paper-and-pencil survey, face-to-face survey), web surveys generate higher response speed, have a lower cost per response, reduce item non-response rate, allow to reach large amounts of respondents, elicit higher data quality, increase data consistency, generate higher levels of personal self-disclosure, reduce social desirability effects, etc. (e.g., Denscombe, 2009; Fricker et al., 2005; Greenlaw & Brown-Welty, 2009; Griffis et al., 2003; Hanna et al., 2005; Heerwegh, 2009; Shin et al., 2012). However, an online self-administered web survey may

also suffer from inevitable limitations, such as lower response rate, higher non-response bias, increase in measurement errors, heightened concerns of privacy, underrepresentation of respondents without internet access, etc. (e.g., Daikeler et al., 2020; Denniston et al., 2010; Dillman et al., 2014; Felderer et al., 2019; Ha & Zhang, 2019; Manfreda et al., 2008). Despite these limitations, web questionnaires remain a very suitable and effective data collection mode for survey research.

Quite recently, scholars have argued that using chatbot technology might improve the quality of survey research even more because of the various advantages of these tools. More precisely, chatbot survey might lead to an reduction in experimenter demand effects (Wijenayake et al., 2020), a decrease in satisfying effects (Celino & Re Calegari, 2020; Kim et al., 2019), an increase in self-disclosure (Lee et al., 2020) and a higher response quality of open answers (Wambsganss et al., 2020). In addition, chatbots combine the convenience of cheap and quick digital data collection (both numeric and text data) on the one hand, and on the other, perform the role of a human interviewer because of the conversational nature of these interfaces (Celino & Re Calegari, 2020; Kim et al., 2019). This means that chatbots can communicate with participants by adopting social scripts and cues that make them appear similar to human-to-human communication (Araujo, 2018; Edwards et al., 2016). However, the latter also highlights a potential shortcoming of chatbots in survey research: the human-like nature of conversational agents might elicit socially desirable answers (more so than with web surveys) (e.g., Schuetzler et al., 2018).

Nevertheless, a conversational survey approach still appears to be a promising development that can lead to survey quality benefits (see earlier), allows to maximize convenience by reaching large numbers of participants on platforms where they spend most of their time (e.g., FB Messenger, WhatsApp, Skype), and yield a more natural survey experience because of the human-like nature of the technology. However, as of yet, limited

knowledge is available about the differences between web and chatbot surveys (in cross-sectional designs), and we know even less about how these modes of data collection differ over time (i.e., in longitudinal designs), which is an important question since chatbots might have the ability to build relationships with participants as time progresses (Wijenayake et al., 2020). Therefore, the overall research aim of this study is to *explore the extent to which there are (cross-sectional and longitudinal) differences in response characteristics and data quality between traditional, web-based surveys and conversational, chatbot-based surveys, and in addition, examine how people evaluate their interactions with these two survey modes.*

Chatbots in cross-sectional survey research

Studies begin to investigate the direct comparison of chatbots and web surveys (e.g., Celino & Re Calegari, 2020; Kim et al., 2019). Using chatbots as research tools has been considered a promising approach because of their increased interactivity. *Interactivity* refers to the degree to which a person considers communication to be bi-directional and occurring in real-time (Kang et al., 2014). In the case of chatbots, respondents engage in a series of back-and-forth exchanges with a chatbot (mimicking a real conversation), which leads to a higher level of interactivity (as compared to a self-administered web survey) (Araujo, 2018). This can increase the participants' engagement with the research. A recent study by Xiao et al. (2020) found that chatbots drove a significantly higher level of participant engagement and elicited significantly better-quality text data (open-ended responses), in terms of their informativeness, relevance, specificity, and clarity. Other studies found that a conversational chatbot survey can reduce satisficing effects in numeric data (Celino & Re Calegari, 2020; Kim et al., 2019), and that they can elicit open answers with a higher readability score (Wambsganss et al., 2020). Another important advantage of chatbots is the fact that they contain anthropomorphistic cues (i.e., human-like characteristics), which can potentially encourage respondents to reveal more personal information (Xiao et al., 2020). In this regard,

chatbots have been found to prompt open answers with an increased participant self-disclosure (Lee et al., 2020), as well as responses with a higher intensity of emotional sentiment (Wambsganss et al., 2020). This personal and sentimental self-disclosure might be the result of chatbots' ability to establish a social bond with respondents while simultaneously allowing for anonymity (Wijenayake et al., 2020).

Although these recent studies have explored initial differences in the effectiveness of chatbots and web tools in administering surveys, much is still to be understood about the differences in i) response characteristics and ii) data quality between the two modes.

Response characteristic variables refer to methodological constructs that indicate how and to which extent responded to the survey. Some important response characteristic variables have not been investigated (yet) in the context of chatbots, such as *response rate* (Baruch & Holtom, 2008), *response time* (Börger, 2016) and *length of the open-ended responses* (Denscombe, 2008; Smyth et al., 2009). *Response rate* is considered an important component for the quality of survey-based research (e.g., Curtin et al., 2000; Groves & Peytcheva, 2008; Sammut et al., 2021). The response rate of a survey provides an indication of response bias and the consequent representativeness of the results (Sammut et al., 2021). Low response rates in survey research may result in samples that are unrepresentative of the larger populations and can lead to an increased likelihood of nonresponse bias (Fosnacht et al., 2017; Groves & Peytcheva, 2008; Pedersen & Nielsen, 2016), i.e., the bias that exists in the data when respondents to a survey are different from those who did not respond in terms of demographic or attitudinal variables (Sax et al., 2003). A high survey response rate is thus crucial to promote the validity of survey-based research findings (Groves & Peytcheva, 2008; Harris, 2010; Pedersen & Nielsen, 2016). A higher *response time* has been shown to have a number of advantages in survey research. For instance, longer durations might be indicative for lower choice randomness, increased precision of research estimates, and a decreased level

of primacy effects (Börger, 2016; Malhotra, 2008). An increased length of open-ended responses has also been linked to greater response quality, because the more words a respondent uses to answer an open-ended question, the more detailed the response will be, and thus, the more useful the information should be for researchers (Barrios et al., 2011; Israel, 2010).

