methods, data, analyses | 2026, pp. 1-36 DOI: 10.25521/mda.772

Measuring the Response Quality of Open-Ended Questions in a Demographic Web Survey Using Linguistic Complexity Features

Xiao Xu[1,2], Antal van den Bosch[3], Anne H. Gauthier[1,2], Gert Stulp[2]

[1] *Netherlands Interdisciplinary Demographic Institute (NIDI-KNAW)*

[2] *University of Groningen*

[3] *Utrecht University*

Abstract

Open-ended questions (OEQs) are important survey tools for social scientists, but their response quality is often disputed due to the additional load they impose on the respondent. To find ideally worded questions and survey strategies that encourage high-quality responses to OEQs, the quality of textual responses is often assessed. Response length and response latency are often used as measures of response quality, but they do not provide enough information on interpretability and richness of the responses. In this study, we propose a novel way of evaluating the data quality of open-ended responses by leveraging approaches from natural language processing, measuring different linguistic complexity features of responses. Using various automatically generated linguistic features, we compared the quality of responses to distinctly worded sets of questions related to people's uncertainty about their intention of having children. Overall, we found that the different wording of questions may affect responses on different aspects of linguistic complexity, which canonical indicators fail to reveal. In addition, we found that the variance in response quality could be attributed to both respondent characteristics and different versions of questions. These findings offer practical strategies for incorporating OEQs into a large-scale demographic survey, as well as providing a new perspective in evaluating responses to OEQs in future surveys.

Keywords: open-ended question, web survey, response quality, linguistic complexity, natural language processing

How do people think about having children in the future, and how certain are they about their plans? Demographers try to answer such questions by measuring people's fertility preferences and intentions in surveys and use such responses as predictors of fertility behavior (Freitas & Testa, 2017; Liefbroer, 2009; Morgan & Rackin, 2010). This research has shown that many respondents are uncertain about their intention and that the predictive power of these responses on fertility behaviors is limited. To get a better understanding of these patterns, open-ended questions (OEQs) have been adopted in a number of qualitative follow-up studies of fertility surveys (Schatz & Williams, 2012; Speizer et al., 2009; Staveteig et al., 2017). This body of research has helped to identify the impact of various factors on fertility behavior, such as gender inequality and religious beliefs.

However, qualitative studies are limited in scale and cannot be extended to wider populations with the same degree of certainty as quantitative ones (Ochieng, 2009). Ideally, we would be able to process responses from many individuals from representative samples to get insight into fertility intentions and uncertainty. This is now possible thanks to advancements in Natural Language Processing (NLP) technology, which allows for automated analysis of large-scale open-ended responses (Gweon et al., 2017; He & Schonlau, 2020). On this account, OEQs in a large-scale survey may offer us new insight into our understanding of fertility intentions, as well as help advance theories on fertility uncertainty (Bhrolcháin & Beaujouan, 2019).

OEQs in surveys often suffer from low-quality responses or high non-response rates (Schuman & Presser, 1996). To extract knowledge from responses to OEQs, a first step is assessing the response quality, i.e., the responses should be substantive and informative. Since there is no precedent in using OEQs on fertility issues in large-scale self-administered online surveys, we believe it is critical to ensure adequate response quality to our questions first.

Many researchers have studied the characteristics that could affect the response quality of online OEQs. From the survey and questionnaire design perspective, characteristics such as the order of questions (Tourangeau et al., 2004), text box size (Smyth et al., 2009) and (in online surveys) interactive web features, such as motivational statements and follow-up probes (Oudejans & Christian, 2010), were found to be important in determining response quality. On the respondent side, the effects of respondent interest (Holland & Christian, 2009; Oudejans & Christian, 2010), age and education (Scholz & Zuell, 2012), and devices used to complete the surveys (Peytchev & Hill, 2010; Struminskaya et al., 2015) have been linked to response quality as well.

Direct correspondence to
Xiao Xu, Netherlands Interdisciplinary Demographic Institute (NIDI-KNAW),
Lange Houtstraat 19, 2511 CV The Hague, The Netherlands
E-mail: xu@nidi.nl

Although these studies have shown robustly which factors matter most in determining response quality, there are two limitations to this body of research. First, these studies mostly relied on process-generated features or descriptive attributes, such as response length or response time. Second, in studies that looked beyond process-generated indicators, such as Smyth et al. (2009), Oudejans and Christian (2010) and Schmidt et al. (2020), human coding is necessary. While this is not a limitation as such, it does mean that this method does not easily extend to larger scale projects. Previous works also did not cover the internal structure and writing style of responses.

This study addresses three research questions and extends previous research in the following ways. Our first question is: Do NLP-generated measures of linguistic complexity offer added value over traditional response quality measures? To this end, we introduce and test a series of linguistic complexity features that are well-known in the field of natural language processing but have not yet been considered in survey research. The responses are evaluated based on six indicators: response length (word count), non-substantive response rate, response time, and three additional linguistic complexity features. Our second question is: How does the formulation of OEQs affect response quality? We developed multiple versions of OEQs and fielded them in a large-scale web survey. Specifically, we adopted two specific interactive strategies that improve responses in online OEQs: motivational statements and follow-up probes (Oudejans & Christian, 2010). In addition, we varied the specific formulation of our questions. Our final contribution is that we examine to what extent respondent characteristics affect the quality of the response. This is important because if there are large differences in response quality across respondents, conclusions derived on the basis of these answers may not be substantive but reflect underlying factors such as motivation or cognitive ability (Krosnick, 1991; Roberts et al., 2019). This allows us to address the question: do respondent characteristics or design features explain more variation in response quality? The ultimate goal of this research project is to get a better understanding of people's uncertainty about fertility intentions, but the focus of the current paper is the quality of answers to OEQs.

Background

Fertility Intention and Uncertainty

Fertility preferences and intentions are routinely measured by closed questions in many series of large-scale demographic surveys, including the Demographic and Health Surveys (DHS; Corsi et al., 2012), the British General Household Survey (GHS; Hough & Mayhew, 1983), and the Generations & Gender Surveys (GGS;

United Nations, 2005), which covers Europe and other developed countries. In the current round of the GGS, a survey that focuses on families, life course trajectories, and gender relations (Gauthier et al., 2018), the following two questions are asked:

- Do you intend to have a/another child during the next three years? (definitely yes, probably yes, don't know, probably not, definitely not)
- Supposing you do not have a/another child during the next three years, do you intend to have any (more) children at all? (definitely yes, probably yes, don't know, probably not, definitely not)

Many researchers have attempted to use such fertility intentions as a predictive factor for whether and when people have children (Freitas & Testa, 2017; Liefbroer, 2009; Morgan & Rackin, 2010) but with limited success. This has led to controversy over their utility. Peoples' uncertainty about their intentions may be a reason for this limited predictive ability; already decades ago it was established that people report large uncertainty about their intentions (Morgan, 1982; Oakley, 1981), and this has changed little across time (Bhrolcháin & Beaujouan, 2011). Despite this, uncertainty has been relatively understudied. Bhrolcháin and Beaujouan (2019) discovered that, among demographic articles on fertility intentions/expectations in low-fertility countries between 2011 and 2015, only 28% reported frequency information or analytical results relating to uncertain responses. Agadjanian (2005) similarly concluded that fertility intentions reported in surveys are "fraught with ambiguity and uncertainty" (p. 617). As a result, although uncertainty has been provided as an option in surveys, we still know little about their qualitative meaning. To gain meaningful insights into attitudes about certainty, a number of demographers have employed mixed methods and conducted follow-up studies (Schatz & Williams, 2012; Staveteig et al., 2017). While these studies tried to tackle the problem on a small scale in-depth interviews, they are also limited by their inherent features, such as a costly and time-consuming analysis and samples that do not necessarily generalize well to the population at large (Queirós et al., 2017). Nowadays, the development of NLP techniques has made automatic analysis on responses to OEQs possible (He & Schonlau, 2020; Schonlau & Couper, 2016; Züll, 2016). These novel techniques can help us get more insight into peoples' (un)certainty in their fertility preferences, which inspired us to incorporate OEQs in a larger-scale survey and apply the new techniques.

OEQs and Underlying Factors of Response Quality

OEQs are useful tools in social surveys, as they allow respondents to offer detailed information that is not constrained by pre-defined response categories

(Schuman & Presser, 1979). Specifically, Fernandez et al. (2016) mentioned that OEQs are particularly useful when researchers are not able to give a finite set of answers and want to obtain top-of-the-head answers from respondents, which is the case for our study of (un)certainty about the intention to have a(nother) child. At the same time, OEQs also require more efforts from respondents and are more burdensome, as respondents need to compose their own response instead of choosing from pre-defined options. Krosnick (1991) claimed that some respondents may be using a satisficing strategy, where they give an answer that takes the lowest possible cognitive effort, which lowers the quality of responses and leads to a higher nonresponse rate (Reja et al., 2003; Scholz & Zuell, 2012) that may be particularly problematic for burdensome OEQs.

Much research has been conducted on improving the response quality of OEQs. Barth and Schmitz (2021) found that the response quality is jointly produced by respondents and interviewers. Factors from the surveys themselves have an impact on the response quality, and many researchers have explored the influence of decisions in designing and presenting questions on response behaviors. It is known that respondents take cues from visual elements of online surveys (Couper et al., 2004; Couper et al., 2007), and according to the interactive nature of online surveys, many specific features, such as text box size (Smyth et al., 2009; Zuell et al., 2015), question wording (Y. Chen, 2017; Reja et al., 2003), motivational statements (Oudejans & Christian, 2010; Smyth et al., 2009), and follow-up probes (Holland & Christian, 2009; Oudejans & Christian, 2010), all impact response quality and may be utilized to motivate respondents to provide answers of better quality.

