Methods for the Classification of Data from Open-Ended Questions in Surveys

Inauguraldissertation zur Erlangung des akademischen Grades einer Doktorin der Sozialwissenschaften der Universität Mannheim

Vorgelegt von

Camille Landesvatter

Dekan der Fakultät für Sozialwissenschaften
Prof. Dr. Michael Diehl

Betreuer
Dr. Paul C. Bauer
Prof. Dr. Florian Keusch

Gutachter
Prof. Dr. Florian Keusch
Prof. Dr. Tobias Gummer

Tag der Disputation: 16.04.2024

Contents

Acknowledgements

Funding for this work has been provided by a grant from the German Research Foundation (DFG) for the project “Measuring and Explaining Trust (TRUSTME)” (2020-2024, 449946260), and by the Mannheim Centre for European Social Research (MZES) at the University of Mannheim.

1 Introduction

Individual-level data represent the most basic data unit of sociology, and the modern social sciences can employ many different methods to collect such data. One of the most popular modes is the survey, which can be defined as a tool "to inquire about an audience and collect ideas, opinions, and thoughts" (Mosca et al., 2022, p. 49). Survey research, despite being a young field compared to many scientific domains, has already progressed through three distinct stages of development (Groves, 2011). Especially the first (1930-1960) and second stages (1960-1990) were characterized by a shift from qualitative interviews to quantitative social science practices (e.g., standardized survey formats). However, recent accounts on the benefits and possibilities of open-ended questions alongside developments in computational methods for the automatic classification of text have sparked a renewed interest in open-ended questions.

Open-ended questions (OEQs) are "survey questions that do not include a set of response options" (Züll, 2016, p. 1) and they "require respondents to formulate a response in their own words and to express it verbally or in writing" (Züll, 2016, p. 1). Typically, open-ended questions are descriptive and ask who, what, when, where, and why questions (Popping, 2015).

They are different from "fixed-alternatives" or closed-ended questions (CEQs), where the answer categories are presented in a closed form (Inui et al., 2001, p. 1). Overall, OEQs are theoretically able to provide researchers with detailed, rich and nuanced insights from respondents into subjective meanings, argumentations, descriptions or associations with concepts (Bauer et al., 2017; Heffington et al., 2019; Scholz & Zuell, 2012; M. Singer, 2011). Schuman and Presser (1979) offer one of the early insights into OEQs, and describe that this question format comes with two main benefits: "One is to discover the responses that individuals give spontaneously; the other is to avoid the bias that may result from suggesting responses to individuals" (Schuman & Presser, 1979, p. 692).

The quantitative analysis of data from open-ended questions, which means formatting the unstructured natural language data into numerical formats, requires methods of classification. Classification with regard to data from open-ended survey responses can be defined as coding material "by assigning numbers and/or

categories to text segments" (Rytting et al., 2023, p. 1). This task is challenging and in some cases may require careful and sophisticated coding approaches, most of which have until recently been performed by humans in a manual workflow (Giorgetti & Sebastiani, 2003). Since this manual approach involves high costs, researchers have recently adapted methods from natural language processing (NLP), such as information extraction, automatic summarization and automatic classification (Inui et al., 2001). Early accounts on the idea of using NLP to analyze data from open-ended questions can be found in Lebart et al. (1997), and modern research is still exploring methods to achieve such tasks (Macanovic, 2022).

Nowadays, the use of computational methods to analyze text data from surveys can be located in the eponymous field of Computational Social Science. Many of these methods include state-of-the-art techniques (i.e., supervised and unsupervised classifiers), and some of them also account for the special structure of survey answers which are typically short and concise (Zhu et al., 2022).

Above all, this thesis pursues the objective of introducing various methods of classifying data from open-ended survey questions and empirically illustrating their application. A central research question addressed in this thesis therefore concerns the analysis of (short) text data generated by open-ended survey questions. Each of the three empirical studies included in this cumulative thesis demonstrates applications of methods to classify this type of data, including semi-automated (e.g., supervised machine learning in Study 1 or zero-shot prompting in Study 3) and fully automated approaches (e.g., unsupervised machine learning in Study 2). Each of the three studies pays attention to the short and concise structure of these responses by applying suitable methods (e.g., word embeddings, structural topic models) and where applicable, discusses advantages of such methods (e.g., Study 3 discusses the advantages of word embeddings compared to more static dictionary approaches). The three studies are based on data from open-ended questions collected in three distinct surveys and thereby also outline different ways that researchers can collect open-ended data in surveys. This includes traditional open-ended questions (Study 2) as well as so-called probing questions (Studies 1, 2, 3). Also, the studies examine and introduce different methods to collect the data in terms of the data input mode. Study 2 explicitly compares answers from text and

