Behavior Research Methods (2021) 53:2238–2251
https://doi.org/10.3758/s13428-021-01574-w

Can intelligent agents improve data quality in online questiosnnaires? A pilot study

Arne Söderström[1] · Adrian Shatte[2] · Matthew Fuller-Tyszkiewicz[1,3]

Accepted: 5 March 2021 / Published online: 5 April 2021

Abstract
We explored the utility of chatbots for improving data quality arising from collection via sonline surveys. Three-hundred Australian adults sampled via Prolific Academic were randomized across chatbot-supported or unassisted online questionnaire conditions. The questionnaire comprised validated measures, along with challenge items formulated to be confusing yet aligned with the validated targets. The chatbot condition provided optional assistance with item clarity via a virtual support agent. Chatbot use and user satisfaction were measured through session logs and user feedback. Data quality was operationalized as between-group differences in relationships among validated and challenge measures. Findings broadly supported chatbot utility for online surveys, showing that most participants with chatbot access utilized it, found it helpful, and demonstrated modestly improved data quality (vs. controls). Absence of confusion for one challenge item is believed to have contributed to an underestimated effect. Findings show that assistive chatbots can enhance data quality, will be utilized by many participants if available, and are perceived as beneficial by most users. Scope constraints for this pilot study are believed to have led to underestimated effects. Future testing with longer-form questionnaires incorporating expanded item difficulty may further understanding of chatbot utility for survey completion and data quality.

Keywords chatbot · autonomous conversational agent · online survey research · questionnaire item confusion · response accuracy · data quality

Introduction

Current estimates suggest that just over half the world's population have access to the Internet via computers and smartphone devices, and that the average daily Internet usage per capita is around 3 h (Statistica, 2021). Access and usage vary by region, with estimates ranging from around 80–94% for first world countries such as Australia, the UK, and USA, to 60–72% for countries with emerging economies such as China, Russia, and Turkey, and much lower rates of < 30% for poorer countries such as Indonesia, Pakistan, and Burkina Faso (Pew Research Centre, 2016). Across regions, it is also apparent that younger adults (18–34 vs. 35+), those with more education and higher income, and those who are male are more likely to report using the Internet at least occasionally (Pew Research Centre). Even so, the global trend is for those with access to use the Internet at least daily (Pew Research Centre). Hence, the sheer volume of people accessible via online means has encouraged development of online platforms, such as Amazon Mechanical Turk (Crowston, 2012) and Prolific (Palan & Schitter, 2018), to capitalize on Internet usage for research purposes.

This Internet-based survey research has become vital to psychology and the social sciences (Gosling & Mason, 2015). The instrumental utility of online questionnaires for research rests upon an assumed correspondence between the self-reported and actual levels of a construct being measured (Collins, 2003; Lietz, 2010). Unfortunately, evidence shows that a participant's propensity to respond either diligently or perfunctorily is influenced—at least in part—by the burden of confusing or unclearly worded questionnaire items (e.g., Lenzner, 2012; Smyth & Olson, 2018). For instance, Lenzner (2012) demonstrated experimentally that participants who received less comprehensible survey items were more

✉ Matthew Fuller-Tyszkiewicz
matthewf@deakin.edu.au

[1] School of Psychology, Deakin University, Geelong, Victoria 3220, Australia

[2] School of Science, Engineering, and Information Technology, Federation University, Berwick, Victoria 3806, Australia

[3] Center for Social and Early Emotional Development, Deakin University, Burwood, Victoria 3125, Australia

likely to break off from the survey (i.e., exit prior to completion), utilize nonsubstantive response options (e.g., 'Don't know' or simply leave a response blank), and were less consistent over time in response to the same questions than participants who received items that were easier to comprehend. This compromises the quality of survey response data, potentially introducing measurement error that distorts correlations involving affected variables, and reduces confidence in the generalizability of estimates from descriptive statistics (DeCastellarnau, 2018; Maniaci & Rogge, 2014) Fig. 1.

While these problems may arise for surveys delivered face-to-face, they are especially salient in online contexts, where investigators are not on hand to address participant confusion. Meanwhile, autonomous conversation agents called chatbots are increasingly being deployed in other online domains (virtual classrooms, technical support, government and business websites) as an effective and scalable means of approximating user support interactions (Androutsopoulou et al., 2019; Georgescu, 2018; Thorne, 2017; Zumstein & Hundertmark, 2017). Although this chatbot technology has the potential to mitigate the influence of questionnaire item confusion on the quality of survey response data, there is scant prior literature on adapting these technologies for that purpose. As such, this pilot study will investigate whether supplementing an online questionnaire with an assistive chatbot might improve the quality of elicited response data, and will explore participants' adoption of—and experiences with—this help feature.

Questionnaires and data quality

Online questionnaires are versatile tools that circumvent the barrier of geographical distance which separates researchers from a diverse global pool of potential participants. These questionnaires can be administered at large scales without incurring the considerable costs of face-to-face survey delivery (Gosling & Mason, 2015). Further, Internet-delivered instruments largely eliminate the error-prone and labor-intensive requirement for investigators to manually input research data into statistical software (Gosling et al., 2004; Riva et al., 2003). Despite the merits of online questionnaires, their value and utility ultimately rest on the assumption that participants will read, understand, and accurately respond to questionnaire items. These assumptions are not always satisfied (DeCastellarnau, 2018; Krosnick, 1999).

A participant's response accuracy (correspondence between reported and actual levels of the construct being