At the respondent level, various individual characteristics also predict answer quality. Previous research has extensively shown that older and less educated respondents have poorer quality responses (Couper & Kreuter, 2013; Denscombe, 2008; Yan & Tourangeau, 2008). In the context of online surveys, people on mobile devices are also found to give shorter answers compared to those on computers (Mavletova, 2013; Revilla & Ochoa, 2016). In addition, respondents' personal topic interest and attitude are also found to be impactful (Holland & Christian, 2009; Zhou et al., 2017). Large differences in response quality depending on respondent characteristics inevitably constrain what can be learned from such responses.

Measuring the Response Quality of OEQs

To properly assess the response quality of OEQs, researchers must find appropriate indicators. Response length (counted by number of words) is the most common indicator of response quality in methodological studies on OEQs, and more words are often believed to be associated with better answers (Denscombe, 2008; Mavletova, 2013). However, this is not always true for open-ended attitude ques-

tions (Schmidt et al., 2020), as there may be an inherent trade-off between the extensiveness and accuracy of a response: the accuracy of a response increases with its length until a certain point, then starts to drop. Therefore, it is insufficient to use response length alone as an indicator.

Response time has also been increasingly used as an indicator of response quality. Shorter response time often indicates lower cognitive effort made by respondents, which could lead to lower data quality (Greszki et al., 2015; Smyth et al., 2009). This pattern was also found on OEQs, where shorter response times were linked to less precise and more nonsense answers (Revilla & Ochoa, 2015). However, response time may also be affected by other factors such as unclear or flawed questions (Bassili & Scott, 1996), for instance, when the question is hard to understand due to wording not being concise or multiple topics being addressed. Therefore, it is suggested not to use response time as the sole indicator of response quality (Schmidt et al., 2020).

Several studies have explored the richness and interpretability of responses to OEQs using different measures. Mavletova (2013) and Revilla and Ochoa (2016) used a non-substantive rate as an indicator[1], while Smyth et al. (2009) and Oudejans and Christian (2010) annotated the answers with the number of themes mentioned. Schmidt et al. (2020) used the interpretability of the responses as a content-related indicator of response quality. All these approaches relied on human coders to manually code answers; as a large-scale online survey including thousands of respondents, it is not always feasible to manually code all answers. Instead, we propose to operationalize the richness and interpretability of responses with linguistic complexity features that are used in NLP.

Various linguistic features that can be automatically calculated have been developed as measures of text complexity. Norris and Ortega (2009) suggested length-based metrics, subclausal complexity, and subordinate occurrence as the three subconstructs for global syntactic complexity. Complementing this, measures have been developed to assess lexical complexity, such as the Measure of Textual Lexical Diversity (MTLD; McCarthy & Jarvis, 2010). These measures, among many others, have been implemented in open-source code by Vajjala and Meurers (2012) for English texts and have shown to be of reasonable quality on large-scale corpora (Alexopoulou et al., 2017).

Indeed, NLP is a rich field in which many more indicators of linguistic complexity have been developed than the few that are applied in evaluating OEQs. These indicators have been applied to assess the readability and writing quality of texts (Alexopoulou et al., 2017; Crossley, 2020; Lu et al., 2019) but have not

[1] Revilla and Ochoa (2016, p. 2504) defined non-substantive answers as "answers for which they do not need to think, like 'don't know' answers (DK) or nonsense answers (i.e., just some numbers or letters without any meaning, or a few words that do not answer at all the question)".

yet been used to assess the quality of open-ended answers. These developments may help us get a better sense of the quality of the responses.

Linguistic Complexity Features

We use the term "linguistic complexity" as the complexity of a text from a linguistics point of view (Vajjala & Meurers, 2012). Linguistic complexity measures have seen a variety of applications, such as mood and sentiment analysis (Joshi et al., 2014), text simplification (Dell'Orletta et al., 2011; Saggion et al., 2015), automatic translation (Vanmassenhove et al., 2021), and also the assessment of text readability for language learners (X. Chen & Meurers, 2019; Kleijn, 2018; Pilán et al., 2016; Vajjala & Meurers, 2012).

The link between linguistic complexity and cognitive effort could also contribute to our understanding of the quality of responses to OEQs. According to the satisficing theory (Krosnick, 1991), respondents incompletely pass through the cognitive process of answering an OEQ to reduce cognitive burden. Although it is usually assumed that satisficing results in shorter and less detailed answers, there is no empirical evidence for that link. Meanwhile, various linguistic complexity features in writing have been found associated with cognitive impairment (Kemper et al., 1993; Sand Aronsson et al., 2021; Tsantali et al., 2013), which makes them potential indicators of cognitive efforts in producing responses in surveys as well.

Traditionally, linguistic complexity has been measured by features that are easy to calculate, such as word length and sentence length (Flesch, 1948). Recent development in computational linguistics has allowed the extraction of more sophisticated features with automated tools. A prominent example for the Dutch language is T-scan (Pander Maat et al., 2014), a computational linguistics toolkit that offers various variables to represent the linguistic complexity of Dutch texts. Vajjala and Meurers (2012) grouped these extracted features into two categories, lexical and syntactical, which are often referred to as "lexical richness" and "syntactic complexity" in the readability and second language acquisition (SLA) literatures.

Although T-scan computes over 400 features, work by Kleijn (2018) on a readability index for Dutch has identified empirically the most prominent linguistic complexity features among them, which were (a) word frequency, (b) content words per clause, and (c) maximum syntactic dependency length (SDL). In this study, we will adopt these measures and explain them in detail below. Although linguistic features have been applied to examine wording effects in surveys (Holleman, 2000), there is still no precedent in analyzing responses to OEQs with them.

Data and Methods

Data Collection

The data in this study come from a survey carried out on the LISS (Longitudinal Internet Studies for the Social Sciences) panel administered by CentERdata (Tilburg University, The Netherlands). The LISS panel covers a representative sample of Dutch individuals who participate in monthly Internet surveys, as it is based on a true probability sample of households drawn from the population register by Statistics Netherlands (CBS). Households that could not otherwise participate are provided with a computer and Internet connection. Only households in which at least one household member spoke Dutch are included. A longitudinal study consisting of ten core surveys is fielded in the panel every year, covering a large variety of domains including work, education, income, housing, time use, political views, values, and personality.

The study is based on the module "Social networks and fertility" (in Dutch: Sociale relaties en kinderkeuzes onderzoek) of the LISS panel, which was first fielded in 2018 and then again in 2021 when we included the OEQs. The module's objective was to investigate fertility intentions and attitudes in relation to people's personal networks [2]. In the first wave of this sample in 2018, among all the women in the LISS panel that were between the ages of 18 and 40 (N = 1,332), 758 women (56.9%) between the ages of 18 and 41 completed the survey. In February 2021, those women from the first wave who were still active in the LISS panel were invited to this second wave. In total, 596 participants were invited to this survey, and 464 women (77.9%) responded. Respondents were very similar to non-respondents based on a comparison of a range of measures that are collected for all respondents and are continuously updated, including birth year, position in household, number of children, marital status, region of living, income, educational level, and (migration) background. The largest but still small differences were observed in migration background: 72.0% of respondents in the second wave were native Dutch, compared to 62.1% of non-respondents[3]. Each respondent received €7.50 for completing the questionnaire, which was advertised as lasting around 15 minutes to finish. This compensation rate is twice as high as the standard rate of the LISS panel (€2.50 per 10 minutes). The survey was conducted in Dutch.

Ethical permission for the study was obtained from the ethical committee of sociology at the University of Groningen (ECS-201123). The dataset is available at dataarchive.lissdata.nl. The full questionnaire, linguistic features generated

[2] See https://doi.org/10.34894/EZCDOA for full details on the study, as well as Stulp (2021) and Buijs and Stulp (2022)

[3] See https://doi.org/10.34894/PQPGDH

by T-scan, and code used to link datasets and conduct analyses in this study are available at https://doi.org/10.34894/XXXSXJ.

Evaluating Responses

Our first research question concerned to what extent NLP-generated measures of linguistic complexity help evaluate the quality of open-ended responses, also in comparison to traditional measures. To capture traditionally used measures of response quality, we used response length (total number of words), non-substantive response rate, and response time to evaluate responses. Although the questionnaire did not allow empty responses, which prevented the presence of "traditional" non-responses, there are still answers that did not offer any knowledge or information. Non-substantive responses were jointly coded by the first and the last author of this paper, and the latter is a native Dutch speaker. As is defined by Revilla and Ochoa (2016), we coded answers that did not require thinking, such as 'don't know' or nonsense symbols, as non-substantive responses. Another common format of non-substantive responses in follow-up questions is "please refer to the first question", which also did not add any new information.

We add three linguistic complexity features: word frequency, content words per clause, and maximum syntactic dependency length, that are commonly accepted to represent data quality in previous studies (Kleijn, 2018). These measures were calculated through the automatic text analysis toolkit T-scan (Pander Maat et al., 2014).

Indicator 1: Word Frequency

Word frequency is one of the most commonly used indicators of lexical richness (X. Chen & Meurers, 2016; Jensen, 2009) and was also the strongest predictor in the work of Kleijn (2018). It concerns how commonly used the words used in a text are: more frequent words are easier to comprehend and to process than low frequent words. Within T-Scan, we use the Dutch subtitle corpus SUBTLEX-NL (Keuleers et al., 2010) to calculate word frequency based on film and television subtitles. Word frequency is negatively related to complexity: the higher the word frequency of a text is, the easier it is to compose and comprehend it.

Some examples from our collected responses may help better understand the concept of word frequency. Among the responses with the highest word frequency, more general and common answers are found, such as "I don't know", "I am still too young" or "Not too much and not too little[4]".

In contrast, we found more varied answers among the ones with the lowest word frequency. One respondent said:

[4] Originally in Dutch, respectively: Ik weet het niet, Ik ben nog te jong, Niet te veel en niet te weinig

> "My partner was sterilized years ago and I would have to go to an IVF clinic for a child.[5]"

which described a more concrete scenario with many less commonly used words. Another example is:

> (In English) "Getting older it is nice to have older children. But having children also takes away a lot of freedom so not sure yet if I want children."