Data quality variables are indicators which can potentially be used to evaluate the participants' response quality. This study will focus on three important data quality variables: *internal consistency* (Henson, 2001), *attention check* (Shamon & Berning, 2019), and *data variability* (Galesic & Bosnjak, 2009; Santesso et al., 2020). Internal consistency refers to the interrelatedness or cohesiveness of a set of items and is considered a prerequisite to assure the accuracy and quality of a measurement instrument (Chau, 1999; Kimberlin & Winterstein, 2008; Schmitt, 1996). Next, it has been found that including an attention check question (e.g., *"please select -strongly disagree- for this item"*) is very suited to identify careless respondents (e.g., respondents that speed through the survey) and exert a motivational influence on answering behavior, hereby contributing to data quality without compromising validity (Kung et al., 2018; Shamon & Berning, 2019). Importantly, when attention checks are not implemented based on *best practices*, they might hurt data quality (see Geisen, 2022); but the consensus in the literature is that including them in surveys can contribute to high quality data. Data variability refers to the level of variation in measurements and is usually expressed in a numerical way (e.g., by looking at the standard deviation or variance of a construct) (Alwin, 2016; Cella et al., 2015). This variability can reflect both measurement error and useful variance components (Stanton, 1998). Although we do not focus for the exact source of variability in this study, this measure can still be useful as an indication of data quality, because it 1) can demonstrate response behavior of unmotivated participants who simply choose the same answer repeatedly, 2) give an indicate of the level of measurement error and

3) highlight potential imprecise results (Galesic & Bosnjak, 2009; Santesso et al., 2020; Stanton, 1998; Wikman & Wärneryd, 1990).

Altogether, this study aims to explore the differences in these response characteristics and data quality variables between a conversational chatbot survey and an online web survey. Therefore, we propose the following research questions:

RQ1: Response characteristics:

RQ1a: do these survey modes differ with regards to the response rate?

RQ1b: do these survey modes differ with regards to the response time?

RQ1c: do these survey modes differ with regards to the length of open-ended responses?

RQ2: Data quality:

RQ2a: do these survey modes differ with regards to the consistency of the data?

RQ2b: do these survey modes differ with regards to the pass/fail rate of the attention check question?

RQ2c: do these survey modes differ with regards to the variability of the data?

Chatbots in longitudinal survey research

As already addressed earlier, an important implication of chatbots is their ability to build conversational relationships over time (Wijenayake et al., 2020), which could be interesting for longitudinal survey administration. This idea of 'relationship formation' is based on the Computers Are Social Actors (CASA) paradigm (Nass and Steuer, 1993; Reeves and Nass, 1996), which posits that humans tend to apply the same social heuristics and scripts used for human interactions to technology (i.e., the media equation hypothesis). That is, people tend to react socially to various technologies, including conversational agents such as chatbots (Araujo, 2018; Ho et al., 2018). An important concept in this context is *perceived anthropomorphism*, or the tendency to attribute human-like features and characteristics to a

chatbot (Go & Sundar, 2019). This means that people tend to "humanize" chatbots and respond socially to them. It has been shown that people tend to humanize chatbots a greater extent than a website, because of how people interact with these chatbots (i.e., in a conversational way via a chat dialogue) (e.g., Araujo, 2018; Zarouali et al., 2020). Put differently, interacting with a chatbot may very well feel like chatting with a human being.

This implies that differences between web surveys and chatbots surveys may change over time (i.e., over longer periods of time), as people might get more acquainted with conversational chatbots (compared to a web survey design) and eventually form social relationships with them (which might not occur in the case of a web survey). For instance, research has shown that in-depth qualitative interviews -containing open-ended questions- carried out by an acquainted interviewer (i.e., an interviewer that has some kind of basic relationship with the interviewee) could lead to "hidden" thought patterns that might otherwise never be discovered with a traditional interview (i.e., an interview where interviewee and interviewer do not know each other) (Garton & Copland, 2010; Owton & Allen-Collinson, 2014). The same logic could also take place with chatbots that are being used in longitudinal data collection (over a longer period of time): repeated use of a chatbot could create a relational bond between participant and conversational agent, which might influence the quality of people's survey responses. However, no study investigated this, and thus, it remains to be explored. Therefore, this research explores whether significant differences in response characteristics between chatbots and web surveys arise during a period of two weeks. To conclude, we propose the following research question:

RQ3. How are the differences in response characteristics (see RQ1) evolving over a time period of two weeks (i.e., longitudinal survey research)?

Chatbot survey and user evaluation

In addition to response characteristics and data quality, survey management should also center around scrutinizing and optimizing *user evaluations*. In this study, we will focus on comparing the two survey modes based on *utilitarian- and hedonic perceptions*, *perceived security* and *perceived cognitive load*.

Drawing on the tenets of the Technology Acceptance model (TAM), Bosnjak et al. (2010) presented a framework consisting of key utilitarian and hedonic factors in the context of surveys that rely on technology. In this framework, the authors included perceived usefulness and perceived ease-of-use as utilitarian factors, and perceived enjoyment as a hedonic factor. Perceived enjoyment refers to the degree to which the respondent likes to participate in the survey research (Rogelberg et al., 2001). A recent study found that participants evaluate a conversational survey as more enjoyable than a traditional web questionnaire (Celino & Re Calegari, 2020; Kim et al., 2019). In addition, it was also found that a higher level of enjoyment in a conversational chatbot survey also leads to a higher response quality (Wambsganss et al., 2020). Perceived usefulness indicates the respondent's perception that a chatbot survey will enable him or her to participate more efficiently in the research, whereas perceived ease-of-use refers to the expectations about the potential efforts needed to participate (Bosnjak et al., 2010). Prior research identified these two cognitive factors -i.e., perceived usefulness and perceived ease-of-use- as key variables in the context of responses to a conversational chatbot (Zarouali et al., 2018).

Perceived security refers to respondents' beliefs that their personal information will not be viewed, stored, and used by unauthorized third parties (Flavián & Guinalíu, 2006; Zarouali et al., 2021). This perception is important because surveys tend to collect personal data (e.g., socio-demographic information, political orientation, personality characteristics, etc.). Therefore, researchers should always try to enhance perceived security in survey research, especially as respondents may already be wary of the security of the Web in general

(de Leeuw, 2018). In the context of conversational agents, prior research has shown that people have difficulties in assessing the actual level of security of the chatbot (van der Goot & Pilgrim, 2020), which might hinder favorable evaluations of chatbot surveys (or even lead to unfavorable evaluations of this mode compared to a web survey).

Perceived cognitive load refers to the respondents' perceptions of the amount of mental effort they invested to fill out the survey (Paas & Van Merriënboer, 1994). In survey research, some respondents tend to give satisfying responses instead of accurate ones to reduce cognitive burden, because responding in an accurate way requires more mental efforts (Krosnick, 1991). Therefore, reducing the cognitive load can have a positive influence on the quality of online surveys responses (Fricker et al., 2005). Recent scholars have suggested that a chatbot might reduce the cognitive burden of a survey because of the conversational flow of chatbot surveys (Kim et al., 2019).

In the light of all this, we formulate the following research questions:

RQ4a: do people evaluate these survey modes differently with regards to perceived ease-of-use and usefulness (utilitarian features)?

RQ4b: do people evaluate these survey modes differently with regards to perceived enjoyment (hedonic feature)?

RQ4c: do people evaluate these survey modes differently with regards to perceived security?

RQ4d: do people evaluate these survey modes differently with regards to cognitive load?