audio conditions and study 3 focuses on data from audio input modes.
This thesis is structured as follows. The introductory Chapter first contains a general introduction to open-ended questions in surveys and how they developed over the course of the last few decades. Section 1.2 of this introduction provides an overview of methods available for the classification of open-ended text data from surveys and aims at providing an introduction to the methods that are applied in the following three empirical studies. Section 1.3 provides a summary of the three studies. The subsequent Chapters (Chapters 5, 6 and 7) include the three studies, following the order below (abbreviation and study title):

- Study 1: "How valid are trust survey measures? New insights from open-ended probing data and supervised machine learning"
- Study 2: "Open-ended survey questions: A comparison of information content in text and audio response formats"
- Study 3: "Asking Why: Is there an Affective Component of Political Trust Ratings in Surveys?"

Finally, in Chapter 8, this dissertation concludes with a discussion by placing the previously presented results in a larger context and discusses starting points for future research.

1.1 Open-Ended Questions in Surveys

Open-Ended Questions in the Context of Modern Survey Research

According to Groves (2011), modern survey research has developed over the course of three distinct stages. The initial phase, spanning from 1930 to 1960, focused on establishing the fundamental methods and infrastructure of the discipline. During this time, the field undertook a movement from unstandardized to standardized questionnaires, and especially in the '30s and '40s the debate about open versus closed forms of questions flourished (Converse, 1984; Geer, 1991; Groves, 2011; Schuman & Presser, 1979). Notably, in 1935, Lazarsfeld introduced open-ended follow-up questions referred to as "why questions".

The following thirty years, the "Era of Expansion" (Groves, 2011), were characterized by a rapid growth in quantitative social sciences employing standardized survey formats. The drivers of this development were probably ease of use, comparison and analysis (Schuman & Presser, 1979). Despite Schuman's pioneering introduction of "Random Probes" in 1966, open-ended questions, with a few exceptions (for example Bailey, 1994; Schuman and Presser, 1979) received little attention, possibly due to the challenges involved in data analysis and the new excitement about standardized, closed-ended questions.

In recent years, there has been a renewed interest in data from open-ended questions (Neuert, Meitinger, Behr, & Schonlau, 2021).[1] Singer and Couper (2017) outlined a list of objectives particularly well-suited for the use of open-ended questions. These objectives include understanding reasons behind reluctance or refusal, testing methodological theories and hypotheses, encouraging truthful answers, providing an opportunity for feedback, and improving response quality.[2]

Open-Ended Probing Questions

One application of open-ended questions is a technique called probing, which originates from cognitive interviewing. As stated above, the first introductions of probes can be dated back to 1935 (Lazarsfeld, 1935) and in 1966 Schuman also used the term "probe" in calling his questions "random probes" (Schuman, 1966). During the second era of survey research, surveys increasingly became quantified, and modern probing techniques provide a tool that allows researchers to delve into the cognitive processes that respondents undergo when answering closed-ended survey questions (cf. Figure 1.1).

Probes are "a particular type of open-ended questions" (Behr et al., 2017, p. 5) that "require narrative answers from the respondents, but always in relation to a foregoing closed-ended question" (Behr et al., 2017, p. 5). The primary goal

[1] Groves (2011) describes a third stage of survey research, spanning from approximately 1990 to around 2011 (as indicated by the date of the publication). For this stage, Groves emphasizes the various challenges the field is facing (e.g., low response rates, lack of telephones and PCs in households, growth of alternative methods of data collection modes). The renewed interest in open-ended questions, as discussed in this thesis, is primarily attributable to the more recent years (post-2011).