Which used English instead of Dutch in this supposedly Dutch-only survey. As a result, this text logged a very low word frequency score, since English words are, by definition, rare in a Dutch corpus.

Indicator 2: Content Words Per Clause

The number of content words per clause (i.e., a sentence or sub-sentence containing a main verb), instead, is positively related to the general complexity in the experiment of Kleijn (2018). Content words, in contrast with function words, refer to words that carry meaning. Nouns, verbs, adjectives, and adverbs are usually considered content words, while words such as pronouns, prepositions, and conjunctions are excluded. It also indicates the lexical richness of a text. The number of content words per clause is related to propositional density: the amount of information conveyed in an object per unit element. Experiments have shown that propositional density is related to cognitive effort in language production (Smolík et al., 2016).

In this study, we calculated the number of content words per (sub)sentence, excluding adverbs. Our examples were also consistent with the idea that content word density and information richness are linked (Johansson, 2008). Here is a response that contains 13 content words per clause, the highest number in our data (content words marked in bold):

> "I am not yet able to **raise** a **child** properly at this **moment** without **stopping** my education and/or **losing most** of my **freedom** and **social circle**.[6]"

The answers that contain few or no content words, on the other hand, are often short, simple expressions or even nonsensical answers, such as "2 is

---

[5] In Dutch: Mijn partner is járen geleden gesteriliseerd en ik zou dan naar een IVF-kliniek moeten voor een kind.

[6] In Dutch (content words in bold): Ik ben nog niet in **staat** om een **kind goed** op te **voeden momenteel**, zonder te **stoppen** met mijn **opleiding** en/of het **grootste deel** van mijn **vrijheid** en **sociale kring** te **verliezen**.

enough[7]", "These are enough for me[8]" or "Nothing[9]", which do not contain any content words according to our definition.

Indicator 3: Maximum Syntactic Dependency Length

Maximum syntactic dependency length (SDL) refers to the maximum distance between a syntactic head (i.e., a verb) and its dependent (i.e., its subject). In contrast with the previous features, it indicates complexity at the syntactic level. It is also related to the widely used text complexity indicator of sentence length (Rudnicka, 2018), as the distance between the head and the dependent cannot be longer than the sentence. However, it is possible for a long sentence to have short SDL. In the experiments of Kleijn (2018), they demonstrated that maximum SDL is a better predictor of text complexity level than the traditional predictor sentence length.

In our examples, the relation between the SDL and sentence length is also apparent. The following answer has a maximum syntactic dependency length of 14, one of the longest in our data. The head and its furthest dependent are marked in bold:

> "**I** like to have time for myself, not caring and want to **keep** myself at number one.[10]"

On the other hand, many responses did not contain a full sentence but only a phrase, such as "No partner[11]", "Too old[12]" or "No need[13]", which lack a main verb in the first place. The SDL is considered zero in that scenario.

Alternative Versions of OEQs

In our second research question we address to what extent the formulation of the OEQ would impact the quality of the responses. Respondents received two different questions, where the first question had two variants and the second question three. Both questions followed directly after a widely used close-ended question on fertility intentions ("Do you intend to have a/another child during the next three years?").

---

7 In Dutch: Genoeg aan 2

8 In Dutch: Deze zijn genoeg voor mij

9 In Dutch: niets

10 In Dutch (the furthest pair of head and dependent in bold): **Ik** heb graag tijd voor mezelf, ben niet zorgzaam en wil mezelf op nummer 1 **houden**.

11 In Dutch: Geen partner

12 In Dutch: Te oud

13 In Dutch: Geen behoefte aan

Our first version of OEQ number one was[14]:

Q1.1 Original: *Can you tell us more about what makes you (un)certain about whether or not to have children?*

Zaller and Feldman (1992) argued that most people possess opposing considerations on most issues and answer survey questions based on ideas that happen to be salient at that moment. Because fertility intentions are highly dynamic (Bhrolcháin & Beaujouan, 2019) and the reported fertility intention of a respondent could change in the same survey (Staveteig et al., 2017), it was particularly interesting to see how an adaptive reminder would change respondents' behavior in answering the OEQs.

Although very little research examined the versatility of responses in one survey, we found the challenge comparable to the "seam effect" found in panel surveys (Callegaro, 2008). Proactive dependent interviewing, which reminds respondents of their previous answers on the same topic, is a technique that produces more accurate and stable data in panel surveys (Jäckle & Eckman, 2020). Therefore, in the second version of the question, we add a prompt to remind respondents of their choice in the previous question on short-term fertility intention. By providing this reminder, we also expect the answers to be more consistent with the closed questions.

Q1.2 Adaptive: *You answered the previous question "Do you plan to have a child in the next three years?" with [definitely yes / probably yes / unsure / probably not / definitely not/ don't know]. Can you tell us more about what makes you (un)certain about whether or not to have children?*

We also evaluated the added value of a second follow-up question. Follow-up probes are commonly used in self-administered surveys (Groves et al., 2011) and are shown to encourage people to provide longer responses with more themes (Holland & Christian, 2009; Oudejans & Christian, 2010). As an exploratory experiment, we tested three versions of follow-up probes in this study:

Q2.1 "Three reasons" follow-up: *Can you mention the three main reasons why you are (un)certain about whether or not to have children?*

Q2.2 "Aspects" follow-up: *Which specific aspects in your life make you feel (un)certain about whether or not to have children?*

[14] All original questions were in Dutch. For the original Dutch materials, see https://doi.org/10.34894/PQPGDH

Q2.3 "More" follow-up: *Can you tell us more about whether you would like (more) children or not?*

Each respondent is randomly allocated one of the two versions of the first question and one of the three follow-ups, generating six versions of the questionnaire in total.

Respondent Characteristics

Our third research question addressed in what way respondent characteristics impacted response quality, also in comparison to the variation in questions asked. At the respondent level, we included age and education level as indicators of the ability of respondents, where the education levels were classified into low/medium/high according to the International Standard Classification of Education (ISCED) based on mapping to the Dutch education system by the United Nations Educational, Scientific and Cultural Organization (UNESCO)[15]. Whether respondents used a PC ($N = 235$) or mobile devices (smartphone or tablet) ($N = 198$) was also included.

Analytical Strategy

Before modeling the complexity of responses, we analyzed non-substantive responses and compared such responses for each OEQ. After removing the non-substantive responses, we examined response complexity (i.e., response length, response time, and linguistic features) across the different variants of the questions.

To examine simultaneously the effects of the different variants of the questions as well as respondent characteristics on response quality, we ran linear regressions for each measure of response complexity. To further evaluate the effects from each level (i.e., questions versus respondent level), we also build models exclusively on variables from each level and compare the R-squared value to assess their contribution to our dependent variables. We performed linear regression for all our features for comparability across measures and because most indicators approximated normal distributions (see Figure 1). The length and time of the response were clearly skewed to the right, which is why these two indicators were transformed via a logarithmic transformation, which resulted in distributions more similar to a normal distribution (see Figure 1 in Appendix A). The linear regression models were run with these logarithmic transformations for these measures. A logarithmic transformation of the maximum SDL

[15] http://uis.unesco.org/sites/default/files/documents/isced_mapping_2015_netherlands_0.xls

for responses to the second question still led to a highly skewed distribution. This is why we also performed negative binomial regression on count variables as a robustness check. Results were similar (see Tables 1 & 2 in Appendix B). An additional way in which we take into account the fact that distributions deviate from normality is that we present a non-parametric Wilcoxon rank sum test for the comparison of the medians between groups. We also present Spearman rank correlations between all our measures to examine to what extent multiple measures tap into similar underlying constructs. All models were implemented using R 4.1.2 (R Core Team, 2021).

Because answers to the first question may impact answers to the second question, we further examined whether the variant on Q1 affected response quality on the different variants of Q2, but we observed no effects of Q1 on Q2 (see Figure 2 in Appendix C). This also meant we could examine Q1 and Q2 separately.

Results

Descriptive Statistics

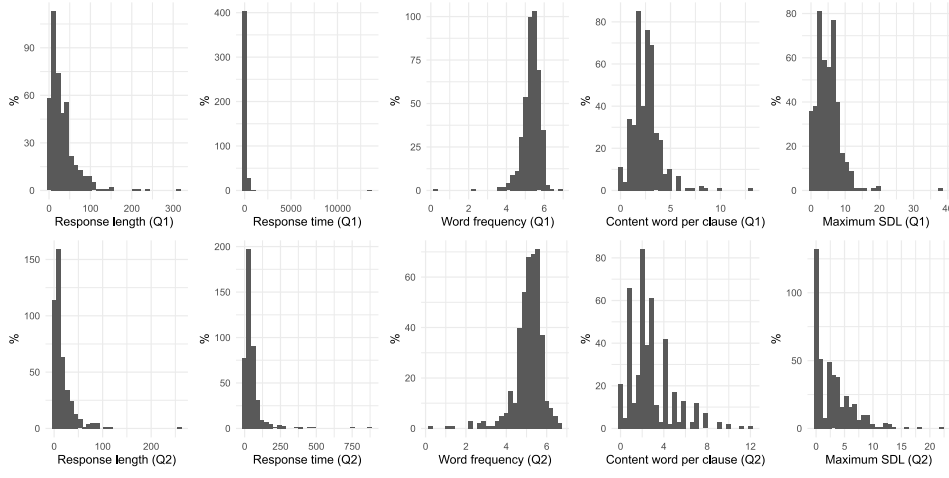

Figure 1 Histograms of responses to Q1 and Q2 on each indicator

In total, 433 women filled in the survey entirely, as 31 women who were pregnant at the time of the survey were not presented the OEQs. Women had a mean age of 32.5 (*SD* = 6.8). When examining education, we observed that 33 respondents had low education level (ISCED level 0-2), 159 had medium education (ISCED level 3-4) and 240 were highly educated (ISCED level 5 or higher). Most women did

the survey on a computer (N = 235) while the rest used a smartphone or tablet (N = 198).