Method

Design and data collection

This study was part of a larger project investigating news exposure and avoidance via chatbots. The recent paper of Ohme et al. (2022) was part of this project. On OSF, the present methodological research was pre-registered as a separate study (see procedure). This study

used a longitudinal experimental design with two *survey mode* conditions: a web survey condition vs. a chatbot survey condition. Participants that met the criteria of being adult (18+ years), owning a smartphone and having a Skype account[1] were invited for this research. In other words, these three criteria were formal requirements to be allowed to participate in this study. All participants were part of a research panel and were recruited by the company Panel Inzicht [2].

In total, 304 participants from The Netherlands were invited to participate in this study. Participants were randomly assigned to one of the two conditions: *chatbot* vs. *web survey*. More precisely, 148 participants were in the chatbot condition, and 156 participants in the web survey condition. The fieldwork was carried out by a market research company between November and December 2020. Participants received an incentive for their participation, which was negotiated to be 10% below the average incentive a respondent would receive for similar studies in the same panel. This was done to prevent the incentive structure from having an influence on the survey responses. The respondents had a mean age of 47.16 (*SD* = 16.50 years), and 48% of them were women. In terms of education, 11% of the respondents had a low education level, 37% had an intermediate educational degree, and 52% had a high education level.

Procedure

As already mentioned, this study is part of a larger project implementing a longitudinal experimental design with two conditions: web survey vs. chatbot survey. In both conditions, participants were invited to answer daily surveys over the period of two weeks. In the web-based survey condition, participants answered the daily surveys using Qualtrics. In the

[1] Participants were allowed to create a new Skype account to take part in the study, if desired. In this case, they received instructions on how to download Skype on their mobile device and to enable notifications, to ensure they received the invite from the chatbot on a daily basis (should they be in the chatbot condition).

[2] Light quotas were added to ensure variance in terms of gender and age to partially reflect the online population in The Netherlands. For education, these light quotas were not enforced given the uneven distribution of Skype users across education levels.

chatbot- based survey condition, participants answered the daily surveys through a conversational chatbot on the platform Skype (see stimuli materials). Participants were randomly assigned to one of the two conditions.

To participate, respondents first had to agree to an informed consent. At the start of the project, they completed a recruitment survey (or, an intake-survey), including control variables and moderators required for the overall project. This survey lasted around 15 minutes. The most important function of this recruitment survey was that it served to randomly assign respondents to one of the two conditions (chatbot or web survey). After completing the recruitment survey, respondents started filling out daily surveys during a period of two weeks, via one of the two survey modes. Thus, on a daily basis, respondents received a notification at a fixed point in time -i.e., afternoon- to complete their daily survey (during a period of 14 days). These daily surveys were short and lasted about 2-3 minutes. The topic of these daily surveys was about respondents' news consumption of the day. With regards to *wave non-response*: when respondents did not answer the daily surveys for more than two days in a row, their participation was ended (all other days were then marked as missing values). Only participants for which measurement on at least two days exist were included in the analysis

At the end of the experiment, i.e., after two weeks of completing daily surveys, participants were invited to complete a debrief survey that lasted about 10 minutes. This debrief survey served to gauge people's survey experience over the past fourteen days (either with the chatbot or the web survey) by assessing general perception variables (e.g., perceived enjoyment, perceived usefulness, etc.). In addition, this debrief survey also served to debrief the participants about the objectives of the study, as well as provide instructions to receive the incentive for participation. The procedure of the whole project was approved by the Ethical Review Board of the University of Amsterdam (Ethical Approval Number: 2020-PCJ-12371).

The full project was pre-registered on the Open Science Framework[3], as well as this methodological study as a separate project in its own right[4]. In this study-specific pre-registration, all the research questions, the design, recruitment details and analysis plan can be found. There are a couple of deviations from our pre-registration plan, which we carefully discussed in a separate document on the main OSF-page.

Stimuli material: chatbot survey vs. web survey

In the first condition, we created a chatbot specifically for the purpose of this study. The chatbot was developed using the Conversational Agent Research Toolkit (Araujo, 2020), which was extended to work with longitudinal designs, and made available via Skype[5]. We chose this chatbot because scholars can easily develop it themselves based on the step-by-step tutorial of Araujo (2020), and moreover, this choice also increases the transparency of our chatbot. In the second condition, participants took surveys in the online survey platform Qualtrics. The platform Qualtrics was chosen because it has become a leading platform to administer web surveys in academic research. Altogether, both the chatbot and the web survey we used should normally be the ones other scholars might use as well in their research, which increases the relevance of our stimuli.

The chatbot and web survey could both be completed by using any device (which we controlled for in the cross-sectional analyses on day 1). In both conditions, participants received a notification at the end of the afternoon to respond to the daily survey - by a message sent by the chatbot, in the chatbot condition, or via email, in the web survey condition. This means that not only the survey mode differed, but also the invitation mode. The reason for this is that chatbots are characterized by their ability to directly recruit participants on messaging platforms by means of instant messages (which is received as an

[3] https://osf.io/ws6ax/

[4] https://osf.io/p7rvj/

[5] The chatbot was published with the Bot Channels Registration service from Microsoft Azure.

instant notification by respondents). Thus, this element of notification cannot be "separated" from a chatbot, because otherwise you would be eliminating one of its most important features. Having said this, we do argue that it is important to consider this notification difference when interpreting the results (e.g., the variable response rate).

The survey flow was designed to collect repeated measures on key variables, such as media usage and news topic evaluation, while at the same time present variability in the question structure, to minimize item expectancy and reactance in responses. On average, similar questions were only asked every fourth day, while every day started with an open question about what participants perceived as the most important news topic on that day.

Chatbot design

The design of the chatbot and the web survey strived for keeping the conditions as similar as possible, while ensuring ecological validity for each condition. As such, the chatbot interacted with participants only via text (i.e., no images or buttons) in the standard Skype instant messenger chat window (with the exact same lay-out as an ordinary chat conversation). The questions asked followed three formats: (1) open-ended questions, where participants had to write the answer in text, (2) closed-ended questions with text options, where participants had to write an answer in text using one of the pre-defined options, and (3) closed-ended questions with numerical options, where participants had to write the answer as a number.

In addition, a validation script compared the participant's answers to the closed-ended (text or numeric) questions with a list of valid options. For text entries, the validation script accepted answers that had 90% or higher similarity[6] with one of the valid options. When the answer did not match, the chatbot asked the participant to type the answer again, highlighting

[6] Calculated as Levenshtein Distance using Python's FuzzyWuzzy package.

the valid options. The validation was done one time per question. If a participant provided an invalid response for a second time, the chatbot moved to the next question.

The survey flow and order of the questions were the same across conditions. In some instances, both the chatbot and the web survey skipped questions when needed (e.g., participants were only asked the general news avoidance question when they indicated low or no news consumption in the previous question). An important difference between the chatbot and the web survey were for scales with multiple items. In the web survey, they were shown as a matrix table allowing to combine multiple questions with the same answers. For the chatbot, they were always asked as individual questions.