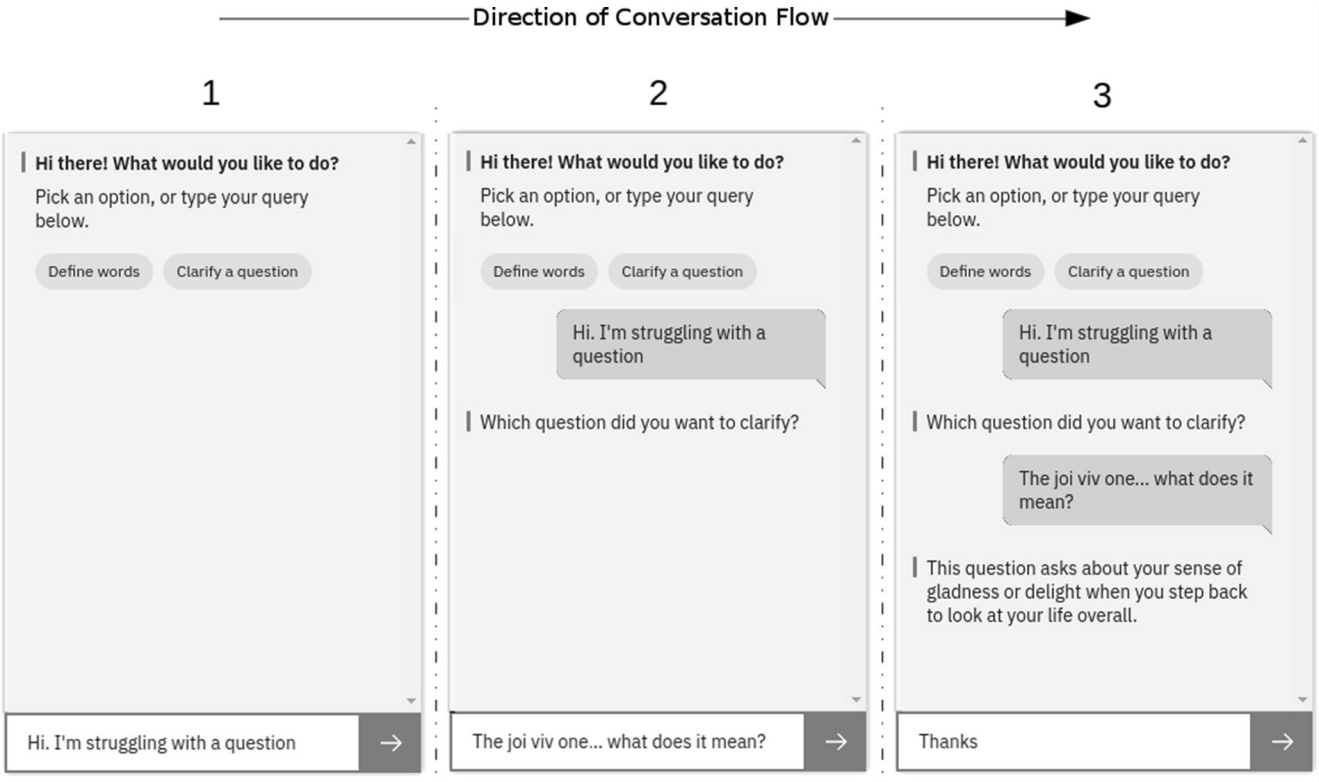

Fig. 1 Indicates an example of a brief help requested episode with the chatbot. Each the three panels is comprised of a conversation history pane (above) and a text entry field for user input (below). Messages from the chatbot are aligned with the left margin, white user-submitted message are right-aligned and visually delineated by a speech bubble. Text appearing in the user input field at the bottom of each panels has not yet been sent to the chatbot. Conversation flow progress left-to-right across the three panels, and top-to-bottom within the conversation history pane of each panel. Note the chatbot's robustness to casual or vague user input, including an incorrect spelling of the item being queried (i.e., *joi viv* instead of *joie de vivre*).

measured) can be influenced by a range of factors. One such influence is the order of item presentation. For instance, within the context of personality research, it has been demonstrated that presenting general items before more specific ones tends to lead to lower satisfaction ratings than if the specific items are presented first (Kaplin et al., 2013). Similarly, for individuals with elevated health risks, presentation of specific items about health domains prior to global assessments of health status tend to produce worse self-rated health (Garbarski et al., 2015).

The characteristics of a given questionnaire item may also influence response accuracy (DeCastellarnau, 2018; Van Vaerenbergh & Thomas, 2013). This happens when the design of a question interferes with the cognitive processes involved in answering it (Lietz, 2010). Commonly cited cognitive models emphasize that completing each item involves inferring the question's objective (comprehension), reflecting on cognitions—thoughts, feelings, remembered experiences—that are relevant to gauging subjective levels of the target construct (retrieval), evaluating whether candidate responses satisfy the stimulus query (judgement), then translating the decided answer into an appropriate outcome (response) captured on the questionnaire (Lietz, 2010).

Questionnaire items that incorporate certain features (obscure words, ambiguous phrases) have been shown to significantly increase the time and effort required to comprehend questions, which interferes with the very first step in the response process (Lenzner et al., 2010, 2011). This lack of clarity impedes participants' ability to quickly interpret what is being asked, making accurate responding difficult (Hamby & Taylor, 2016), even for motivated, conscientious individuals (Anson, 2018; Behrend et al., 2011; Hauser & Schwarz, 2016). Moreover, the confusion can induce respondents to engage in suboptimal behaviors, such as skipping items and utilizing nonsubstantive response options (such as 'Don't know' or 'Not applicable') (Lenzner, 2012).

Confusion-induced responding poses a considerable threat to the integrity of online survey research by artificially inflating or deflating observed means, score variability, and the strength of correlations, as well as eroding the statistical power of analyses (Baumgartner & Steenkamp, 2001; Maniaci & Rogge, 2014; Van Vaerenbergh & Thomas, 2013). It is possible to lessen the impact of such errors to some extent by incorporating attention-checking items into questionnaires to detect unsound response data (Abbey & Meloy, 2017; Curran, 2016; Niessen et al., 2016; Oppenheimer et al., 2009; Paas et al., 2018). However, the effective prevention of item confusion—and its erosive effect on data quality—is hindered considerably by the sheer scale and remote nature of the online context, where the survey setting is far removed from researchers who could quickly provide needed clarity.

Past efforts to reduce the burden of responding have targeted the structure of questionnaire items, such as by separating the elements of grid items into discrete questions (Roßmann et al., 2018). Others have instead attempted to dissuade speeded, inaccurate responses through the use of implied observers and admonitory survey instructions (Ward & Pond, 2015). Yet there has been no research assessing the effectiveness of empowering participants to improve their understanding by directly seeking clarification. One possible approach to enhancing item clarity might thus be the provision of resources that enable survey respondents to inquire about the meanings of confusing questionnaire items, thereby facilitating greater response accuracy. Although this is typically actioned via an attendant researcher in phone-based or face-to-face settings, such a resource could potentially be approximated in online surveys through the use of a chatbot.

Scalable intelligent support

A nascent area of research inquiry is the use of technology-enabled approaches to facilitate research conduct and dissemination. Such approaches have been applied at: (1) the back-end to identify and remove poor quality data following survey completion yet prior to intended analyses, and (2) the front-end as a substitute or complement to human-driven interactions, in order to cost-effectively facilitate dispensing of psychoeducation and treatment resources, and to help match the appropriate treatment resource to the individual.

Several statistical approaches and automated programs have been developed to identify patterns in collected data that may signal careless responding. Meade and Craig (2012) defined 12 possible indices to detect poor quality data, including time taken to complete the survey, total score across a number of sham items used to detect careless responding, correlations among item pairs tapping into the same construct, response to attention items, and number of identical responses in a row. Subsequent research has utilized bots to detect response characteristics such as those defined by Meade and Craig. For instance, Buchanan and Scofield (Buchanan & Schofield, 2018) developed a Google Chrome plug-in to automatically code respondent data that were suspicious in terms of time taken to complete each page of the online survey, performance on manipulation check items, distribution of participant responses and number of response options used, and click count to detect engagement with the items on each page of the survey. When using a cut-off of at least two indicators of untrustworthy data (out of a possible five), the authors found that their algorithm had excellent sensitivity and specificity, detecting 100% of automated data designed to be problematic and 99% of low effort data, and falsely flagging only 2% of the high effort data. It was argued that this may be used as an effective and quick way to filter out data subsequent to data collection but prior to analysis.

Chatbots have been employed at the front-end to engage directly with participants as they engage with one's product

(whether an intervention, healthcare service, or survey). Chatbots are autonomous conversational agents underpinned by algorithms that enable computers to learn human languages through examples and experience (Natural Language Processing [NLP]; Bengio et al., 2003; Cambria & White, 2014; Chowdhury, 2003; Nadkarni et al., 2011). These agents then use their acquired capabilities (learned patterns for linguistic usage; Tait & Wilks, 2019) to autonomously engage in conversational exchanges; a chatbot is effectively a computer-driven technology that approximates human interaction—you communicate with it and it responds. As robust and highly scalable intelligent technologies, chatbots are becoming ubiquitous on the Internet in a variety of conversation-based roles (virtual assistants, customer service agents; Chakrabarti & Luger, 2015; Hasler et al., 2013; Keeling et al., 2010; Radziwill & Benton, 2017).

A recent review of the literature underscores both the potential for wide-ranging psychological applications of chatbots and the current paucity of empirical evaluations of their utility (n = 10 studies) (Vaidyam et al., 2019). Three studies demonstrated the efficacy of chatbot-disseminated treatment content for alleviating depressive and anxiety symptoms (Fitzpatrick et al., 2017), reducing stress-related alcohol consumption and improving fruit consumption (Gardiner et al., 2017), and enhancing psychological well-being and reducing perceived stress (Ly et al., 2017). However, each of these studies comprised non-clinical samples, and the comparison condition comprised self-guided information resources or a wait-list control period.