The non-substantive response rate (i.e., "don't know," "NA," "nothing") was 2.1% for Q1 and 8.3% for Q2. Our survey recorded considerably lower non-substantive response rates than many similar previous studies (Holland & Christian, 2009; Oudejans & Christian, 2010). In addition to "pure" empty answers not being allowed in our survey design, this could be attributed to a higher compensation rate and usage of various interactive features (such as motivational statements). The mean length of the responses was 32.5 words (SD = 34.3) for Q1 and 16.5 (SD = 21.7) for the additional Q2. The response length for Q1 is about twice as many words as reported in a previous survey also using the LISS panel (Oudejans & Christian, 2010). Across both questions, respondents took on average 86 seconds to answer, had a score of 5.2 (SD = 0.6) on word frequency, used 2.7 (SD = 1.7) content words, and had a maximum SDL of 3.9 (SD = 3.6). See Table 1 for a breakdown of those numbers for each variant of the questions.

Table 1 Descriptive statistics of responses to Q1 and Q2, by group

Question	Response length	Response time (s)	Word frequency[1]	Content word	Maximum SDL
			Mean (*SD*)		
			Min/Median/Max		
Q1.1 "original"	36.1 (41.2)	164.0 (926)	5.25 (0.43)	2.77 (1.44)	5.2 (3.06)
	1/27/315	7.35/55.9/13481	3.67/5.29/6.26	0/2.62/10	0/4.5/20
Q1.2 "adaptive"	29.1 (25.5)	73.9 (72)	5.37 (0.57)	2.46 (1.28)	5.4 (3.82)
	1/21.5/151	7.32/49.6/481	0.32/5.43/6.94	0/2.33/13	0/4.75/38
Q2.1 "3 reasons"	14.1 (15)	63.6 (105)	4.92 (0.86)	3.33 (2.29)	2.33 (3.17)
	1/10/116	5.12/41.4/877	0.32/5.08/6.57	0/3/12	0/1/22
Q2.2 "aspects"	14.6 (27)	47.3 (63.7)	4.97 (0.66)	2.84 (2)	2.49 (3.42)
	0/7/258	6.38/24.4/495	2.28/5.06/6.75	0/2.25/9	0/1/18
Q2.3 "more"	20.7 (21.4)	47.6 (56.7)	5.39 (0.55)	2.2 (1.26)	3.93 (2.92)
	1/13/107	3.73/30.3/401	2.28/5.46/6.40	0/2/7	0/3.62/14

[1] Higher word frequency suggests lower complexity.

Overlap in Measures

To address our first research question regarding the utility of NLP-generated measures, we examined the relationships between the traditional indicators (length, time) and the linguistic complexity features on Q1 (see Figure 2).

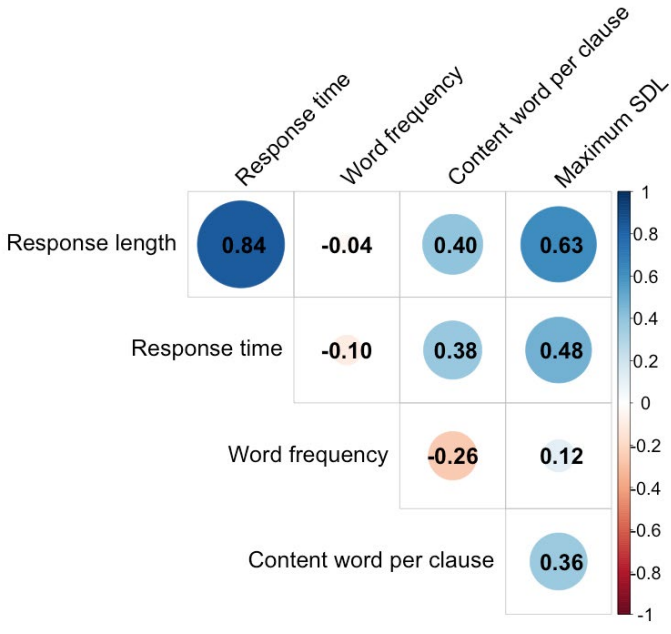

Figure 2 Spearman rank correlations between response quality measures.

We observe that response length is strongly positively associated with response time (r_s = 0.84, p < .001), moderately positively associated with maximum SDL (r_s = 0.63, p < .001) and content words per clause (r_s = 0.40, p < .001), and barely but negatively associated with word frequency (r_s = -0.04, p = .700). Similar, but weaker, patterns are observed for response time as it is also positively associated with the number of content words (r_s = 0.38, p < .001) and maximum SDL (r_s = 0.48, p < .001) and negatively with word frequency (r_s = -0.10, p = .040). This suggests that while longer answers, which take more time to write, are indicative of better answer quality, the moderate to low correlations suggest that longer answers that allow for more syntactic complexity do not necessarily imply the use of more diverse and sophisticated vocabulary. Small to moderate correlations between linguistic complexity measures further suggest that the chosen measures capture different aspects of the response process. Response time does not seem to provide much additional information over and above response length.

The Impact of the Formulation of the Questions on Response Quality

Non-Substantive Response

The non-substantive response on Q1 was minimal (9 out of 433; Table 2), and also such response did not differ for the two variants of Q1 (4 out of 213 versus 5 out of 220; Table 2). Non-response was higher for Q2 compared to Q1 (36 versus 9 out of 433, a significant difference χ^2 (1, N = 433) = 15.8, $p < .001$). Several respondents claimed that they answered the question before and thus had nothing to add by responding "See last answer[16]" or similar expressions. Also, across the different variants of Q2, there were differences in non-substantive response rates, ranging from 5.4% to 11.0% (Table 2). In pairwise comparison of versions of Q2, we found that the non-substantive response rate of Q2.1 ("three reasons") was significantly lower than Q2.3 ("more") (χ^2 (1, N = 433) = 4.7, $p < .05$). No other significant difference was found.

Table 2 Non-substantive response rates of Q1 and Q2, by group

Response rate	Non-substantive responses	Total responses	Non-substantive response rate
Q1.1 "original"	4	213	0.019
Q1.2 "adaptive"	5	220	0.023
All Q1	9	433	0.021
Q2.1 "3 reasons"	8	148	0.054
Q2.2 "aspects"	12	139	0.086
Q2.3 "more"	16	146	0.110
All Q2	36	433	0.083

Response Length, Time, and Linguistic Complexity

Wilcoxon rank sum tests revealed that for the two variants of Q1 (see Figure 3), significant median differences were observed only for word frequency (U = 18115, $p < .001$), and content words per clause (U = 26396, $p = .022$). This means that answers to the original question (Q1.1) contained more rarely used words and more content words. Thus, the adaptive variant of the question, in which respondents were reminded of their previous answer, led to answers with lower complexity at the lexical level and shorter response times. Linear regression with only the variant of the question as a predictor variable led to similar results (see Table 3). The highest, but still low, explained variance was for word frequency (2.8%), then content words (1.8%) with less than 1.0% of variance

[16] In Dutch: Zie voorgaand antwoord

explained for the other outcomes. Thus, from a linguistic complexity perspective, we found that Q1.1 offers better responses compared to Q1.2, although differences were minimal.

For Q2 (see Figure 4), we find larger differences across variants. Asking for "more" (Q2.3) compared to asking for "three reasons" (Q2.1) or "which aspects" (Q2.2) leads to longer answers (see Table 2; $U = 8948$, $p = .011$ compared to Q2.1 and $U = 7048$, $p < .001$ compared to Q2.2 respectively), higher word frequency ($U = 5716$, $p < .001$ compared to Q2.1 and $U = 4974$, $p < .001$ compared to Q2.2), lower number of content words ($U = 13920$, $p < .001$ compared to Q2.1 and $U = 11558$, $p = .041$ compared to Q2.2), and higher maximum SDL ($U = 6768$, $p < .001$ compared to Q2.1 and $U = 6540$, $p < .001$ compared to Q2.2). These findings suggest that while the response length is higher when respondents are simply asked for more to add, people tend to use more common, less sophisticated words in their answers. However, despite this lower lexical complexity, their answers have higher syntactic complexity as measured by maximum SDL.

Comparing variants Q2.1 ("three reasons") and Q2.2 ("aspects"), we observe that asking respondents for three reasons leads to slightly shorter answers ($U = 12038$, $p = .013$), longer response times ($U = 12912$, $p < .001$), and more content words ($U = 11609$, $p = .058$).

Results from the linear regression with only the variant of Q2 as a predictor corroborate these results (Table 4) and also show that the impact of the variants of Q2 was higher than for Q1. Roughly 10.7% of the variation in word frequency was explained by Q2, whereas this was 6.6% for the number of words, 5.6% for the number of content words, 1.9% for response time, and 1.0% for maximum SDL.

Conclusions about the preferred variant of Q2 are therefore more complex, as the performance on different indicators or dimensions conflicts with each other for different variants. This trade-off is also illustrated by the following examples. The first example is a response to the question Q2.3, which has one of the highest maximum SDL and lowest lexical complexity:

> "The chance that a next child has the same as our first child has ensured that we certainly do not go for a third, even though our 2nd child is perfectly healthy.[17]"

In contrast, here is an answer to Q2.1, with lower syntactic complexity (maximum SDL) but higher lexical complexity:

[17] Original Dutch text: "De kans dat een volgend kind hetzelfde heeft als ons eerste kindje heeft er bij ons ervoor gezorgd dat we zeker niet meer voor een derde gaan ondanks dat ons 2de kindje kerngezond is."

> "1. Money 2. Keeping a happy family seems difficult to me 3. Do not find life itself interesting enough to wish the same for a child[18]"

It is evident that, while the first example is longer, it only mentioned one reason for uncertainty about having children (the first child having a particular condition), while the second example listed three reasons as indicated in the question despite being shorter. This proves that linguistic complexity features offered us more insight into the quality of responses that could be overlooked by only comparing response length. Thus, from a linguistic complexity perspective, both Q2.1 ("three reasons") and Q2.3 ("more") have their own advantages respectively: responses to Q2.1 have more complex and diverse vocabulary, which means they are relatively rich in content, while responses to Q2.3 are longer and more syntactically complex, which means they are relatively more narrative.