Measures

Response characteristics and data quality variables

We computed the *response rate* by calculating the respondents who took the survey relative to all respondents invited to participate in the study. This rate was calculated for every single (survey) day during the 14-day research period. *Response time* was computed by calculating the total time spent in seconds for every survey in the 14-day longitudinal study. *Length of open-ended questions* was computed by counting the number of words answered in an open-ended question. The resulting value is a number (discrete variable). Higher values represent longer answers. *Attention check* was assessed by looking at whether participants passed or failed the control question. The item was: "To ensure data quality, we want to find out if you are paying attention. Please select "strongly disagree" for this item. The resulting value is binary in nature. A value 1 represents a "pass" for the control question, and thus higher data quality. *Internal consistency* was assessed by looking at the Cronbach alpha value of a construct that we included in both conditions: psychological reactance (Shen & Dillard, 2005). This scale was assessed based on a 7-point Likert scale with response options ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). Using this same reactance scale, we used the

R-package called *cocron* (comparing Cronbach alpha's) developed by Diedenhofen & Mush (2016) to compare the consistency estimates across groups for this variable. Finally, *variability* was assessed by looking at coefficients of variation (e.g., standard deviation) of the reactance scale. We used the R-package *cvequality* (Marwick and Krishnamoorthy 2018) to compare these coefficients across the two conditions.

User evaluation variables

To measure perceived ease-of-use, perceived usefulness and perceived enjoyment, we used validated scales originating from the study of Bosnjak et al. (2010), where these instruments have been used in the context of mobile surveys. The three scales consisted of three items each, with answer options ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). A sample item for perceived ease-of-use is: *"It was easy to answer the questions in the chatbot surveys";* for perceived usefulness: "*Chatbot surveys would largely facilitate my participation in research"*; and for perceived enjoyment: *"It was fun for me to fill out the questions in the chatbot surveys"*. Perceived security was measured by using two items from Hartono et al. (2014). The items were also measured on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The items were: *"My personal information is securely managed in these chatbot surveys"* and *"These chatbot surveys are safe for my personal information"*. Finally, perceived cognitive load was assessed based on a single-item self-report measure (Paas & Van Merriënboer, 1994; Park, 2015) on a nine-point scale ranging from 1 (*very low mental effort*) to 9 (*very high mental effort*), asking: *"How much mental effort did you allocate to complete the chatbot surveys?"*.

Data Analysis

All analyses were conducted in RStudio with R version 4.0.2. Prior to data analysis, data wrangling was performed using the *tidyverse* package (Wickham et al., 2019). This process included the merging of all daily surveys, recruitment survey and debrief survey into

one single dataset (wide format dataset). The full data wrangling code is publicly available on the main OSF-page mentioned above. After data wrangling, the data was ready for analysis. To compare estimates across the two conditions (chatbot vs. web survey), we conducted between-group significance testing (e.g., ANOVA, Chi-Square, t-tests, etc.). $RQ1_{abc}$ and $RQ2_{abc}$ were tested based on cross-sectional data, i.e., the data of the web survey and chatbot survey on the first day of interaction. The test results from these analyses indicate differences between the two modes at one point in time. RQ3 was answered by looking at the longitudinal data (see next paragraph). $RQ4_{a-d}$ were answered with between-group comparisons based on the data of the final debrief survey, which includes all the evaluation variables (e.g., enjoyment, usefulness, etc.).

This study relied on a latent growth modeling (LGM) approach to assess individual-level longitudinal changes in response characteristics (RQ3). Put simply, LGM can model the form of change in an outcome variable as time passes by (Hox et al., 2017). In addition, it allows to look at the influence of individual differences between respondents on the intercept and slope of the dependent variable. In this study, we used LGM with the response characteristic variables as outcomes (i.e., response time and length of open-ended responses), and the factor "condition" (web survey vs. chatbot) as a *time-invariant* covariate. We chose to simplify these longitudinal analyses by focusing on 5 time points: day 1, day 4, day 7, day 11 and day 14. This should give a good idea of how response characteristics evolve over time across the two conditions. The LGM used maximum likelihood estimation, and full-information maximum likelihood was selected as the method for dealing with missing data. The LGM analyses were conducted with the R-package *lavaan* (Rosseel, 2012).

Results

Cross-sectional results: web survey vs. chatbot survey

The first analysis compares the response rate (RQ1a) of both survey modes. This refers to the proportion of respondents who took the survey relative to all respondents invited

(i.e., total of invited respondents: $N = 304$). The results can be found in Table 1, together with a Chi-Square test to compare the two proportions. In terms of response rate, the chatbot survey did not differ significantly from the web survey condition. In the context of response time (RQ1b), first inspected for possible outliers that were more extreme than the mean + or - three times the standard deviation (Tabachnick & Fidell, 2013). Based on this procedure, we deleted 19 outliers. As shown in Table 1, the chatbot condition had a significantly higher response time than the conventional web survey. Finally, when looking at the length of the open-ended response (RQ1c), we see that respondents wrote significantly more words in the web survey condition. These results provide an answer to RQ1a-c.

[INSERT TABLE 1 AROUND HERE]

The next step in the analyses was to check the differences in data quality variables between the two survey modes (RQ2a-c). RQ2a and RQ2c will be answered by using the *psychological reactance* scale as a latent construct. It is important to note that we found no systematic differences in the mean values of this variable across the two survey modes.
For RQ2a, we looked at the internal consistency of psychological reactance. This resulted in a Cronbach's alpha of $\alpha = .80$ in the chatbot condition, and $\alpha = .88$ in the web survey condition. Based on a significance test with *cocron*, we found that this difference between the two estimates was significant ($F(155, 147) = 1.67$, $p < .01$).

RQ2b addressed the potential differences in pass/fail rate of the attention check question in the two survey modes. The analysis showed that 62.19% of the respondents passed the attention check (i.e., they selected the answer "*strongly disagree*") in the chatbot condition, whereas this proportion drops to only 50.00% in the web survey condition. This difference was found to be statistically significant ($\chi^2(1) = 3.78$, $p = .05$), indicating that more people pass the attention check in the chatbot condition.

Finally, we aimed to explore data variability, i.e., the extent to which there is variation in a construct across the two survey modes. We used the Modified signed-likelihood ratio test (M-SLRT) (Krishnamoorthy & Lee, 2014) with *cvequality*. Again, the variable psychological reactance was used to perform this test. This analysis resulted in a estimate of .93, indicating that there is no significant difference in variation across the two conditions (p = .33).

Longitudinal analyses

To answer RQ3, we look at how the response characteristics of $RQ1_{a\text{-}c}$ evolve over a period of two weeks. RQ3 is fundamentally a question about within-person change for the variables *response time* and *length of open-ended-responses*; however, for response rate (which is not calculated on a person-level), it is about comparing the aggregate rates at the different time points. In the light of this, Figure 1 presents the response rates in both conditions at the different time points, whereas Table 3 shows the results of the latent growth modelling (LGM) for the within-person longitudinal changes in response time and length of open-ended-responses.