Evaluations among clinical populations have demonstrated utility of chatbots for identification of clinical symptoms of depression (Philip et al., 2017) and posttraumatic stress disorder (Lucas et al., 2017), dispensing of discharge information to patients following a period of hospital stay (Bickmore et al., 2010a, b), and adherence to prescribed medication among patients with schizophrenia (Bickmore et al., 2010b). The use of chatbots for military service members with potential PTSD symptoms was demonstrated to elicit more symptom information than face-to-face interviews (Lucas et al., 2017), suggesting that under some conditions (preference for anonymity, anxiety around people, etc.) chatbots may be particularly advantageous. Across these reviewed studies found generally high acceptability and reported ease of use of the chatbot. Less is currently known about patient safety implications of chatbot use, though Vaidyam et al.'s review highlights very limited adverse reactions (two complaints out of ~800 cases) from the few studies that have documented this.

Autonomous conversational agents offer several potential benefits to researchers, from reductions in research costs (e.g., personnel remuneration), to the enhancement of available support for remotely located participants (Zumstein & Hundertmark, 2017). Being deployed over the Internet to interact autonomously and intelligently with human survey respondents is precisely the type of problem that chatbots have been designed and demonstrated to solve (Araujo, 2018; Ciechanowski et al., 2019; Clément & Guitton, 2015; Inkster et al., 2018; Pereira & Díaz, 2019). Indeed, interest in the potential applications for intelligent technologies within research contexts is growing, such as the employment of textual or embodied (i.e., graphically simulated) agents as virtual research personnel, and for approximating the administration of face-to-face conversational interviews (Conrad et al., 2015; Hasler et al., 2013). Despite these investigations however, the use of assistive text-based chatbots for mitigating questionnaire item confusion to improve data quality in online survey research remains largely unexplored.

The present study

Thus, the aim of the present study is to explore the data quality, feature adoption, and participant experience associated with using a chatbot to assist completion of an online survey. The investigation seeks to answer three research questions (RQ): (1) will participants access the help feature if available? (2) do chatbot users perceive the feature as helpful? and (3) does chatbot use improve data quality?

Propensity to use the chatbot (RQ1) will be measured as the proportion of participants in the chatbot condition who choose to initiate a valid query (pertaining to present questionnaire) via the provided help feature. Meanwhile, participant satisfaction with the chatbot (RQ2) will be gauged via optional user feedback; this exploratory research question is primarily aimed at informing future research and practice. Finally, the influence of chatbot use on data quality (RQ3) will be determined through between-group comparisons of response accuracy, as assessed independently for data from the chatbot and control groups.

Response accuracy is operationalized as the strength of positive correlation between a validated measure with known properties (target variable), and a tailored test item (challenge item) that has been formulated to correlate strongly with that target (mirroring its scale and meaning) while deliberately incorporating item characteristics that induce item confusion and predispose satisficing behaviors (e.g., obscure words, ambiguous phrases; Lenzner et al., 2010). Postulating that confusing items increase measurement error and weaken correlations, a comparison of the relationship between challenge items and their target variables should yield different strengths of association for individuals who gain assistance to resolve their confusion, than for those who receive no support. A group disparity in data quality is therefore defined as a significant difference between two corresponding Pearson's correlations observed independently in the data from each condition (chatbot, control) for an identical pairing of a given challenge item and target variable (employing Fisher's r-to-z

transformation to enable comparisons in standardized units with a normal distribution; Fisher, 1915).

Participants were randomized to either a chatbot or self-guided online survey completion condition. In both instances, a subset of the items was intentionally worded in a vague and confusing manner to impact data quality (control condition) and encourage chatbot use (chatbot condition). It was predicted that individuals in the chatbot condition would:

1. Use the chatbot feature specifically for these ambiguous items;
2. Report satisfaction with use of the chatbot; and
3. Demonstrate improved data quality relative to the control condition.

Method

Participants

A sample of 300 English-speaking adults representing the general Australian population were recruited into the present study. This was based on power calculations with power set at .80 and alpha at .05 (two-tailed) showing that a small, non-trivial standardized mean differences ($d > .32$) could be detected with 150 per group. This was deemed sufficient, balancing desire for a well-powered sample to detect meaningful effects against unnecessary oversampling and likelihood of significant results due to an over-powered study.

Table 1 displays the distribution of participants demographics across experimental conditions. Also included are results from chi-square analyses conducted to assess the significance of demographic differences between the chatbot and control groups. Fisher's exact test (Fisher, 1915) Exact Sig. (two-sided) indicated that the experimental conditions were demographically comparable, revealing no significant between-groups differences with respect to age group ($p = .062$), gender identity ($p = .325$), or educational attainment ($p = .928$).

Table 1 Distribution of participant demographics across experimental conditions

Variable	Condition		χ^2	Cramer's V
	Chatbot	Control		
Age bracket (years)			10.52	.19
18–25	43	57		
26–35	60	55		
36–45	19	27		
46–55	16	9		
56–65	9	3		
Gender identity			4.65	.13
Male	76	84		
Female	71	65		
Transgender	0	1		
Variant / non-conforming	2	0		
Other	0	1		
Educational attainment			2.49	.09
Secondary dropout	1	0		
Secondary student	8	5		
Year 12 or equivalent	30	34		
Certificate level	13	15		
Diploma / Advanced Diploma	11	9		
Graduate Certificate / Diploma	2	3		
Postgrad. (Hons, Mast., PhD)	51	51		

Note. Group differences assessed using Fisher's Exact Test due to unsatisfied cell count assumptions.

* $p < .05$

Comparing chatbot users and abstainers

Despite random allocation across chatbot and control conditions (with 1:1 allocation), 60 of the respondents allocated to the chatbot condition subsequently opted to abstain from engaging with the experimental manipulation. A within-group Chi-square analysis was conducted using Fisher's exact test to assess whether individuals who used the chatbot differed from the abstainers in terms of demographic factors. Fisher's exact test Exact Sig. (two-sided) indicated that those who engaged with the chatbot differed significantly from those who did not on the factor of age group ($\chi^2 = 12.19$, $p = .02$, $V = .29$), but not on the factors of gender identity ($\chi^2 = 2.01$, $p = .36$, $V = .13$) or educational attainment ($\chi^2 = 10.40$, $p = .13$, $V = .26$). Seventy of the 89 valid cases (79%) in the chatbot group were from age groups 18–25 ($n = 32$) and 26–35 ($n = 38$); rates of active participation were lower in the abstaining group (55% overall; $n = 11$ for 18–25 and $n = 22$ for 26–35). Thus, chatbot use abstainers tended to be older (see Table S1 in the supplementary file). The two groups also did not differ in terms of overall well-being score ($M_{users} = 61.86$ and $M_{abstainers} = 61.17$; $t(147) = .225$, $p = .823$, Cohen's $d = .04$), extraversion ($M_{users} = 2.92$ and $M_{abstainers} = 3.16$; $t(147) = .988$, $p = .325$, Cohen's $d = .16$), or conscientiousness ($M_{users} = 4.72$ and $M_{abstainers} = 4.85$; $t(147) = .538$, $p = .591$, Cohen's $d = .09$).

Limiting demographic comparisons to those chatbot participants who used the feature ($n = 89$) and a propensity score matched subsample of the control group ($n = 89$), the pattern of non-significant demographic differences for age, gender identity, and educational attainment reported in Table 1 for the whole sample replicated for

this subsample (see Table S2 in the supplementary file for further details).

Apparatus

Chatbot

An autonomous conversational agent was specifically designed and developed for this study by the first author using the IBM Watson Assistant service. The chatbot's knowledge domain was constrained to a subset of the present questionnaire (see measures overview below), along with anticipated queries related to its use (e.g., privacy, meaning of words, etc.). Primary design and performance criteria included (i) absolute quarantine from participant data; (ii) sensitivity in entity identification (specific word or item being queried); (iii) specificity in user intent discrimination (seeking definition, clarification, information); (iv) accuracy in response selection; (v) consistency in response delivery; and (vi) curtailment of irrelevant digressions during conversations. The preclusion of behaviors that might make the chatbot an unanticipated source of measurement error (varying explanations between users, allowing or enabling distraction from the questionnaire) were a particular focus during the iterative training and testing process.