The Impact of Respondent Characteristics on Response Quality

For Q1, where we found (very) small effects of the variant of the outcome on the question, we observed that respondent characteristics were more important than the variant. About 8.0% of variation was explained by respondent characteristics for both the number of words and maximum SDL, with 5.3% being explained in the number of content words (Table 3). Age was the most important respondent characteristic: older people tend to answer the question with fewer words, fewer content words on average, less complicated sentences, but slightly more advanced vocabulary. Lower educated people also gave shorter responses and spent less time on the question, but no effects of education were observed for the other measures.

For Q2, respondent characteristics were much less important, also compared to the variants of the question. Consistent with Q1, we observed that higher educated and younger people provided longer answers. For Q2, these groups also provided answers with higher maximum SDL. Platform characteristics had no effect in either Q1 or Q2.

[18] Original Dutch text: "1. Geld 2. Gelukkig gezinnetje blijven lijkt me moeilijk 3. Vind het leven zelf niet boeiend genoeg om een kind dat ook te gunnen"

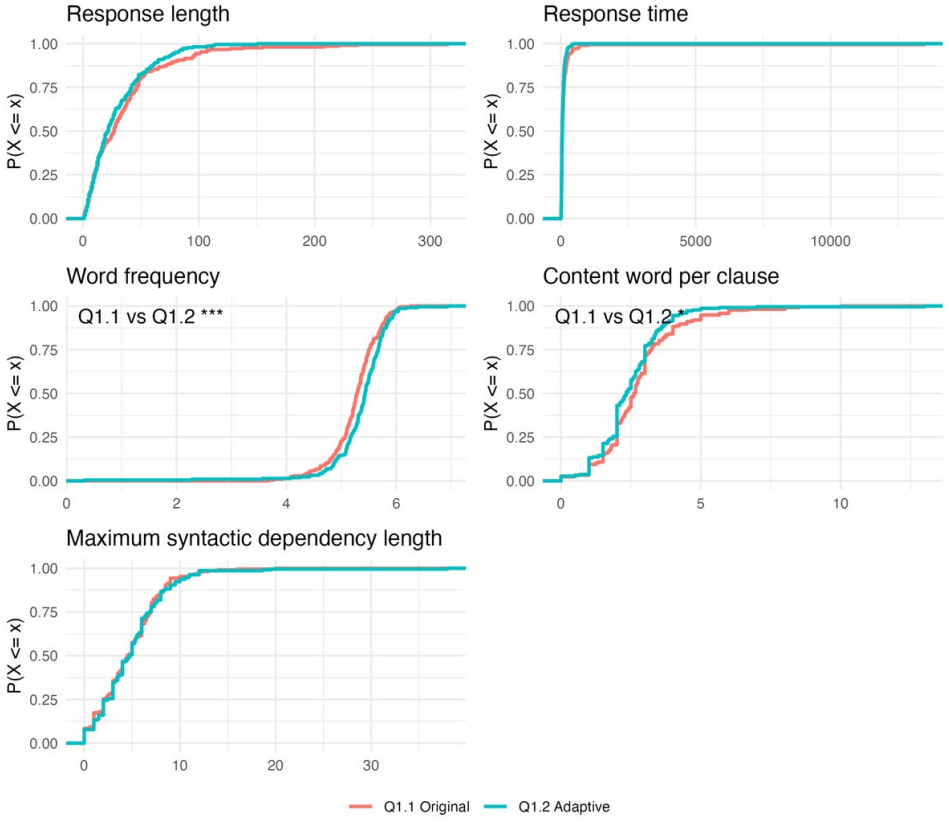

Figure 3 CDF graphs for the response length, time and 3 linguistic complexity features of the answers by Q1 versions.

Table 3 Regression on linguistic complexity features of responses to Q1 on question and respondent characteristics

Variable	Model 1 Response length	Model 2 Response time	Model 3 Word frequency	Model 4 Content words	Model 5 Maximum SDL
Q1.2: adaptive (ref: Q1.1)	−0.105 (0.09)	−0.184* (0.09)	0.147*** (0.04)	−0.335** (0.13)	0.293 (0.34)
Age	−0.035*** (0.01)	−0.011. (0.01)	−0.006* (0.00)	−0.044*** (0.01)	−0.125*** (0.03)
Education: low (ref: high)	−0.563** (0.17)	−0.351* (0.17)	0.121 (0.08)	−0.350 (0.24)	−0.944 (0.67)
Education: middle (ref: high)	−0.269** (0.09)	−0.188* (0.09)	0.017 (0.04)	0.136 (0.13)	−0.563 (0.36)
Mobile platform (ref: desktop)	−0.026 (0.09)	0.113 (0.09)	0.001 (0.04)	−0.111 (0.13)	0.127 (0.34)
Intercept	4.46*** (0.23)	4.589*** (0.22)	5.429*** (0.10)	4.275*** (0.32)	9.451*** (0.86)
AIC	1126.594	1104.721	453.861	1429.76	2089.443
R^2	0.086	0.024	0.036	0.066	0.056
R^2 for respondent -level model	0.085	0.016	0.007	0.053	0.080
R^2 for question -level model	0.001	0.008	0.028	0.013	−0.001

Note. ***p < .001, **p < .01, *p < .05, .p < .1
Standard errors in brackets.
Dependent variable was log-transformed.

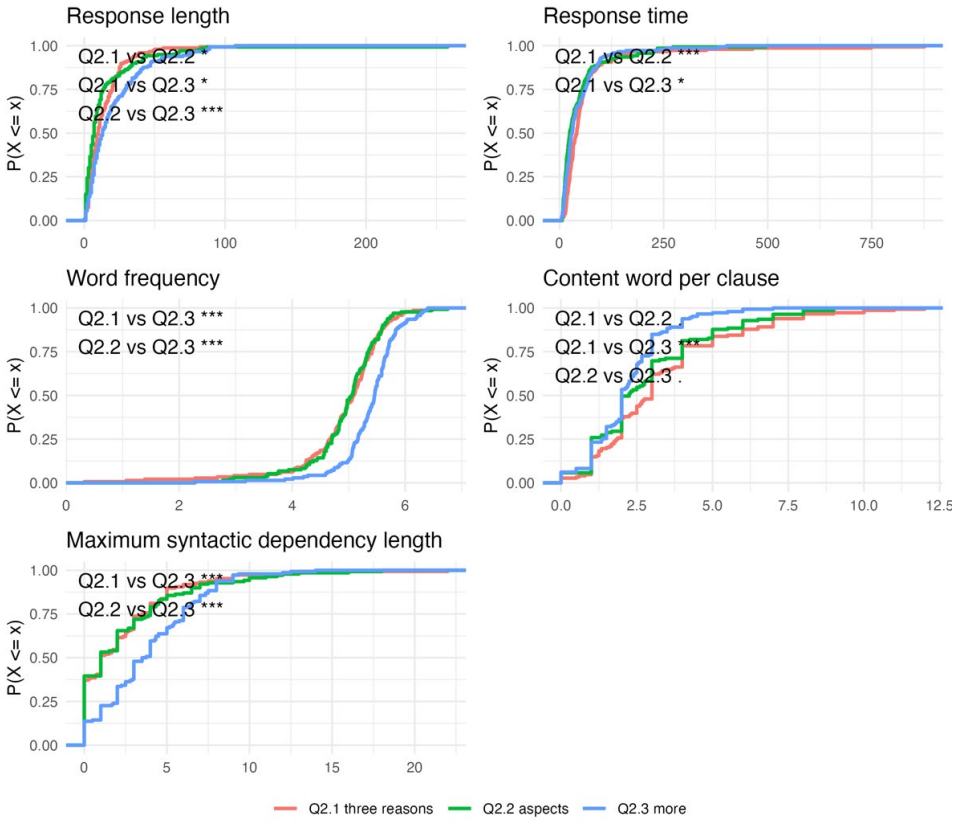

Figure 4 CDF graphs for the response length, time and 3 linguistic complexity features of the answers by Q2 versions.

Table 4 Regression on linguistic complexity features of responses to Q2 on question and respondent characteristic

Variable	Model 1 Response length	Model 2 Response time1	Model 3 Word frequency	Model 4 Content words	Model 5 Maximum SDL
Q2.2: aspects (ref: Q2.1)	−0.210. (0.11)	−0.317** (0.10)	0.016 (0.08)	−0.437. (0.23)	0.495 (0.47)
Q2.3: more (ref: Q2.1)	0.392*** (0.11)	−0.195* (0.10)	0.473*** (0.07)	−1.155*** (0.23)	0.924* (0.43)
Age	−0.015* (0.01)	−0.000 (0.01)	0.001 (0.00)	−0.003 (0.01)	−0.054* (0.03)
Education: low (ref: high)	−0.330. (0.18)	−0.180 (0.16)	0.216. (0.12)	−0.505 (0.38)	−2.282** (0.77)
Education: middle (ref: high)	−0.259** (0.09)	−0.126 (0.09)	0.052 (0.07)	−0.368. (0.20)	−0.663. (0.39)
Mobile platform (ref: desktop)	−0.096 (0.09)	0.118 (0.08)	−0.045 (0.06)	−0.168 (0.19)	0.123 (0.36)
Intercept	3.079*** (5.32)	3.823*** (0.21)	4.924*** (0.16)	3.817*** (0.49)	5.767*** (0.93)
AIC	1036.941	963.0787	744.0103	1635.095	1504.672
R^2	0.093	0.021	0.107	0.059	0.041
R^2 for respondent -level model	0.026	0.000	−0.004	0.002	0.032
R^2 for question -level model	0.066	0.019	0.107	0.056	0.010

Note. ***p < .001, **p < .01, *p < .05, .p < .1
Standard errors in brackets.
Response length and time were log-transformed.