In Figure 1, the first data point (d1) refers to the proportion -in percentages- of participants that filled out the survey (relative to the total number of invited respondents, i.e., 304 participants). Figure 1 shows that the response rates for both survey modes gradually decreased as time progressed, and that the decrease is rather synchronous in both conditions (i.e., roughly an equal amount of people dropped out of the study in both conditions), except for the last day (d14), where the chatbot condition experienced a greater decline. This was also the main finding based on the between-group analyses (Chi-Square): we found no significant differences in response rate between the chatbot and web survey, at no point in time, except on day 14 (p = .001). When it comes to within-person patterns of missingness, there was not a single account of a participant that had missed a day but answered again the next day. Thus, once respondents dropped out, they did not come back at a later time.

Importantly, Figure 1 also gives a good overview of our missing data. Based on the response rate proportions (relative to the 156 invited participants in the web survey condition, or 148 in the chatbot survey condition), the total number of observations per condition, per day can be calculated. For instance, on the last day, a total of 41 respondents had complete data for all days in the web survey condition (calculation: 156 * 26.28%) and 17 in the chatbot condition (calculation: 148 * 11.49%). In total, we had complete data for 19.08% of the invited respondents (missing ratio: 80.92%).

[INSERT FIGURE 1 AROUND HERE]

Then, we adopted an LGM-approach. When it comes to the model specification, *the level-1 (within-person) model* is expressed as $y_{ti} = \eta_{0i} + \eta_{1i}\, x_t + \varepsilon_{it}$, where y = outcome variable (i.e., response time and length of open-ended responses), x = time score, η_0 = baseline level growth factor (intercept), η_1 = trend growth factor (slope), ε = within-subjects error, i = individual, and t = time point (i.e., d1, d4, d7, d11 and d14). Subscripts 0 and 1 represent the baseline level (intercept) and trend (slope) growth factors. The level-2 (between-person) model equations are $\eta_{0i} = \alpha_0 + \gamma_0\, w_i + \zeta_{0i}$ and $\eta_{1i} = \alpha_1 + \gamma_1\, w_i + \zeta_{1i}$, where α = growth factor mean (intercept & slope), w = time invariant predictor (chatbot survey vs. web survey), and ζ = between-subjects error.

The latent growth model (linear) for *response time* had a good overall fit $\chi^2(13) = 14.41$, $p = .35$; RMSEA =.02 (CI: .000 - .062), CFI = .99, TLI = .99: SRMR = 0.08, which suggests that the model hypothesizing linear growth in response time over the five points in time is a good description of the data. The parameter estimates can be found in Table 2. As shown in this table, the average response time at the first measurement was 260.30 sec, with variance being non-significant. The response time decreased by 40.95 on average between each of the different time points, with a significant amount of variation in people's level of

decline (-3539.21). A significant covariance between the slope and intercept was found (5866.85), meaning that the growth-factor is significantly impacted by the initial level of response time. More precisely, a higher initial response time on d1 is related to slower or less decline over time. We also see that on average, people in the chatbot condition had a higher response time (+131.16) at the beginning of the measurements. These (chatbot) respondents exhibited a stronger decrease (compared to web survey respondents) as time passed by (-42.06).

The latent growth model for the variable length of open-ended responses had a good fit $\chi^2(13) = 20.65$, $p = 0.08$; RMSEA =.04 (CI: .000 - .078); CFI = .98: TLI = .98; SRMR = .05 (see Table 2 for parameter estimates). The results reveal that respondents wrote 7.15 words on average on day 1, with significant individual variation in this initial stage (11.36). The growth in number of words as time passed by was significant (+.45), and the variance in this growth factor (+2.13) was also found to be significant. The covariance (2.95) between intercept and slope was significant, meaning that a higher initial number of written words (on d1) is related to a higher rate of change over time. Finally, we note that respondents in the chatbot condition, on average, wrote less words in the open-ended questions (-1.95). In addition, a chatbot survey (compared to a web survey) has a negative effect on the linear growth trend of the number of words people write in open-ended questions (-.64). Thus, over time, the chatbot respondents will end up writing less words in open-ended question.

[INSERT TABLE 2 AROUND HERE]

User evaluations

To provide an answer to RQ4$_{a\text{-}d}$, a series of t-tests were conducted to compare the means across the two survey mode conditions. These results are presented in Table 3. This table shows that the web survey via Qualtrics generates a significantly higher perceived usefulness,

perceived enjoyment and perceived security among respondents. When it comes to the perceived cognitive load and perceived ease-of-use, no significant difference was found between the two survey modes.

[INSERT TABLE 3 AROUND HERE]

General Discussion

The aim of this study was to investigate the differences in response characteristics and data quality between a traditional, web-based survey on the one hand, and a conversational, chatbot-based survey on the other. Furthermore, we examined how people evaluate their interactions with these two survey modes. This study should be a starting point of research that aims to fuel a broader discussion about whether a chatbot (and other conversational interfaces) can serve as a reliable survey administration tool that might lead to higher data quality. These insights can help us make inferences as to whether a chatbot can be used as a valid substitute for a traditional web survey. It is important to note that this study was conducted with respondents from an online research panel, and thus, results (and their generalizability) should also be interpreted in the light of this.

First, one of the main contributions of this study is that it presents the methodological strengths and weaknesses of chatbots in large-scale survey research. More precisely, it provides unprecedented insights about the differences in *response characteristics and data quality* between chatbots surveys and web surveys. For *response rate*, it was found that the chatbot had a similar response rate than the Qualtrics web survey. Since response rate is very important to avoid response biases (Groves & Peytcheva, 2008), it can be concluded that chatbots seem to be on par with web surveys from the perspective of nonresponse biases. It was also found that chatbots had a *higher response time* as compared to web surveys, which may indicate a lower choice randomness and an increased precision of research answers (Börger, 2016). But as we will see in the next paragraph, this higher response time is most

likely the result of participants' unfamiliarity with the chatbot format (which decreased over the course of time). Chatbots also seem to trigger a lower *length in open-ended responses*. This somehow challenges the findings of Xiao et al., (2020), who found that chatbots can encourage people to reveal more personal information, and Lee et al. (2020), who found that chatbots prompt deeper self-disclosure. The fact that chatbots trigger shorter responses could be explained by the instant messaging format, where people might answer in a similar way as they would do in a regular chat conversation with friends via the instant messaging app (which is usually very brief and to-the-point) (Hill et al., 2015).

In terms of data quality, it was found that chatbots suffered from a lower internal consistency in latent constructs (i.e., Cronbach's alpha), but at the same time, generated a significantly better pass/fail ratio for the attention check pass/fail ratio compared to web surveys. Overall, these results show a more nuanced view of the effectiveness of chatbots as a data collection tool, and as compared to other recent studies (e.g., Celino & Re Calegari, 2020; Kim et al., 2019), we do not find evidence that chatbots are superior to traditional survey methods across the board.