We were deliberate in our decisions about how to incorporate the chatbot into the questionnaire. When not in use, both conditions saw identical page layouts except for an unobtrusive button for summoning the chatbot. When summoned, the chatbot took up the whole screen until dismissed. This was done to ensure: (1) a uniform chatbot user experience regardless of screen size, and (2) that the presentation of and interactions with the questionnaire itself would be identical for every participant across both conditions.

Questionnaire

An online questionnaire (Qualtrics) was configured to integrate the chatbot with one of two survey participation conditions. The chatbot condition provided optional access to chatbot assistance during questionnaire completion, while the control condition was unassisted. Aside from chatbot access and feedback items specific to chatbot use, the questionnaire content and layout were identical across conditions.

We could not guarantee that participants would complete the study on a computer rather than a mobile device. Several steps were taken to mitigate the risk that participant experience might differ according to device used to access the survey. The questionnaire was developed in accordance with Qualtrics' instructions for mobile optimization. Custom code was added to implement an intuitive but uncluttered user interface that looked and worked the same way on both small and large screens. Testing (n = 5 end users) confirmed the functional equivalence between mobile and desktop in terms of design, user experience, and integration of the chatbot.

Measures overview

In order to establish the potential influence of chatbot use on data quality, the present study required measures that were expected to correlate based on a body of prior studies. For this purpose, validated constructs of subjective well-being and life satisfaction were chosen from the well-being literature (Personal Well-being Index [PWI]; Cummins et al., 2003), and have been shown to correlate strongly with each other (Cummins et al., 2018). Personality correlates of well-being—as demonstrated in a recent meta-analysis (Anglim et al., 2020) —were also included as an additional check on response accuracy (Ten-Item Personality Inventory [TIPI]; Gosling et al., 2003). The conscientiousness construct in particular provides further opportunity to evaluate whether individuals differ in their use of the chatbot, with the expectation that more conscientious individuals will be more compliant with study instructions (Bowling et al., 2016). Further, challenge items devised to impact data quality were worded to mirror the well-being constructs such that, absent item confusion, these items would be expected to correlate with the target variables. Measures unrelated to data quality were based on chatbot session logs and user feedback.

Challenge items

Two challenge items were crafted to correlate strongly with the PWI variables by approximating their underlying meaning and scales, while deliberately incorporating syntactic and semantic features that increase difficulty in responding (Lenzner et al., 2011; Lietz, 2010). One item ($ITEM_{AMBIG}$) employed ambiguous phrase structure by conflating several disparate life domains ("*I am satisfied with my work, home life, and relationships*"). In construction of the item, we sought to follow the examples of double-barreled questions by asking about several distinct things but offering the participant only a single response. We chose three domains (work, home life, and relationships) that are common to answer, but that a participant may have different levels of satisfaction with.

The second item ($ITEM_{OBSCURE}$) invoked obscure terminology by roughly paraphrasing a statement of joyous life satisfaction using foreign words ("*I feel a sense of joie de vivre when thinking about my life*"). We chose an uncommon phrase to increase the challenge of the item, with the expectation that lack of familiarity with the phrase, and wording that did not provide context to guess the meaning, would produce the confusion we sought to elicit. Google Books Ngram Viewer confirmed the infrequency of this term in books in its library (< 0.00002% frequency).

The items were scored on a scale identical to that of the PWI (cf. Target variables below) except for textual anchors modified to fit the item wording ("*Completely Disagree*", "*Completely Agree*"). Higher scores indicate greater satisfaction.

Target variables

PWI The PWI is a short inventory that measures perceived satisfaction across each of seven life domains (*living standard, health, achievement, relationships, safety, community, security*; Cummins et al., 2003) along an 11-point scale (0-to-10, delimited by anchors "*Not Satisfied*", "*Completely Satisfied*"). These core domains yield a single score (PWI_{SWB}) representing an individual's subjective well-being (α = .70) (International Wellbeing Group, 2013). An auxiliary item ("*How satisfied are you with your life as a whole*") provides an additional—and correlated—measure of general life satisfaction (PWI_{LIFE}). Higher scores indicate greater levels of subjective satisfaction with the life domain being measured. The validity of this measure has been established previously, with support for a single factor structure and demonstrated predictive utility of the PWI for assessing global life satisfaction in a large Australian sample of 45,192 adults (Richardson et al., 2016).

TIPI A four-item subset of the Ten-Item Personality Inventory (Gosling et al., 2003) was used to measure levels of trait extraversion ($TIPI_{EXTRV}$) and conscientiousness ($TIPI_{CONSC}$). Each item used a seven-point scale ("*Disagree Strongly*", "*Agree Strongly*") to record the extent of a participant's self-identification with trait-specific attributes (e.g., "*I see myself as reserved, quiet*"). Higher scores indicate greater identification with the trait attribute being measured. The TIPI is optimized for content validity rather than internal consistency, and due to having only two items per scale, reliability estimates appear low for both the Extraversion (α = .68) and Conscientiousness (α = .50) scales (Gosling et al., 2003). The initial validation study for the TIPI demonstrated acceptable model fit for the proposed factor structure, strong correlation with the same constructs using a longer, previously validated personality measure, strong test–retest reliability over 2 weeks, external validity in terms of correlations with constructs previously linked to personality (e.g., self-esteem and depressive symptoms) (Gosling et al., 2003). Factor structure and convergent validity have been confirmed subsequently on new samples (e.g., Ehrhart et al., 2009; Myszkowski et al., 2019).

Feedback Optional feedback items at the end of the questionnaire gauged participant satisfaction with the chatbot. Items measured ease of use ("*I found it easy to use the help feature*"), perceived utility ("*The help feature made it easier to understand and answer the questions*"), and attitudes toward wider chatbot availability in online surveys ("*I would like more questionnaires to provide this kind of help feature*"). Participant responses were captured on a non-numerical scale delimited by textual anchors indicating either negative or positive feedback ("*Strongly Disagree*", "*Strongly Agree*"), and bisected by neutral response option ("*Neither agree nor disagree*"). Scores for feedback items were used separately to represent three dimensions of participant satisfaction (ease of use, utility, feature acceptance).

Usage A count of chatbot interactions with unique users was automatically generated by the IBM Watson Assistant service that powers the chatbot, quantifying the questionnaire sessions during which a participant submitted at least one query to the chatbot. Chatbot activity logs record the particular questionnaire items targeted by valid queries (entities that the chatbot's underlying neural network is trained to recognize), along with the form of assistance requested (the user's intent—defining words, clarifying phrases).

Procedure

Participants were recruited via Prolific Academic (www.prolific.ac), a crowdsourcing platform providing access to pre-screened participants curated to satisfy the ethical and methodological requirements of academic researchers (Palan & Schitter, 2018; Peer et al., 2017). Survey invitations were distributed via Prolific. Interested respondents were redirected to Qualtrics, whereupon consenting participants were randomized across chatbot and control conditions. Both conditions completed a questionnaire comprised of demographics (age, gender, education) and measures items (PWI, TIPI, challenge items) in the same order. Aside from chatbot-specific instructions and feedback questions, both conditions employed identical content.

While the control group progressed unassisted from one section to the next until done, the chatbot group was notified about the help feature before entering the measures section, and could repeatedly summon and dismiss it as needed via an on-screen button. Chatbot use was entirely optional, and the help feature was only accessible for items within the measures section. Participation concluded upon questionnaire completion. In accordance with a minimum pro rata hourly compensation stipulated by Prolific to preclude exploitation, participants were reimbursed with a nominal payment unlikely to induce coerced or risky behavior. Involvement in the study remained entirely anonymous and voluntary, and respondents were free to discontinue at any time. Ethics approval was obtained prior to conducting the study. Informed consent was obtained in advance, and the privacy rights of all participants were observed. While ethics approval does not permit public

availability of participant data, coding of the chatbot may be requested to the corresponding author.