Discussion

In this study, we demonstrated the application of linguistic complexity features in measuring response quality of online OEQs. While previous studies either relied on automatic indicators that focus only on one aspect of response quality (e.g., response length) or human coding (e.g., the number of themes), with the help of NLP tools, we provided a new outlook for analyzing textual responses from more perspectives, with minimal extra expenses. With these new mea-

sures, we also assessed the impact of our questions on response quality and prepared for larger-scale surveys in future studies.

Our results show that using single indicators to measure the response quality of OEQs may be misleading (Schmidt et al., 2020). While some particular questions led to long responses, they did not necessarily have the best quality as measured by other indicators. For example, one variant of our question (Q2.3) that led to the longest answers also suffered from higher non-substantive response rates, contained more frequently used words, and had fewer complex words. Although mostly positive correlations existed between all measures of quality, meaning that on average, longer responses took longer to write, had fewer frequently used words, and had higher complexity at the syntactic level, correlations between most measures were low to moderate in magnitude, suggesting that these measures tap into different aspects of text. Longer sentences certainly did not imply higher lexical sophistication. Response time as a measure of quality does not seem to add much beyond response length.

With proper analysis tools, automatically generated indicators may by themselves be sufficient to understand the response quality of OEQs. While Schmidt et al. (2020) suggested that content-related measures as annotated by humans lead to different insights than automatically generated indicators, we believe that these insights could also be obtained by studying multiple sides of linguistic complexity. As we discussed in Section 4.2, the correlation analysis confirms that while response length often overlaps significantly with the number of content words and syntactic complexity, it does not necessarily imply higher lexical sophistication, which suggests that linguistic complexity features offered us more insight into the quality of responses than comparing response length. Meanwhile, linguistic complexity features are free and less time-consuming to obtain than manual codings.

In our analyses based on response length, time, and linguistic features, we assessed the impact of both respondent-level and question-level features on various indicators covering different sides of the concept of linguistics complexity. On the level of questions, for the first OEQ in our survey, we found that Q1.1 ("Can you tell us more about what makes you (un)certain about whether or not to have children?"), the version without the adaptive statement, obtained higher-quality responses in all measures. With the adaptive statement added, we found the response lengths were shorter, with less sophisticated words and fewer content words per clause. Therefore, from a purely quantitative point of view, we conclude that the original question in Q1 would be the preferred version for receiving better answers. This is surprising because both Holland and Christian (2009) and Oudejans and Christian (2010) chose to present the previous answer when asking follow-up questions. Our first OEQ can also be considered as a follow-up to the closed-ended question; however, when respondents' choices were

mentioned, the response quality actually became lower, which differs from previous literature.

For the follow-up question, it is more difficult to argue which variant produces responses of the highest quality: while response length suggests that we would get the longest responses by simply asking for "more" (Q2.3), our newly introduced measures demonstrated conflicting results: when inspecting word frequency, both Q2.1 (three reasons) and Q2.2 (aspects) have more sophisticated vocabularies; in terms of number of content words per clause, we obtained the most content words when asking for three reasons (Q2.1); while answers to Q2.3 have longer maximum SDL, i.e., are syntactically more complex.

The expectation that Q2.3 might evoke relatively poor responses was corroborated by our analyses based on non-substantive response rates. While Q2.3 ("more") received the longest responses, they do not necessarily have the best quality. As our indicators demonstrate, this version of the question suffers from a higher non-substantive response rate (e.g., people answering various forms of "please refer to the last question", which is longer than typical non-substantive responses but does not provide additional information). Meanwhile, although Q2.1 ("three reasons") received shorter responses on average than Q2.3, they are still longer than Q2.2 ("aspects"). The lower non-substantive response rate, usage of more complex words, and longer response times all indicate that respondents put more effort into answering Q2.1 than the other versions.

At the level of respondents, we corroborate the finding that younger and more educated respondents offered answers of better quality. However, we did not find any difference in response quality between respondents using desktop and mobile devices. In previous research, some scholars reported shorter answers among mobile users (Mavletova, 2013; Peterson et al., 2013), while others found no significant differences (Bosnjak et al., 2013; Buskirk and Andrus, 2014). Since this is also an unsettled question under dispute, perhaps experiments where respondents are assigned either a mobile device or a computer can shed further light on how response quality is impacted by the devices they use.

By comparing the magnitude of effects from respondent characteristics and the different variants of questions, we did not find consistent results: respondent characteristics were more important for answers to questions 1 but less important for answers to Q2. This could be attributed to how much the wording of each version differs from each other in each question. Nevertheless, neither respondent characteristics nor the formulation of the question contributed substantially to variation observed, and our models could explain no more than 10% of the variation.

With the results above, we are able to formulate recommendations for future larger-scale surveys on fertility intentions that want to make use of OEQs, such as our case study exemplifies. For the first questions, almost all our indicators suggest that adding an adaptive statement would harm instead of improve

response quality, and we recommend the original format (Q1.1). For the follow-up, despite relative shorter response lengths, respondents to the "three reasons" version (Q2.1) supposedly put more effort in answering the question, and this version would also be recommended as the follow-up question to the first OEQ in the future survey on uncertainty about having children.

It should be noted that our research so far only focused on the quantitative quality of responses, without scrutinizing the content. While no difference in response complexity was found in the two versions of the first OEQ, it is still possible that the different wordings would cause a shift in focus on topic and timeframe in life. In the next stage of this research project, we will use Topic Modeling (Blei et al., 2003) to extract themes from the text and also compare them to the close-ended questions, and a more thorough comparison between the different versions of the questions can be made.

Overall, we believe this paper could be a first step of incorporating NLP techniques into survey data analysis and hope that future studies could be inspired to improve on our methods. However, we also realize that there are several limitations of this study, and they should be considered by future scholars. First, while T-scan offers more than 400 linguistic features, we selected our features based on findings from previous research focusing on the genre of educational texts (Kleijn, 2018). Since our genre of texts, answers to open survey questions, differs in many aspects from educational texts, future analysis may involve selecting and experimenting with various other features. Secondly, while we attempted to distinguish effects from respondent and question level, we did not analyze the interaction of survey and respondent features. With responses seen as joint production of respondent and interviewer (Barth & Schmitz, 2021), multi-level analysis may be considered in future studies as well. With these initial results and experience, we are looking forward to furthering the application of linguistic features in survey design.

Data and code availability statement

The dataset collected for this study is available to download at dataarchive.lissdata.nl. Full questionnaire, linguistic features generated by T-scan, code used to link datasets and conduct analyses in this study can be accessed through DataverseNL at https://doi.org/10.34894/XXXSXJ.

Conflict of interest statement

The authors declare no conflicts of interest.

Funding

The authors received no specific funding for this work.

Ethics statement

Ethical permission for the study was obtained from the ethical committee of sociology at the University of Groningen (ECS-201123).

References

Agadjanian, V. (2005). Fraught with ambivalence: Reproductive intentions and contraceptive choices in a sub-saharan fertility transition. *Population Research and Policy Review, 24*(6), 617–645. https://doi.org/10.1007/s11113-005-5096-8

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning, 67*(S1), 180–208. https://doi.org/10.1111/lang.12232

Barth, A., & Schmitz, A. (2021). Interviewers' and respondents' joint production of response quality in openended questions. a multilevel negative binomial regression approach. *Methods, Data, Analyses, 15*(1), 34. https://doi.org/10.12758/mda.2020.08

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly, 60*(3), 390–399. https://doi.org/10.1086/297760

Bhrolcháin, M. N., & Beaujouan, É. (2011). Uncertainty in fertility intentions in Britain, 1979-2007. *Vienna Yearbook of Population Research*, 99–129. https://doi.org/10.1553/populationyearbook2011s99

Bhrolcháin, M. N., & Beaujouan, É. (2019). Do people have reproductive goals? constructive preferences and the discovery of desired family size. *Analytical Family* Demography (pp. 27–56). Springer. https://doi.org/10.1007/978-3-319-93227-9_3

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993-1022.

Bosnjak, M., Poggio, T., Becker, K. R., Funke, F., Wachenfeld, A., & Fischer, B. (2013, May). Online survey participation via mobile devices. In *The American Association for Public Opinion Research (AAPOR) 68th Annual Conference*.

Buijs, V. L., & Stulp, G. (2022). Friends, family, and family friends: Predicting friendships of Dutch women. *Social Networks, 70*, 25–35. https://doi.org/10.1016/j.socnet.2021.10.008

Buskirk, T. D., & Andrus, C. H. (2014). Making mobile browser surveys smarter: Results from a randomized experiment comparing online surveys completed via computer or smartphone. *Field Methods, 26*(4), 322–342. https://doi.org/10.1177/1525822X14526146

Callegaro, M. (2008). Seam effects in longitudinal surveys. *Journal of Official Statistics, 24*(3), 387.

Chen, X., & Meurers, D. (2016). Characterizing text difficulty with word frequencies. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 84–94. https://doi.org/10.18653/v1/W16-0509

Chen, X., & Meurers, D. (2019). Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning, 32*(4), 418–447. https://doi.org/10.1080/09588221.2018.1527358

Chen, Y. (2017). *The effects of question customization on the quality of an open-ended question*. Nebraska Department of Education, Data, Research, Evaluation.

Corsi, D. J., Neuman, M., Finlay, J. E., & Subramanian, S. (2012). Demographic and health surveys: A profile. *International journal of epidemiology, 41*(6), 1602–1613. https://doi.org/10.1093/ije/dys184

Couper, M. P., Conrad, F. G., & Tourangeau, R. (2007). Visual context effects in web surveys. *Public Opinion Quarterly, 71*(4), 623–634. https://doi.org/10.1093/poq/nfm044

Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 271–286. https://doi.org/10.1111/j.1467-985X.2012.01041.x

Couper, M. P., Tourangeau, R., & Kenyon, K. (2004). Picture this! exploring visual effects in web surveys. *Public Opinion Quarterly, 68*(2), 255–266. https://doi.org/10.1093/poq/nfh013

Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research, 11*(3). https://doi.org/10.17239/jowr-2020.11.03.01

Dell'Orletta, F., Montemagni, S., & Venturi, G. (2011). Read–it: Assessing readability of italian texts with a view to text simplification. *Proceedings of the second workshop on speech and language processing for assistive technologies*, 73–83.