Second, this study is one of the first studies that explores chatbots' response characteristics in survey research over time (mostly based on LGM). We see that chatbots are fairly similar to web surveys when it comes to response rate during a period of two weeks (except for the last day). For response time, a significant difference was found at day 1 (with chatbot respondents requiring more time to complete the survey), but this difference levelled of over time since the chatbot condition had a stronger decline in response. This might show a trend in which respondents first have to get familiar with the conversational format of the chatbot survey; after being familiar with these chatbot surveys on direct messaging platforms, we found that the time needed to complete the daily chatbot survey decreased. Finally, we see a pattern where a chatbot survey generates lower word entries in open-ended questions

compared to a web survey, and this difference tends to increase as time passes by. These shorter answers among chatbot survey respondents might lead to less detailed and insightful conclusions for researchers (e.g., Barrios et al., 2011; Israel, 2010).

Third, contrary to what other studies showed (e.g., Celino & Re Calegari, 2020; Kim et al., 2019), we found that chatbots have lower perceived enjoyment levels as compared to a web survey. So chatbots might not always be the best approach to engage respondents and ensure an enjoyable survey administration. In addition, chatbots were also perceived to be less of a useful tool for participating in survey research. This means that people still think that a traditional web survey are more efficient as a survey mode than the conversational chatbot. When it comes to ease-of-use and the perceived cognitive load, we found no difference between the chatbot and the web survey. This illustrates that respondents have no real difficulties in participating in conversational chatbot survey research. However, a worrying finding was that the chatbot did exhibit a lower level of perceived security. This might be explained based on the instant messaging platform in which the chatbot survey was operating. A recent study showed that privacy and security in instant messaging apps is very important for people; and that a non-negligible share of these people actually has data security concerns about these apps (Ali, 2021). It is important to consider that such app-related security concerns might have a negative influence on the security perceptions of chatbot surveys.

Overall, we did not find evidence that chatbots might be better survey research tools than web surveys. On the contrary, the "traditional" web survey often seemed to be the better approach to engage respondents in survey administration. However, we do not want to rule out that chatbot surveys can still be useful for research purposes, as other studies have found them to be a credible alternative or even a preferred method with respect to traditional tools (e.g., Celino & Re Calegari, 2020; Kim et al., 2019). An important point to address here is that our results (and the results of other studies exploring a conversational approach to survey

research) are dependent on the type of chatbot used. Using a different chatbot might for instance have an influence on people's survey experience (e.g., perceived enjoyment, perceived security), since every chatbot has its own style, features, and affordances. However, we think that other findings in this study should stay (fairly) stable when using a different type of chatbot (e.g., response characteristics). For instance, the length of open-ended responses was found to be lower in the chatbot condition than in the web survey condition, which was explained by the instant messaging format of the chatbot, which invites or triggers short replies instead of long answers. Even if scholars were to use another type of chatbot, we would still expect it to trigger shorter answers than a textbox in a traditional web survey. So, we are confident that this difference is due to the chatbot mode in general and not due to our specific chatbot and all the (potential) confounds which come with it. Another illustration of a finding that should also remain stable in a different chatbot context is the response rate. We believe that the features of a chatbot will not fully determine whether a participant will participate in a survey: They might influence the completion rate (e.g., if the features are bad or annoying, people might not finish the survey), but they should not have a major influence on the number of people starting the actual survey. In longitudinal designs, however, we might expect that the frequency of use of the social media platform or messaging application by the respondent in their daily routine may influence their likelihood to answer the chatbots on a more regular basis. To summarize, there are certain findings in this study that we believe would still hold in a different chatbot context (e.g., response characteristics), whereas other elements might heavily depend on the chatbot being used in a specific study (e.g., chatbot survey perceptions, or response frequency). Therefore, we argue that more studies should be conducted to provide an unequivocal answer about these chatbot differences, in order to have a more solid base of evidence about the promises and pitfalls of a conversational survey approach.

On a practical level, we believe that one of the advantages for scholars considering conducting surveys via chatbots is the possibility to reach large numbers of participants on platforms that are important in people's daily lives (including certain subpopulations that are harder to reach via conventional survey recruitment), with potential benefits for longitudinal designs. Although we did experience some practical difficulties to recruit participants for the Skype chatbot -because not everyone has a Skype account-, we believe that this problem will not occur in popular messaging apps, such as WhatsApp and Facebook Messenger. Thus, chatbots might very well become tools that can increase convenience in data collection and reach certain parts of society that might be under-represented in traditional survey research and, that scholars may want to consider adopting while being aware of the potential data quality challenges discussed above. In addition, scholars should also consider the cost and technical complexity of developing their own chatbots using open source frameworks – which allows for higher levels of control over the design and of the data –, or using off-the-shelf solutions provided by third-party suppliers or digital platforms that allow organizations to configure their own chatbots.

It is important to mention that a research design involving chatbots raises ethical questions with regards to informational *privacy*. For instance, ensuring participants' anonymity might be a great challenge on certain messaging platforms. To enact conversations, chatbots usually store personal information that comes from the messaging platform, such as the name (as informed by a user in their profile), location, and so on. This gives researchers the important responsibility of guarding participants privacy in these situations by being aware of which data are being collected *by default* by the different messaging platforms. In addition, researchers must be aware that participants' personal data are being created in *(and potentially for)* a platform. In-depth information is asked *within* the walled garden of a social and messaging media platform - for example Facebook Messenger-

where researchers do not have full data control. Having the chatbot asking the participant personality characteristics, political preference, ethnicity, or sexual orientation might create valuable data that can ultimately be *reused* by the platform for other objectives (e.g., user profiling). In addition, these data might also be used by a commercial third-party. It is important to be aware of whether such practices are taking place on the platform before starting data collection efforts via chatbots. These ethical and privacy-related questions should be tackled by researchers before setting up a chatbot research design, because ultimately, it is the researcher's responsibility to protect participants' data security and personal privacy.

Finally, this study has a couple of limitations that open up pathways for future research. First, we did not control for device effects in the longitudinal analyses. Although we did have this data in the survey on day 1 (which we controlled for in the cross-sectional analyses, which did not change the significance of the results), we did not have data about the device used in all the consecutive days of the study (days 2-14). Therefore, it is important that future research focus on the extent to which the device has an impact on the response characteristics and data quality over time when using chatbots as data collection tools. Second, we used a chatbot that was deployed on the instant messaging interface of Skype. As already addressed earlier, quality variables and user perceptions will most likely be different depending on the type of chatbot (e.g., a self-programmed academic chatbot, a commercial tech company chatbot, etc.) and the type of platform (e.g., famous messaging apps such as Facebook Messenger and WhatsApp, a private self-developed platform, etc.). We therefore encourage scholars to see how a chatbot survey performs when it comes to these different contexts and modalities. Third, given the uneven distribution of Skype users in the general population in terms of education, we did not enforce quotas on this level. This resulted in a higher proportion of higher educated people in our sample. Although we did not find education-related differences in our study, we should still acknowledge that educational bias can be a

major problem in survey research. Therefore, in future research, we encourage scholars to focus on instant messaging platforms that are more evenly used by people from different educational categories to avoid potential biases (e.g., WhatsApp, FB Messenger).