Data analysis

Prior to analysis, data underwent assumption checking. There were no missing values for survey items. The data distributions for all variables after cleaning were sufficiently normal, based on cut-offs for absolute skew and kurtosis (Mishra et al., 2019). Several outliers were identified, but these were low in number (< 1% of overall sample), and scores were within possible scale ranges. As such, these cases were retained.

A series of analyses was conducted to address the three key research hypotheses. Chatbot usage (Hypothesis 1) and user satisfaction (Hypothesis 2) relied on descriptive statistics from those in the chatbot condition who utilized the chatbot feature. As a substantial portion of individuals within the chatbot condition (60 out of 149) did not utilize the chatbot function, two strategies were used for evaluation of group differences between chatbot and control participants in correlation strength (Hypothesis 3). First, an intention-to-treat (ITT) approach was utilized, in which correlations were compared across groups using the whole sample. This retains randomization, but likely underestimates group differences since it includes people in the chatbot group who did not use this feature. A second approach to augment these ITT results was to limit the chatbot group to those who used the chatbot feature (n = 89) and use propensity score matching in an attempt to balance the chatbot and control groups. We used the R package *MatchIt* (Ho et al., 2011) for propensity score matching, with 1:1 nearest neighbor matching across groups based on demographic variables available within our dataset (age, gender, educational attainment). As shown in Table S2 (supplementary file), matching resulted in non-significant differences between groups for these demographics, mirroring non-significant group differences for the whole sample produced by randomization (see Table 1).

Results

Hypothesis testing

H1: Chatbot usage

On average, individuals in the chatbot condition took longer to complete the survey (M = 5.83 mins, SD = 4.85) relative to the control group (M = 3.92 mins, SD = 4.85); $t(238) = 2.94$, p = .004, Cohen's d = 0.39. Eighty-nine participants (60% of initial chatbot group) chose to utilize the chatbot. Query validity was verified via aggregated usage data from the IBM Watson Assistant service. Overall, there were 251 individual messages received by the chatbot across 89 unique query sessions (M = 2.82 messages). Approximately 87% of queries targeted challenge items ($ITEM_{OBSCURE}$ = 69 queries [78%]; $ITEM_{AMBIG}$ = 8 queries [9%]). Meanwhile, only about 4% of queries addressed PWI items, and another 9% were irrelevant.

H2: User satisfaction

Table 2 summarizes feedback from chatbot users, revealing the distribution of negative, neutral, and positive sentiments relating to their user experience. The aspect of chatbot use that received the highest frequency of negative responses was ease of use, but this accounted for less than one-sixth of users. Chatbot usefulness and the desirability for wider chatbot availability each attracted fewer unfavorable responses. Although a considerable number of participants responded neutrally on each feedback item, the majority of feedback across every user experience dimension was positive.

H3: Data quality

Bivariate correlation analyses were conducted separately for $ITEM_{AMBIG}$ and $ITEM_{OBSCURE}$ to assess relationships between each challenge item and the target variables (PWI_{SWB}, PWI_{LIFE}, $TIPI_{EXTRV}$, $TIPI_{CONSC}$). Correlations within each of the two analyses were assessed separately for the chatbot and control groups. The results are displayed in Table 3 (for full sample) and Table 4 (propensity matched sample), including confidence intervals around each value for Pearson's r, and relative difference in corresponding values r between experimental conditions (Δr).

Full sample results Following Cohen (1988), absolute value differences in Pearson's r between the two groups indicated small but meaningful effects ($\Delta r > |.1|$) in data from chatbot users were observed for the full sample between $ITEM_{OBSCURE}$ and $TIPI_{EXTRV}$, as well as between $ITEM_{OBSCURE}$ and $TIPI_{CONSC}$. In both instances, the larger value for Pearson's r was observed in data for the chatbot condition.

All challenge items were significantly related to every target variable within data from both chatbot and control conditions. Positive values for Δr were observed across all intercorrelation comparisons except for the relationship between $ITEM_{OBSCURE}$ and PWI_{LIFE}, revealing a reasonably consistent trend for stronger underlying correlations for the chatbot group. However, none of the observed differences reached significance using comparisons based on Fisher's z-transformation of correlations.

Propensity matched sample results When comparing those who used the chatbot feature against a propensity matched control group, three results showed small but meaningful differences in magnitude of correlations across the groups ($\Delta r >$

Table 2 Summary of Participants Feedback about their User Experience with the Chatbot

Aspect of User Experience	Feedback		
	Negative	Neutral	Positive
To help feature was easy to use	15 (17%)	29 (33%)	45 (51%)
Chatbot made it easier to understand and answer questions	8 (9%)	34 (38%)	47 (53%)
More surveys should provide a similar help feature	8 (9%)	20 (22%)	61 (69%)

Note: Values in the table represent the breakdown of chatbot group participants by counts (and percentage) by feedback valence (i.e., negative, neutral, positive) for each user experience dimension. There were no cases with missing Values for the chatbot user feedback data (n = 89).

|.1|): correlations between (1) $ITEM_{AMBIG}$ and $TIPI_{EXTRV}$, (2) $ITEM_{OBSCURE}$ and $TIPI_{EXTRV}$, and (3) $ITEM_{OBSCURE}$ and $TIPI_{CONSC}$. Across all correlation pairs, magnitude of correlation was greater for the chatbot group, though only the difference for $ITEM_{OBSCURE}$ and $TIPI_{EXTRV}$ was significant.

Discussion

Online questionnaires are indispensable tools for contemporary social and psychology research (Gosling & Mason, 2015), but their ubiquity belies fundamental methodological caveats. Questionnaire items that are ambiguous, obscure, or otherwise confusing, are thought to impose increased cognitive demands on survey respondents (Lenzner et al., 2010, 2011), predisposing them to engage in compensatory behaviors theorized to alleviate the heightened burden (Krosnick et al., 1996). Such behaviors can compromise the accuracy of responses captured in the survey data, undermining subsequent statistical analyses and the research findings they inform (Van Vaerenbergh & Thomas, 2013). Predicated on the established applicability of autonomous conversational agents to analogous problem domains (Radziwill & Benton, 2017; Zumstein & Hundertmark, 2017), this pilot study thus set out to explore the potential utility of user support chatbots for bolstering the integrity of Internet-based survey research. The present findings broadly support the chatbot help feature, showing that it was used, perceived as useful, and enhanced data quality.

Table 3 Correlations and confidence intervals by condition between challenge and target measures for intention-to-treat sample (n = 300)

Variables	Condition									Difference		
	Chatbot Queriesa	Chatbot				Control				Fisher's z		Δr[b]
		n	r	95% C.I.		n	r	95% C.I.		z	p	
				Lower	Upper			Lower	Upper			
$ITEM_{AMBIG}$	8											
PWI_{SWB}[c]		149	.85***	.80	.89	151	.80***	.74	.85	1.35	.089	.05
PWI_{LIFE}[d]		149	.82***	.75	.86	151	.79***	.72	.84	0.73	.232	.03
$TIPI_{EXTRV}$[e]		149	.36***	.21	.49	151	.31***	.16	.45	0.48	.315	.05
$TIPI_{CONSC}$[f]		149	.42***	.27	.54	151	.36***	.22	.50	0.61	.272	.06
$ITEM_{OBSCURE}$	78											
PWI_{SWB}		149	.75***	.67	.81	151	.75***	.67	.81	0.00	.500	.00
PWI_{LIFE}		149	.71***	.62	.78	151	.75***	.67	.81	0.74	.231	-.04
$TIPI_{EXTRV}$		149	.45***	.32	.57	151	.34***	.19	.48	1.12	.132	.11
$TIPI_{CONSC}$		149	.39***	.24	.52	151	.28***	.13	.42	1.06	.144	.11

Note. Positive values for Fisher's z test (i.e., between-groups differences in standardized effects) and Δr indicate stronger correlations in chatbot group. One-tailed significance is reported for Fisher's z due to directional hypotheses.