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

Fernandez, T., Godwin, A., Doyle, J., Verdin, D., Boone, H., Kirn, A., Benson, L., & Potvin, G. (2016). More comprehensive and inclusive approaches to demographic data collection. https://doi.org/10.18260/p.25751

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221. https://doi.org/10.1037/h0057532

Freitas, R., & Testa, M. R. (2017). *Fertility desires, intentions and behaviour: A comparative analysis of their consistency* (tech. rep.). Vienna Institute of Demography Working Papers. https://doi.org/10.1553/0x003cd010

Gauthier, A. H., Caba¸co, S. L. F., & Emery, T. (2018). Generations and gender survey study profile. *Longitudinal and Life Course Studies, 9*(4), 456– 465. https://doi.org/10.14301/llcs.v9i4.500

Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the effects of removing "too fast" responses and respondents from web surveys. *Public Opinion Quarterly, 79*(2), 471–503. https://doi.org/10.1093/poq/nfu058

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey Methodology* (Vol. 561). John Wiley & Sons.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics, 33*(1), 101–122. https://doi.org/10.1515/jos-2017-0006

He, Z., & Schonlau, M. (2020). Automatic coding of open-ended questions into multiple classes: Whether and how to use double coded data. *Survey Research Methods, 14*(3), 267–287. https://doi.org/10.18148/srm/2020.v14i3.7639

Holland, J. L., & Christian, L. M. (2009). The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. *Social Science Computer Review, 27*(2), 196–212. https://doi.org/10.1177/0894439308327481

Holleman, B. (2000). *The forbid/allow asymmetry: On the cognitive mechanisms underlying wording effects in surveys*. Brill.

Hough, M., & Mayhew, P. (1983). *The British crime survey: First report* (Vol. 76). HM Stationery Office London.

Jäckle, A., & Eckman, S. (2020). Is that still the same? Has that changed? On the accuracy of measuring change with dependent interviewing. *Journal of Survey Statistics and Methodology, 8*(4), 706-725. https://doi.org/10.1093/jssam/smz021

Jensen, K. T. (2009). Indicators of text complexity. *Mees, IM; F. Alves & S. Göpferich (eds.)*, 61–80.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics, 53*, 61–79.

Joshi, A., Mishra, A., Senthamilselvan, N., & Bhattacharyya, P. (2014). Measuring sentiment annotation complexity of text. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 36–41. https://doi.org/10.3115/v1/P14-2007

Kemper, S., LaBarge, E., Ferraro, F. R., Cheung, H., Cheung, H., & Storandt, M. (1993). On the preservation of syntax in Alzheimer's disease: Evidence from written sentences. *Archives of Neurology, 50*(1), 81–86. https://doi.org/10.1001/archneur.1993.00540010075021

Keuleers, E., Brysbaert, M., & New, B. (2010). Subtlex-nl: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Kleijn, S. (2018). *Clozing in on readability: How linguistic features affect and predict text comprehension and on-line processing*. LOT.

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

Liefbroer, A. C. (2009). Changes in family size intentions across young adulthood: A life-course perspective. *European Journal of Population/Revue européenne de Démographie, 25*(4), 363–386. https://doi.org/10.1007/s10680-008-9173-7

Lu, C., Bu, Y., Dong, X., Wang, J., Ding, Y., Larivière, V., Sugimoto, C. R., Paul, L., & Zhang, C. (2019). Analyzing linguistic complexity and scientific impact. *Journal of Informetrics, 13*(3), 817–829. https://doi.org/10.1016/j.joi.2019.07.004

Mavletova, A. (2013). Data quality in pc and mobile web surveys. *Social Science Computer Review, 31*(6), 725–743. https://doi.org/10.1177/0894439313485201

McCarthy, P. M., & Jarvis, S. (2010). Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Morgan, S. P. (1982). Parity-specific fertility intentions and uncertainty: The United States, 1970 to 1976. *Demography, 19*(3), 315–334. https://doi.org/10.2307/2060974

Morgan, S. P., & Rackin, H. (2010). The correspondence between fertility intentions and behavior in the United States. *Population and Development Review, 36*(1), 91–118. https://doi.org/10.1111/j.1728-4457.2010.00319.x

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578. https://doi.org/10.1093/applin/amp044

Oakley, D. (1981). Reflections on the development of measures of childbearing expectations. *Predicting fertility. Demographic Studies of Birth Expectations*, 11–26.

Ochieng, P. A. (2009). An analysis of the strengths and limitation of qualitative and quantitative research paradigms. *Problems of Education in the 21st Century, 13*, 13.

Oudejans, M., & Christian, L. M. (2010). Using interactive features to motivate and probe responses to open-ended questions. *Social and behavioral research and the internet: Advances in applied methods and research strategies*, 304–332.

Pander Maat, H., Kraf, R., van den Bosch, A., Dekker, N., Gompel, M. v., Kleijn, S. d., Sanders, T., & Sloot, K. (2014). T-scan: A new tool for analyzing Dutch text.

Peterson, G., Mechling, J., LaFrance, J., Swinehart, J., & Ham, G. (2013). Solving the unintentional mobile challenge. *CASRO Online Research Conference, March, 6*(8).

Peytchev, A., & Hill, C. A. (2010). Experiments in mobile web survey design: Similarities to other modes and unique considerations. *Social Science Computer Review, 28*(3), 319–335. https://doi.org/10.1177/0894439309353037

Pilán, I., Vajjala, S., & Volodina, E. (2016). A readable read: Automatic assessment of language learning materials based on linguistic complexity. *arXiv preprint arXiv:1603.08868*.

Queirós, A., Faria, D., & Almeida, F. (2017). Strengths and limitations of qualitative and quantitative research methods. *European Journal of Education Studies*. https://doi.org/10.5281/zenodo.887089

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. *Developments in Applied Statistics, 19*(1), 159–177.

Revilla, M., & Ochoa, C. (2015). What are the links in a web survey among response time, quality, and auto-evaluation of the efforts done? *Social Science Computer Review, 33*(1), 97–114. https://doi.org/10.1177/0894439314531214

Revilla, M., & Ochoa, C. (2016). Open narrative questions in pc and smartphones: Is the device playing a role? *Quality & Quantity, 50*(6), 2495– 2513. https://doi.org/10.1007/s11135-015-0273-2

Roberts, C., Gilbert, E., Allum, N., & Eisner, L. (2019). Research synthesis: Satisficing in surveys: A systematic review of the literature. *Public Opinion Quarterly*, 83(3), 598-626. https://doi.org/10.1093/poq/nfz035

Rudnicka, K. (2018). Variation of sentence length across time and genre. *Diachronic corpora, genre, and language change*, 220–240. https://doi.org/10.1075/scl.85.10rud

Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., & Drndarevic, B. (2015). Making it simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing (TACCESS), 6*(4), 1–36. https://doi.org/10.1145/2738046

Sand Aronsson, F., Kuhlmann, M., Jelic, V., & Östberg, P. (2021). Is cognitive impairment associated with reduced syntactic complexity in writing? evidence from automated text analysis. *Aphasiology, 35*(7), 900–913. https://doi.org/10.1080/02687038.2020.1742282

Schatz, E., & Williams, J. (2012). Measuring gender and reproductive health in Africa using demographic and health surveys: The need for mixed-methods research. *Culture, Health & Sexuality, 14*(7), 811–826. https://doi.org/10.1080/13691058.2012.698309

Schmidt, K., Gummer, T., & Roßmann, J. (2020). Effects of respondent and survey characteristics on the response quality of an open-ended attitude question in web surveys. *Methods, Data, Analyses, 14*(1), 32. https://doi.org/10.12758/mda.2019.05

Scholz, E., & Zuell, C. (2012). Item non-response in open-ended questions: Who does not answer on the meaning of left and right? *Social Science Research*, *41*(6), 1415–1428. https://doi.org/10.1016/j.ssresearch.2012.07.006

Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods*, *10*(2), 143–152. https://doi.org/10.18148/srm/2016.v10i2.6213

Schuman, H., & Presser, S. (1979). The open and closed question. *American Sociological Review*, 692–712.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Smolík, F., Stepankova, H., Vyhnálek, M., Nikolai, T., Horáková, K., & Matějka, Š. (2016). Propositional density in spoken and written language of Czech-speaking patients with mild cognitive impairment. Journal of Speech, Language, and Hearing Research, 59(6), 1461-1470. https://doi.org/10.1044/2016_JSLHR-L-15-0301

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

Speizer, I. S., Irani, L., Barden-O'Fallon, J., & Levy, J. (2009). Inconsistent fertility motivations and contraceptive use behaviors among women in Honduras. *Reproductive Health*, *6*(1), 1–10. https://doi.org/10.1186/1742-4755-6-19

Staveteig, S., Aryeetey, R., Anie-Ansah, M., Ahiadeke, C., & Ortiz, L. (2017). Design and methodology of a mixed methods follow-up study to the 2014 Ghana demographic and health survey. *Global Health Action*, *10*(1), 1274072. https://doi.org/10.1080/16549716.2017.1274072

Struminskaya, B., Weyandt, K., & Bosnjak, M. (2015). The effects of questionnaire completion using mobile devices on data quality. evidence from a probability-based general population panel. *Methods, Data, Analyses*, *9*(2), 32. https://doi.org/10.12758/mda.2015.014

Stulp, G. (2021). Collecting large personal networks in a representative sample of Dutch women. *Social Networks*, *64*, 63–71. https://doi.org/10.1016/j.socnet.2020.07.012

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*(3), 368–393. https://doi.org/10.1093/poq/nfh035

Tsantali, E., Economidis, D., & Tsolaki, M. (2013). Could language deficits really differentiate mild cognitive impairment (mci) from mild Alzheimer's disease? *Archives of Gerontology and Geriatrics*, *57*(3), 263–270. https://doi.org/10.1016/j.archger.2013.03.011

United Nations. (2005). *Generations & gender programme: Survey instruments*. New York: United Nations.