Disclosure statement: the chatbot was developed by one of the authors, adapting code published in: Araujo, T. (2020). Conversational Agent Research Toolkit: An alternative for creating and managing chatbots for experimental research. *Computational Communication Research, 2*(1), 35–51.

References

Ali, C. (2021). Do People Actually Care About Data Privacy in Messaging Apps? *Userlike*. https://www.userlike.com/en/blog/messaging-data-privacy-survey

Alwin, D. F. (2016). Survey Data Quality and Measurement Precision. In C. Wolf, D. Joye, T. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. SAGE Publications Ltd. https://doi.org/10.4135/9781473957893

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183–189. https://doi.org/10.1016/j.chb.2018.03.051

Araujo, T. (2020). Conversational Agent Research Toolkit: An alternative for creating and managing chatbots for experimental research. *Computational Communication Research*, *2*(1), 35–51.

Barrios, M., Villarroya, A., Borrego, Á., & Ollé, C. (2011). Response Rates and Data Quality in Web and Mail Surveys Administered to PhD Holders. *Social Science Computer Review*, *29*(2), 208–220. https://doi.org/10.1177/0894439310368031

Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human Relations*, *61*(8), 1139–1160. https://doi.org/10.1177/0018726708094863

Börger, T. (2016). Are Fast Responses More Random? Testing the Effect of Response Time on Scale in an Online Choice Experiment. *Environmental and Resource Economics*, *65*(2), 389–413. https://doi.org/10.1007/s10640-015-9905-1

Bosnjak, M., Metzger, G., & Gräf, L. (2010). Understanding the Willingness to Participate in Mobile Surveys: Exploring the Role of Utilitarian, Affective, Hedonic, Social, Self-

Expressive, and Trust-Related Factors. *Social Science Computer Review, 28*(3), 350–370. https://doi.org/10.1177/0894439309353395

Celino, I., & Re Calegari, G. (2020). Submitting surveys via a conversational interface: An evaluation of user acceptance and approach effectiveness. *International Journal of Human-Computer Studies*, 102410. https://doi.org/10.1016/j.ijhcs.2020.102410

Cella, D., Hahn, E. A., Jensen, S. E., Butt, Z., Nowinski, C. J., Rothrock, N., & Lohr, K. N. (2015). Method and Mode of Administration, Data Collection, and Analysis. In *Patient-Reported Outcomes in Performance Measurement*. RTI Press. https://www.ncbi.nlm.nih.gov/books/NBK424382/

Chau, P. Y. K. (1999). On the use of construct reliability in MIS research: A meta-analysis. *Information & Management*, *35*(4), 217–227. https://doi.org/10.1016/S0378-7206(98)00089-5

Daikeler, J., Bošnjak, M., & Lozar Manfreda, K. (2020). Web Versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates. *Journal of Survey Statistics and Methodology*, *8*(3), 513–539. https://doi.org/10.1093/jssam/smz008

de Leeuw, E. D. (2018). Internet Surveys as Part of a Mixed-Mode Design. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet* (1st ed., pp. 45–76). Routledge. https://doi.org/10.4324/9780203844922-3

Denniston, M. M., Brener, N. D., Kann, L., Eaton, D. K., McManus, T., Kyle, T. M., Roberts, A. M., Flint, K. H., & Ross, J. G. (2010). Comparison of paper-and-pencil versus Web administration of the Youth Risk Behavior Survey (YRBS): Participation, data quality, and perceived privacy and anonymity. *Computers in Human Behavior*, *26*(5), 1054–1060. https://doi.org/10.1016/j.chb.2010.03.006

Denscombe, M. (2008). The Length of Responses to Open-Ended Questions: A Comparison of Online and Paper Questionnaires in Terms of a Mode Effect. *Social Science Computer Review*, *26*(3), 359–368. https://doi.org/10.1177/0894439307309671

Denscombe, M. (2009). Item non-response rates: A comparison of online and paper questionnaires. *International Journal of Social Research Methodology*, *12*(4), 281–291. https://doi.org/10.1080/13645570802054706

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

Edwards, A., Edwards, C., Spence, P. R., Harris, C., & Gambino, A. (2016). Robots in the classroom: Differences in students' perceptions of credibility and learning between "teacher as robot" and "robot as teacher." *Computers in Human Behavior*, *65*, 627–634. https://doi.org/10.1016/j.chb.2016.06.005

Felderer, B., Kirchner, A., & Kreuter, F. (2019). The Effect of Survey Mode on Data Quality: Disentangling Nonresponse and Measurement Error Bias. *Journal of Official Statistics*, *35*(1), 93–115. http://dx.doi.org/10.2478/jos-2019-0005

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

Garton, S., & Copland, F. (2010). 'I like this interview; I get cakes and cats!': The effect of prior relationships on interview talk. *Qualitative Research*, *10*(5), 533–551. https://doi.org/10.1177/1468794110375231

Geisen, E. (2022). Improve data quality by using a commitment request instead of attention checks. *Qualtrics*. https://www.qualtrics.com/blog/attention-checks-and-data-quality/

Greenlaw, C., & Brown-Welty, S. (2009). A Comparison of Web-Based and Paper-Based Survey Methods: Testing Assumptions of Survey Mode and Response Cost. *Evaluation Review*, *33*(5), 464–480. https://doi.org/10.1177/0193841X09340214

Griffis, S. E., Goldsby, T. J., & Cooper, M. (2003). Web-Based and Mail Surveys: A Comparison of Response, Data, and Cost. *Journal of Business Logistics*, *24*(2), 237–258. https://doi.org/10.1002/j.2158-1592.2003.tb00053.x

Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, *72*(2), 167–189. https://doi.org/10.1093/poq/nfn011

Ha, L. S., & Zhang, C. (2019). Are computers better than smartphones for web survey responses? *Online Information Review*, *43*(3), 350–368. https://doi.org/10.1108/OIR-11-2017-0322

Hanna, R. C., Weinberg, B., Dant, R. P., & Berger, P. D. (2005). Do internet-based surveys increase personal self-disclosure? *Journal of Database Marketing & Customer Strategy Management*, *12*(4), 342–356. https://doi.org/10.1057/palgrave.dbm.3240270

Harris, M. A. (2010). Invited Commentary: Evaluating Epidemiologic Research Methods—The Importance of Response Rate Calculation. *American Journal of Epidemiology*, *172*(6), 645–647. https://doi.org/10.1093/aje/kwq219

Heerwegh, D. (2009). Mode Differences Between Face-to-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects. *International Journal of Public Opinion Research*, *21*(1), 111–121. https://doi.org/10.1093/ijpor/edn054

Henson, R. K. (2001). Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha. *Measurement and Evaluation in Counseling and Development*, *34*(3), 177–189. https://doi.org/10.1080/07481756.2002.12069034

Hill, J., Randolph Ford, W., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, *49*, 245–250. https://doi.org/10.1016/j.chb.2015.02.026

Hox, J. J., Moerbeek, M., & Schoot, R. van de. (2017). *Multilevel analysis: Techniques and applications* (Third edition). Routledge.