[a] Chatbot queries represent a raw count of help requests received in connection with a given challenge item (i.e., from unique users in the chatbot group).

[b] Values for Δr that are greater than |.1| represent small but meaningful effects.

[c] Subjective well-being score computed from the 7 core PWI items.

[d] Life satisfaction is the score from an auxiliary PWI item that captures a general life satisfaction rating.

[e] Extraversion and [f] Conscientiousness are scores from the relevant subscales of the TIPI.

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 4 Correlations and confidence intervals by condition between challenge and target measures for propensity score matched subsample ($n = 178$)

Variables	Condition									Difference		
	Chatbot Queriesa	Chatbot				Control				Fisher's z		Δrb
		n	*r*	95% C.I.		*n*	*r*	95% C.I.		*z*	*p*	
				Lower	Upper			Lower	Upper			
$ITEM_{AMBIG}$	8											
PWI_{SWB}c		89	.86***	.79	.90	89	.82***	.74	.88	0.90	.186	.04
PWI_{LIFE}d		89	.85***	.78	.90	89	.79***	.69	.85	1.21	.113	.06
$TIPI_{EXTRV}$e		89	.48***	.30	.63	89	.30***	.09	.48	1.40	.081	.18
$TIPI_{CONSC}$f		89	.43***	.24	.58	89	.37***	.17	.54	0.47	.320	.06
$ITEM_{OBSCURE}$	78											
PWI_{SWB}a		89	.80***	.72	.87	89	.76***	.65	.83	0.67	.251	.04
PWI_{LIFE}		89	.82***	.73	.88	89	.77***	.67	.85	0.90	.186	.05
$TIPI_{EXTRV}$		89	.57***	.41	.70	89	.35***	.15	.52	1.85	.032	.22
$TIPI_{CONSC}$		89	.38***	.19	.55	89	.25***	.04	.43	0.95	.172	.13

Note. Positive values for Fisher's z test (i.e., between-groups differences in standardized effects) and Δ*r* indicate stronger correlations in chatbot group. One-tailed significance is reported for Fisher's z due to directional hypotheses.

[a] Chatbot queries represent a raw count of help requests received in connection with a given challenge item (i.e., from unique users in the chatbot group).

[b] Values for Δ*r* that are greater than |.1| represent small but meaningful effects.

[c] Subjective well-being score computed from the 7 core PWI items.

[d] Life satisfaction is the score from an auxiliary PWI item that captures a general life satisfaction rating.

[e] Extraversion and [f] Conscientiousness are scores from the relevant subscales of the TIPI.

* $p < .05$, ** $p < .01$, *** $p < .001$

Key findings

The majority of participants with access to assistance did indeed use the chatbot, doing so primarily for help with the confusing challenge items (Lenzner et al., 2010). Feedback from chatbot users was largely positive, with participants reporting that the chatbot made survey completion easier. This might be explained by the chatbot's overt role in resolving item confusion, thereby pre-empting the perceived response burden associated with satisficing (Barge & Gehlbach, 2012). Usability was rated positively, and most chatbot users endorsed the prospect of wider availability of similar assistance in online surveys, mirroring prior findings of chatbot acceptance (Clément & Guitton, 2015).

Of the two challenge items that were included to evoke confusion-induced suboptimal responding (Lenzner et al., 2010), $ITEM_{OBSCURE}$ generated the majority of requests for chatbot assistance. This disparity in chatbot utilization revealed an interesting relationship between chatbot usage and data quality; data quality was improved (relative to controls) where participants actually used the chatbot to resolve item confusion, but the greatest gains in correlation magnitude occurred for those relationships that were found to be smaller in both groups (i.e., those involving the personality target measures rather than the well-being target measures). Thus, the benefits of chatbot functionality may be most pronounced where the population correlation values are small, and subtle effects due to item confusion may make the difference between a significant and non-significant finding in one's study.

It is also noteworthy that two small yet-meaningful improvements in correlation magnitude for the chatbot group were found for the $ITEM_{OBSCURE}$ item compared to one for the $ITEM_{AMBIG}$ item. In light of the rarity of chatbot queries targeting $ITEM_{AMBIG}$, this finding was unsurprising. One possible reason for the unanticipated ineffectiveness of this challenge item for eliciting requests for assistance might be the generally healthy sample recruited into this study (scoring near or above Australian well-being average; Cummins et al., 2003). It is conceivable that individuals who are broadly satisfied with each of the life domains that are conflated by $ITEM_{AMBIG}$ might perceive no conflict when rating their satisfaction across those domains collectively.

Finally, an age effect was found for use of the chatbot feature among those given access to it. Whereas 79% of participants in the chatbot condition who used the feature were in the 18–25 and 26–35 age groups, only 55% of chatbot abstainers were within these age groups, suggesting that older

participants were less likely to use the chatbot feature. We propose several plausible explanations for these age-related effects. First, age may act as a proxy for acquired knowledge. Younger participants may be less likely to have been exposed to the phrase *joie de vivre* (the challenge item most often prompting chatbot use in the current study), and hence need for the chatbot function may skew towards younger participants. Second, there may be age-related differences in help seeking approach. There is some evidence to suggest that older individuals may have less interest in using technology (Ellis & Allaire, 1999). Further, older individuals may prefer to interact with a human to find out information (Nadarzynski et al., 2019; Van der Groot & Pilgrim, 2020).

Study limitations

In designing this study, the authors chose to formulate the challenge items (for impacting data quality) on item characteristics most often linked to confusion and response errors (obscure or ambiguous wording; Lenzner et al., 2010, 2011; Lietz, 2010), but there are other ways to manipulate item difficulty that might lead to different usage patterns (response scales; DeCastellarnau, 2018). The use of only two challenge items in an otherwise brief and uncomplicated questionnaire is not reflective of long, onerous surveys. This likely led to an underestimation of the protective influence of chatbot use on data quality in the present findings, particularly in light of the lack of item confusion linked to $ITEM_{AMBIG}$.

A second limitation is the choice of items for manipulating participant confusion. In testing the utility of a chatbot feature, the present study design required a trade-off between experimental control (as optimized in the present study) and ecological validity. Our design was chosen under the assumptions that: (i) uncommon words would elicit confusion, and (ii) this confusion would be a prompt for chatbot utilization. The chatbot use logs suggest we were successful in generating this confusion for the *joie de vivre* item. In contrast, in many existing scales the participant confusion is likely to be harder to predict since scale developers will often generate items that seek to minimize jargon, vague, or uncommon terminology. Thus, there would be considerable uncertainty in how many participants were needed to evaluate chatbot utility for existing scales, where the items are understood by the majority of participants, rendering this approach less practical for research. Even so, we recognize that level of confusion may determine whether an individual seeks help, either from a human or chatbot. It is presently unclear what level of confusion is needed to elicit help seeking behavior; further research is warranted to address this.

A third limitation of this study was the high level of educational attainment in the sample (over half had graduate-level qualifications). This likely counteracted the impact of challenge items, enhancing response accuracy for the control group, and limiting the scope for detecting group differences in data quality. For example, participants might have resolved the seeming ambiguity of $ITEM_{AMBIG}$ by drawing inferences about the item's likely objective based on its similarity to the preceding well-being questions (Tourangeau et al., 2000). However, the consistently weaker relationships among challenge and target measures in control group data—in terms of absolute strength, including where no significant group differences were observed—suggests that a sample with more representative educational attainment might widen the potential difference between unassisted and supported response accuracy, improving the ability to detect an effect for chatbot use.