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the seventh workshop on building educational applications using NLP*, 163–173.

Vanmassenhove, E., Shterionov, D., & Gwilliam, M. (2021). Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. *arXiv preprint arXiv:2102.00287*.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *22*(1), 51–68. https://doi.org/10.1002/acp.1331

Zaller, J., & Feldman, S. (1992). A simple theory of the survey response: Answering questions versus revealing preferences. *American journal of political science*, 579–616. https://doi.org/10.2307/2111583

Zhou, R., Wang, X., Zhang, L., & Guo, H. (2017). Who tends to answer open-ended questions in an e-service survey? the contribution of closed-ended answers. *Behaviour & Information Technology, 36*(12), 1274–1284. https://doi.org/10.1080/0144929X.2017.1381165

Zuell, C., Menold, N., & Körber, S. (2015). The influence of the answer box size on item nonresponse to open-ended questions in a web survey. *Social Science Computer Review, 33*(1), 115–122. https://doi.org/10.1177/0894439314528091

Zuell, C. (2016). Open-ended questions. *GESIS Survey Guidelines, 3*. https://doi.org/10.15465/gesis-sg_en_002

Appendix A

Distribution of Transformed Variables

As noted in Section 3.4, response length and response time were right-skewed. Logarithmic transformation was applied to normalize these distributions for linear regression analysis.

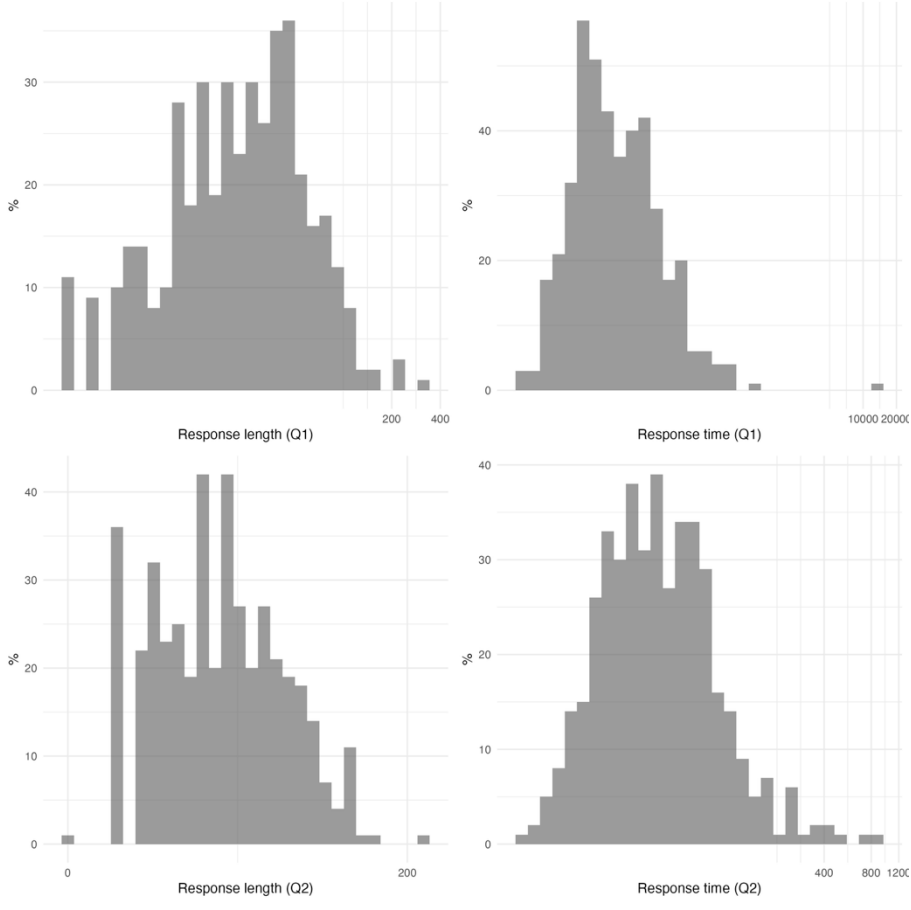

Figure A1 Histograms of log-transformed response length and response time

Appendix B

Robustness Checks (with Negative Binomial Regression)

To verify the robustness of our linear regression results on count data (specifically response length, maximum SDL and number of content words), we performed negative binomial regression models. The result patterns were consistent with the linear models presented in the main text.

Table B1 Negative binomial regression on counted complexity features of responses to Q1 on question and respondent characteristics

Variable	Model 1 Response length	Model 4 Content words	Model 5 Maximum SDL
Q1.2: adaptive (ref: Q1.1)	-0.205* (0.08)	-0.126* (0.06)	0.060 (0.34)
Age	-0.029*** (0.01)	-0.02*** (0.01)	-0.030*** (0.01)
Education: low (ref: high)	-0.583*** (0.16)	-0.146 (0.12)	-0.243. (0.13)
Education: middle (ref: high)	-0.228* (0.09)	0.048 (0.06)	-0.148* (0.07)
Mobile platform (ref: desktop)	-0.025 (0.09)	-0.041 (0.06)	0.009 (0.06)
Intercept	4.65*** (0.22)	1.578*** (0.15)	2.602*** (0.16)
AIC	3791.5	1424.7	2136.1

Table B2 Negative binomial regression on counted complexity features of responses to Q2 on question and respondent characteristics

Variable	Model 1 Response length	Model 4 Content words	Model 5 Maximum SDL
Q2.2: aspects (ref: Q2.1)	-0.015 (0.11)	-0.140. (0.07)	0.037 (0.13)
Q2.3: more (ref: Q2.1)	0.390*** (0.11)	-0.411*** (0.08)	0.559*** (0.13)
Age	-0.012. (0.01)	-0.02 (0.00)	-0.020* (0.08)
Education: low (ref: high)	-0.610** (0.19)	-0.188 (0.13)	-0.774*** (0.23)
Education: middle (ref: high)	-0.332*** (0.10)	0.143* (0.07)	-0.315** (0.11)
Mobile platform (ref: desktop)	-0.054 (0.09)	-0.048 (0.06)	0.029 (0.11)
Intercept	3.30*** (0.24)	1.377*** (0.16)	1.708*** (0.27)
AIC	3102.4	1546.2	1806

Appendix C

Independence of Question Versions

To ensure that the version of the first question (Q1) did not influence the quality of responses to the second question (Q2), we tested for carry-over effects. As shown in Table 3, there was no significant impact in any response quality measure to Q2 based on the version of Q1 received.

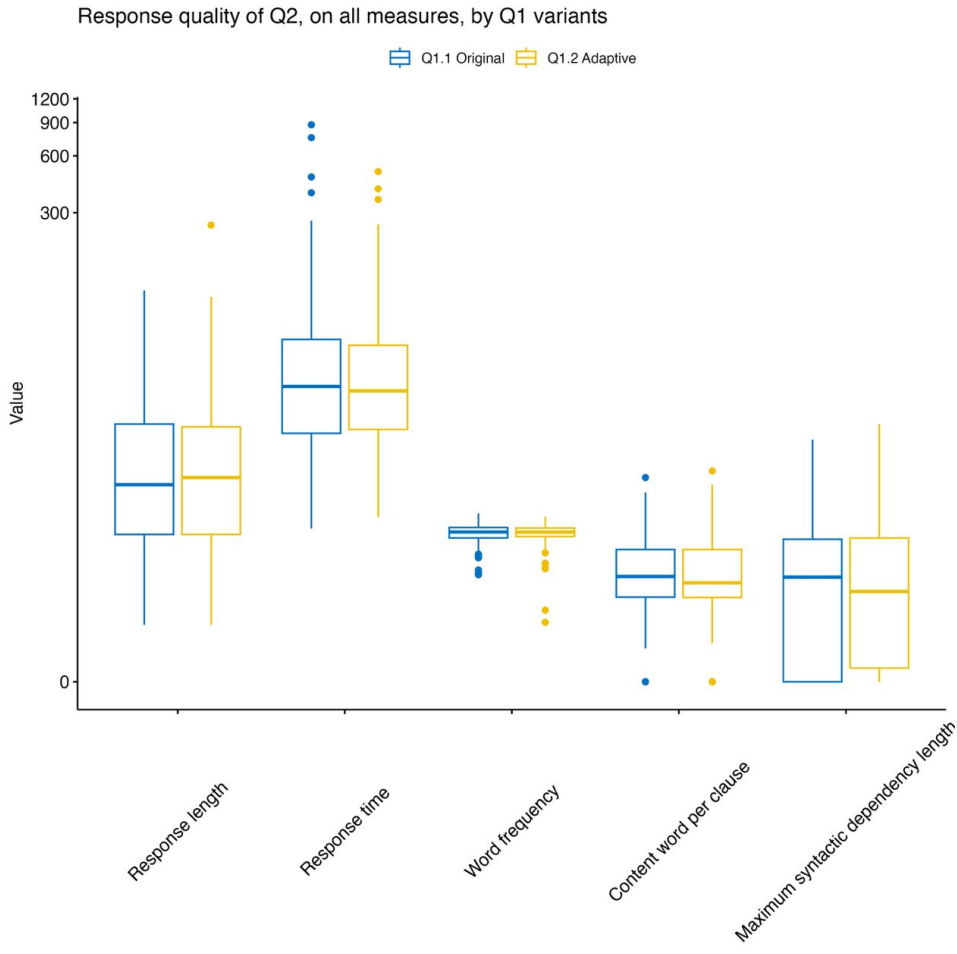

Figure C1 Pairwise comparison of Q2 response quality on all five measures by the version of Q1 received. Pairwise Wilcoxon rank sum test was also conducted, and no significant difference was found.

Online First publication date: 2026-07-10