Ischen, C., Araujo, T., van Noort, G., Voorveld, H., & Smit, E. (2020). "I Am Here to Assist You Today": The Role of Entity, Interactivity and Experiential Perceptions in Chatbot Persuasion. *Journal of Broadcasting & Electronic Media*, *64*(4), 615–639. https://doi.org/10.1080/08838151.2020.1834297

Israel, G. D. (2010). Effects of Answer Space Size on Responses to Open-ended Questions in Mail Surveys. *Journal of Official Statistics*, *26*(2), 271–285.

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018). Evaluating and Informing the Design of Chatbots. *Proceedings of the 2018 Designing Interactive Systems Conference*, 895–906. https://doi.org/10.1145/3196709.3196735

Kim, S., Lee, J., & Gweon, G. (2019). Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300316

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284. https://doi.org/10.2146/ajhp070364

Krishnamoorthy, K., & Lee, M. (2014). Improved tests for the equality of normal coefficients of variation. *Computational Statistics*, *29*(1), 215–232. https://doi.org/10.1007/s00180-013-0445-2

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Kung, F. Y. H., Kwok, N., & Brown, D. J. (2018). Are Attention Check Questions a Threat to Scale Validity? *Applied Psychology*, *67*(2), 264–283. https://doi.org/10.1111/apps.12108

Lee, Y.-C., Yamashita, N., Huang, Y., & Fu, W. (2020). "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376175

Malhotra, N. (2008). Completion Time and Response Order Effects in Web Surveys. *Public Opinion Quarterly*, *72*(5), 914–934. https://doi.org/10.1093/poq/nfn050

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web Surveys versus other Survey Modes: A Meta-Analysis Comparing Response Rates. *International Journal of Market Research*, *50*(1), 79–104. https://doi.org/10.1177/147078530805000107

Ohme, J., Araujo, T., Zarouali, B., & de Vreese, C. H. (2022). Frequencies, Drivers, and Solutions to News Non-Attendance: Investigating Differences Between Low News Usage and News (Topic) Avoidance with Conversational Agents. *Journalism Studies*, *0*(0), 1–21. https://doi.org/10.1080/1461670X.2022.2102533

Owton, H., & Allen-Collinson, J. (2014). Close But Not Too Close: Friendship as Method(ology) in Ethnographic Research Encounters. *Journal of Contemporary Ethnography*, *43*(3), 283–305. https://doi.org/10.1177/0891241613495410

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*(1), 122–133. https://doi.org/10.1037/0022-0663.86.1.122

Park, S. (2015). The Effects of Social Cue Principles on Cognitive Load, Situational Interest, Motivation, and Achievement in Pedagogical Agent Multimedia Learning. *Journal of Educational Technology & Society*, *18*(4), 211–229.

Parnham, C. (2021). *Using Chatbots for better customer engagement – benefits, use cases and examples*. https://cognition.certussolutions.com/blog/using-chatbots-for-better-customer-engagement

Rogelberg, S. G., Fisher, G. G., Maynard, D. C., Hakel, M. D., & Horvath, M. (2001). Attitudes toward Surveys: Development of a Measure and Its Relationship to Respondent Behavior. *Organizational Research Methods*, *4*(1), 3–25. https://doi.org/10.1177/109442810141001

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Santesso, N., Barbara, A. M., Kamran, R., Akkinepally, S., Cairney, J., Akl, E. A., & Schünemann, H. J. (2020). Conclusions from surveys may not consider important biases: A systematic survey of surveys. *Journal of Clinical Epidemiology*, *122*, 108–114. https://doi.org/10.1016/j.jclinepi.2020.01.019

Shamon, H., & Berning, C. (2019). *Attention Check Items and Instructions in Online Surveys with Incentivized and Non-Incentivized Samples: Boon or Bane for Data Quality?* (SSRN Scholarly Paper ID 3549789). Social Science Research Network. https://doi.org/10.2139/ssrn.3549789

Shen, L., & Dillard, J. P. (2005). Psychometric Properties of the Hong Psychological Reactance Scale. *Journal of Personality Assessment*, *85*(1), 74–81. https://doi.org/10.1207/s15327752jpa8501_07

Shin, E., Johnson, T. P., & Rao, K. (2012). Survey Mode Effects on Data Quality: Comparison of Web and Mail Modes in a U.S. National Panel Survey. *Social Science Computer Review*, *30*(2), 212–228. https://doi.org/10.1177/0894439311404508

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Smyth, J. D., Dillman, D. A., Christian, L. M., & Mcbride, M. (2009). Open-Ended Questions in Web Surveys: Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality? *Public Opinion Quarterly*, *73*(2), 325–337. https://doi.org/10.1093/poq/nfp029

Van den Broeck, E., Zarouali, B., & Poels, K. (2019). Chatbot advertising effectiveness: When does the message get through? *Computers in Human Behavior*, *98*, 150–157. https://doi.org/10.1016/j.chb.2019.04.009

van der Goot, M. J., & Pilgrim, T. (2020). Exploring Age Differences in Motivations for and Acceptance of Chatbot Communication in a Customer Service Context. In A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, O.-C. Granmo, E. Luger, & P. B. Brandtzaeg (Eds.), *Chatbot Research and Design* (Vol. 11970, pp. 173–186). Springer International Publishing. https://doi.org/10.1007/978-3-030-39540-7_12

Wambsganss, T., Winkler, R., Söllner, M., & Leimeister, J. M. (2020). A Conversational Agent to Improve Response Quality in Course Evaluations. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3334480.3382805

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wijenayake, S., Berkel, N. van, & Goncalves, J. (2020). Bots for Research: Minimising the Experimenter Effect. *Adjunct Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 1–9. https://vbn.aau.dk/en/publications/bots-for-research-minimising-the-experimenter-effect

Xiao, Z., Zhou, M. X., Liao, Q. V., Mark, G., Chi, C., Chen, W., & Yang, H. (2020). Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions. *ACM Transactions on Computer-Human Interaction*, *27*(3), 15:1-15:37. https://doi.org/10.1145/3381804

Zarouali, B., Makhortykh, M., Bastian, M., & Araujo, T. (2020). Overcoming polarization with chatbot news? Investigating the impact of news content containing opposing views on agreement and credibility: *European Journal of Communication*. https://doi.org/10.1177/0267323120940908

Zarouali, B., Van den Broeck, E., Walrave, M., & Poels, K. (2018). Predicting Consumer Responses to a Chatbot on Facebook. *Cyberpsychology, Behavior, and Social Networking*, *21*(8), 491–497. https://doi.org/10.1089/cyber.2017.0518