Finally, as a pilot study, it was unclear how many individuals would engage with the chatbot if assigned to this treatment group. The high number of participants assigned to the chatbot condition who did not utilize this feature led to a reduction in power that may contribute to several null findings. Present findings may thus help future studies to more accurately calculate required sample size. This noncompliance threatens validity by potentially disrupting balancing due to randomization, and it is also unclear what reasons account for failure to use the chatbot (e.g., simply failing to see the feature or refusal to use it). It is assuring to see that omitted participants in the chatbot group did not differ from those were retained in terms of demographics and psychological constructs, with the exception of age. Lack of group difference for conscientious scores, in particular, suggests against potential biases in response such that more conscientious participants may be assumed to comply more fully with instructions of the study. Even so, more expansive testing of individual difference factors in future studies may enhance understanding of who is more likely to avail themselves of chatbot support for their survey completion.

Future directions

Both the pattern of observed findings and the abovementioned study limitations provide directions for further research. First, as a pilot study, tests of the chatbot's functionality were constrained to two types of confusing items. Evaluation of the provision of chatbot functionality across a wider range of manipulated survey design features (e.g., DeCastellarnau, 2018) in future studies could help to elucidate the contexts in which such functionality is likely to be adopted and have greatest impact on data quality. A key challenge in such a study is to be able to anticipate the types of queries that participants are likely to provide the chatbot with.

Second, the present study evaluated demographic and personality factors that may be associated with acceptance and utilization of the chatbot feature. Age-related effects in chatbot usage were evident, but require further examination. Factors such as level of motivation for completing the survey, preference for anonymous online communication versus human-to-

human interactions, and comfort level with technology may serve as moderators of chatbot utilization and benefit. Level of access to and frequency of use of the Internet may also influence uptake of the chatbot. For instance, individuals who regularly use the Internet may have previously encountered chatbots, and may thus have prior experience and formed expectations that guide their interactions with chatbots in future.

Relatedly, an individual's reading level, cognitive ability, and mental state may also determine the value they may derive from a chatbot. While chatbot functionality could plausibly help respondents to provide more accurate data in clinical contexts (e.g., to inform diagnosis and treatment allocation; Vaidyam et al., 2019), the text-based elements of a chatbot assume a level of reading ability (and motivation) that may preclude some potential end-users. Co-design principles may be useful for determining the appropriate amount of text-based content and the word comprehension levels that would make the product accessible to a broader audience.

Conclusions

In summary, findings showed that chatbot assistance, when utilized, may make modest contributions to enhancing data quality. Participants were inclined to seek help from a chatbot if given the option, generally found the feature to be beneficial, and broadly endorsed its wider adoption in online surveys. However, parameter constraints due to the exploratory nature of this pilot study are believed to have led to an underestimated effect for chatbot use. Several noted limitations warrant further study, including expanding the types and volume of challenge items used, sampling in a more diverse group (especially with respect to educational attainment), and need for more targeted focus on reasons for non-compliance among participants who fail to utilize the chatbot feature. These lines of inquiry would enhance understanding of the potential utility of chatbots for survey completion and data quality.

Open Practices Statements None of the data or materials for the experiments reported here is available, and none of the experiments was preregistered.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01574-w.

References

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, 53–56(1), 63–70. https://doi.org/10.1016/j.jom.2017.06.001

Androutsopoulou, A., Karacapilidis, N., Loukis, E., & Charalabidis, Y. (2019). Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly*, *36*(2), 358–367. https://doi.org/10.1016/j.giq.2018.10.001

Anglim, J., Horwood, S., Smillie, L.D., Marrero, R.J., & Wood, J.K. (2020). Predicting psychological and subjective well-being from personality: A meta-analysis. *Psychological Bulletin, 146*(4), 279-323. https://doi.org/10.1037/bul00000226

Anson, I. G. (2018). Taking the time? Explaining effortful participation among low-cost online survey participants. *Research & Politics*, *5*(3), 2053168018785483. https://doi.org/10.1177/2053168018785483

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

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Researchin Higher Education*, *53*(2), 182–200. https://doi.org/10.1007/s11162-011-9251-2

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A Cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800–813. https://doi.org/10.3758/s13428-011-0081-0

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Bickmore, T.W., Mitchell, S.E., Jack, B.W., Paasche-Orlow, M.K., Pfeifer, L.M., & O'Donnell, J. (2010a). Response to a relational agent by hospital patients with depressive symptoms. *Interacting with Computers, 22*, 289–298. https://doi.org/10.1016/j.intcom.2009.12.001

Bickmore, T.W., Puskar, K., Schlenk, E.A., Pfeifer, L.M., & Sereika, S.M. (2010b). Maintaining reality: Relational agents for antipsychotic medication adherence. *Interacting with Computers, 22*, 276–288. https://doi.org/10.1016/j.intcom.2010.02.001

Bowling, N.A., Huang, J.L., Bragg, C.B., Khazon, S., Liu, M., & Blackmore, C.E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111*(2), 218–229. https://doi.org/10.1037/pspp0000085

Buchanan, E., & Schofield, J.E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods, 50*(3), 2586–2596. https://doi.org/10.3758/s13428-0180193506

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, *9*(2), 48–57.

Chakrabarti, C., & Luger, G. F. (2015). Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metrics. *Expert Systems with Applications*, *42*(20), 6878–6897. https://doi.org/10.1016/j.eswa.2015.04.067

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2019). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*, *92*, 539–548. https://doi.org/10.1016/j.future.2018.01.055

Clément, M., & Guitton, M. J. (2015). Interacting with bots online: Users' reactions to actions of automated programs in Wikipedia. *Computers in Human Behavior*, *50*, 66–75. https://doi.org/10.1016/j.chb.2015.03.078

Cohen, J. (Ed.). (1988). *Statistical power analysis for the behavioral sciences ED 2nd ed*. Lawrence Erlbaum Associates.

Collins, D. (2003). Pretesting survey instruments: an overview of cognitive methods. *Quality of Life Research*, *12*(3), 229–238. https://doi.org/10.1023/A:1023254226592

Conrad, F. G., Schober, M. F., Jans, M., Orlowski, R. A., Nielsen, D., & Levenstein, R. (2015). Comprehension and engagement in survey interviews with virtual agents. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01578

Crowston, K. (2012). Amazon Mechanical Turk: A research tool for organizations and information systems scholars. In: Bhattacherjee A., Fitzgereald B. (eds) Shaping the future of ICT Research. Methods and Approaches. IFIP Advances in Information and Communication Technology, volume 389. Springer. https://doi.org/10.1007/978-3-642-35142-6_14

Cummins, R.A., Capic, T., Fuller-Tyszkiewicz, M., Hutchinson, D., Olsson, C.A., & Richardson, B. (2018). Why self-report variables inter-correlate: The role of homeostatically protected mood. *Journal of Well-Being Assessment, 2*: 93–114. https://doi.org/10.1007/s41543-018-0014-0

Cummins, R. A., Eckersley, R., Pallant, J., van Vugt, J., & Misajon, R. (2003). Developing a national index of subjective wellbeing: The Australian Unity Wellbeing Index. *Social Indicators Research*, *64*(2), 159–190. https://doi.org/10.1023/A:1024704320683

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

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

Ellis, R.D., & Allaire, J. (1999). Modeling computer interest in older adults: The role of age, education, computer knowledge, and computer anxiety. *Human Factors, 41*(3), 345–355. https://doi.org/10.1518/001872099779610996

Ehrhart, M. G., Ehrhart, K. H., Roesch, S. C., Chung-Herrera, B. G., Nadler, K., & Bradshaw, K. (2009). Testing the latent factor structure and construct validity of the Ten-Item Personality Inventory. *Personality and Individual Differences*, *47*(8), 900–905. https://doi.org/10.1016/j.paid.2009.07.012

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507–521. https://doi.org/10.2307/2331838

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, *4*(2), e19. https://doi.org/10.2196/mental.7785

Garbarski, D., Schaeffer, N.C., & Dykema, J. (2015). The effects of response option order and question order on self-rated health. *Quality of Life Research, 24*(6), 1443–1453. https://doi.org/10.1007/s11136-014-0861-y

Gardiner, P.M., McCue, K.D., Negash, L.M., Cheng, T., White, L.F., Yinusa-Nyahkoon, L., Jack, B.W., & Bickmore, T.W. (2017). Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: A feasibility randomized control trial. *Patient Education and Counseling, 100*(9), 1720–1729. https://doi.org/10.1016/j.pec.2017.04.015

Georgescu, A.-A. (2018). Chatbots for education - trends, benefits and challenges. *ELearning & Software for Education*, *2*, 195–200. https://doi.org/10.12753/2066-026X-18-097

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, *66*(1), 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, *59*(2), 93–104. https://doi.org/10.1037/0003-066X.59.2.93

Hamby, T., & Taylor, W. (2016). Survey satisficing inflates reliability and validity measures: An experimental comparison of college and Amazon Mechanical Turk samples. *Educational and Psychological Measurement*, *76*(6), 912–932. https://doi.org/10.1177/0013164415627349

Hasler, B. S., Tuchman, P., & Friedman, D. (2013). Virtual research assistants: Replacing human interviewers by automated avatars in virtual worlds. *Computers in Human Behavior*, *29*(4), 1608–1616. https://doi.org/10.1016/j.chb.2013.01.004

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8), 1–28. https://doi.org/10.18637/jss.v042.i08

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR MHealth and UHealth*, *6*(11), e12106. https://doi.org/10.2196/12106

International Wellbeing Group. (2013). *Personal Wellbeing Index: 5th Edition* (p. 38). Retrieved from Centre on Quality of Life, Deakin University website: http://www.acqol.com.au/uploads/pwi-a/pwi-a-english.pdf

Kaplin, S., Luchman, J., & Mock, L. (2013). General and specific question sequence effects in satisfaction surveys: Integrating directional and correlational effects. *Journal of Happiness Studies, 14*, 1443-1458. https://doi.org/10.1007/S10902-012-9388-5

Keeling, K., McGoldrick, P., & Beatty, S. (2010). Avatars as salespeople: Communication style, trust, and intentions. *Journal of Business Research*, *63*(8), 793–800. https://doi.org/10.1016/j.jbusres.2008.12.015

Krosnick, J. A. (1999). Survey research. *Annual Review Of Psychology*, *50*, 537–567.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, (70), 29–44. https://doi.org/10.1002/ev.1033

Lenzner, T. (2012). Effects of survey question comprehensibility on response quality. *Field Methods, 24*(4), 409-428. https://doi.org/10.1177/1525822X12448166

Lenzner, T., Kaczmirek, L., & Galesic, M. (2011). Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research*, *23*(3), 361–373. https://doi.org/10.1093/ijpor/edq053

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*, *24*(7), 1003–1020. https://doi.org/10.1002/acp.1602

Lietz, P. (2010). Research into questionnaire design. A summary of the literature. *International Journal of Market Research*, *52*(2), 249–272. https://doi.org/10.2501/S147078530920120X

Lucas, G.M., Rizzo, A., Gratch, J., Scherer, S., Stratou, G., Boberg, J., & Morency, L.-P. (2017). Reporting mental health symptoms: Breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI, 4*: 51. https://doi.org/10.3389/frobt.2017.00051

Ly, K. H., Ly, A.-M., & Andersson, G. (2017). A fully automated conversational agent for promoting mental well-being: A pilot RCT

using mixed methods. *Internet Interventions*, *10*, 39–46. https://doi.org/10.1016/j.invent.2017.10.002

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Meade, A.W., & Craig, S.B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. https://doi.org/10.1037/a0028085

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals Of Cardiac Anaesthesia*, *22*(1), 67–72. https://doi.org/10.4103/aca.ACA_157_18

Myszkowski, N., Storme, M., & Tavani, J.-L. (2019). Are reflective models appropriate for very short scales? Proofs of concept of formative models using the Ten-Item Personality Inventory. *Journal of Personality*, *87*(2), 363–372. https://doi.org/10.1111/jopy.12395

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital Health, 5*, 1-12. https://doi.org/10.1177/2055207619871808

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Paas, L. J., Dolnicar, S., & Karlsson, L. (2018). Instructional manipulation checks: A longitudinal analysis with implications for MTurk. *International Journal of Research in Marketing*, *35*(2), 258–269. https://doi.org/10.1016/j.ijresmar.2018.01.003

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Pereira, J., & Díaz, Ó. (2019). Using health chatbots for behavior change: A mapping study. *Journal of Medical Systems*, *43*(5), 135. https://doi.org/10.1007/s10916-019-1237-1

Pew Research Center (2016). *Smartphone ownership and Internet usage continues to climb in emerging economies*. https://www.pewresearch.org/global/2016/02/22/Internet-access-growing-worldwide-but-remains-higher-in-advanced-economies/

Philip, P., Micoulaud-Franchi, J.-A., Sagaspe, P., De Sevin, E., Olive, J., Bioulac, S., & Sauteraud, A. (2017). Virtual human as a new diagnostic tool, a proof-of-concept study in the field of major depressive disorders. *Scientific Reports, 7*: 42656. https://doi.org/10.1038/srep42656

Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents.

Richardson, B., Fuller Tyszkiewicz, M., Tomyn, A., & Cummins, R. (2016). The psychometric equivalence of the Personal Wellbeing Index for normally functioning and homeostatically defeated Australian adults. *Journal of Happiness Studies*, *17*(2), 627–641. https://doi.org/10.1007/s10902-015-9613-0

Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of the Internet in psychological research: comparison of online and offline questionnaires. *Cyberpsychology & Behavior*, *6*(1), 73–80.

Roßmann, J., Gummer, T., & Silber, H. (2018). Mitigating satisficing in cognitively demanding grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics and Methodology*, *6*(3), 376–400. https://doi.org/10.1093/jssam/smx020

Smyth, J.D., & Olson, K. (2018). The effects of mismatches between survey question stems and response options on data quality and responses. *Journal of Survey Statistics and Methodology, 7*(1), 34–65. https://doi.org/10.1093/jssam/smy005

Statistica (2021). *Average daily time spent per capita with the Internet worldwide from 2011 to 2021*. https://www.statista.com/statistics/1009455/daily-time-per-capita-Internet-worldwide/

Tait, J., & Wilks, Y. (2019). Anniversary article: Then and now: 25 years of progress in natural language engineering. *Natural Language Engineering*, *25*, 405–418. https://doi.org/10.1017/S1351324919000081

Thorne, C. (2017). Chatbots for troubleshooting: A survey. *Language and Linguistics Compass*, (10). https://doi.org/10.1111/lnc3.12253

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

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 0706743719828977. https://doi.org/10.1177/0706743719828977

Van der Groot, M.J., & Pilgrim, T. (2020). Exploring age differences in motivations for and acceptance of chatbot communication in a customer service context. In: Følstad A. et al. (eds) *Chatbot research and design*. Conversations 2019. Lecture Notes in Computer Science, vol 11970. Springer, Cham. https://doi.org/10.1007/978-3-030-39540-7_12

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. https://doi.org/10.1093/ijpor/eds021

Ward, M. K., & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior*, *48*, 554–568. https://doi.org/10.1016/j.chb.2015.01.070

Zumstein, D., & Hundertmark, S. (2017). Chatbots—an interactive technology for personalized communication, transactions and services. *IADIS International Journal on WWW/Internet*, *15*(1), 96–109.