[2] For example, problems that commonly accompany closed-ended questions, such as "straightlining" and other forms of "satisficing" can be eliminated through open-ended questions.

of probing is to collect information on how respondents perceive and understand survey questions or single expressions within questions (Willis, 2004). There are also so-called closed-ended probes (Neuert, Meitinger, & Behr, 2021; Scanlon, 2019), however, in open-ended probes the lack of predefined closed response categories might make them an especially powerful tool to gather rich data (Iyengar, 1996). Web probing constitutes the use of probing questions in large-scale web surveys and compared to interviewer-administered methods, it improves data quality by eliminating interviewer effects (Behr et al., 2012, 2017; Meitinger & Kunz, 2022). Sturgis et al. (2019) introduce the concept of regression-based response probing, a method that "combines the strengths of intensive small-sample qualitative approaches with the inferential power of large-scale field trials and experimental manipulations" (Sturgis et al., 2019, p. 575). Overall, probing is a well-established method, enriched by various subforms and probing techniques. Figure 1.1 illustrates some of the different types of probing questions.

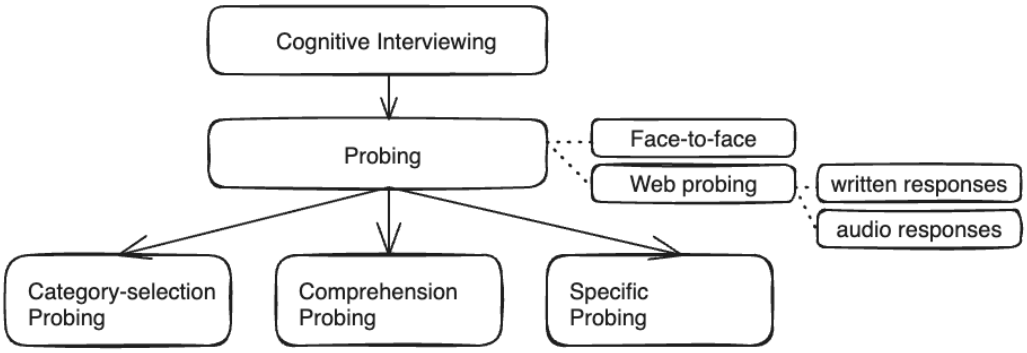

Figure 1.1: Types of Probing in the Context of Cognitive Interviewing.

Category-selection probes ask respondents for their reasons for having chosen a specific response category (Behr et al., 2017). Specific probes ask respondents about a particular detail of a term or another specific aspect that was part of the closed-ended question (Behr et al., 2017). Both category-selection probing as well as specific probing will be applied in the empirical contributions of this thesis (Study 2 and 3 for category-selection, and Study 1 for specific probes).

Current Issues in Research on Open-Ended Questions

Notwithstanding the various use cases and opportunities associated with open-ended questions, many questionnaires include very few or none at all. Neuert et al. (2021) describe how "a general recommendation in survey research is to use open-ended questions sparingly" (Neuert, Meitinger, Behr, & Schonlau, 2021, p. 3). Other, earlier accounts on open-ended responses described that studies often use “excerpts from open-ended responses only to illustrate or underscore quantitative findings” (Mossholder et al., 1995, p. 337) instead of actually analyzing them. Stoneman, Sturgis and Allum (2013) describe that answers to OEQs are often only displayed as “illustrative” quotations alongside a statistical analysis such as regression analysis. Even in examples where OEQs could provide more valuable insights, closed-ended questions are often preferred, sometimes with an extensive list of multiple-choice options and an occasional opportunity for an additional open-ended keyword.

The reasons for this behavior are diverse, but there are two main challenges with open-ended questions: First, compared to closed response formats, they complicate the response process and thus place a higher cognitive burden on the respondent (Tourangeau & Rasinski, 1988; Tourangeau et al., 2000). For example, respondents must formulate their answers in their own words (Keusch, 2014). Second, data from open-ended questions are labor intensive for researchers because in many cases a coding (or categorisation) scheme must be developed to classify qualitative text responses (Züll, 2016).

This thesis aims to address both challenges, i.e., the increased cognitive burden as well as the difficulties associated with analyzing such data. The idea of reducing the response burden of OEQs, for example, is pursued in Study 2, where we asked respondents to use an oral response mode for answering OEQs which might facilitate the data entry process. Concerning the challenge of analyzing such data, this thesis presents and applies different automated methods, including unsupervised and supervised classifiers, throughout the three studies.

1.2 Methods for Analyzing Data from Open-Ended Questions

Social science research objectives with textual data usually have in common the goal of quantitative content analysis. Content analysis is the task of assigning annotation codes to the open responses (Züll, 2016) to convert them into numerical data. For this task, the current landscape of methods in survey research offers a spectrum, which can be broadly distinguished between three approaches: fully manual, semi-automated and fully automated. Table 1.1 depicts these different approaches alongside a selection of available methods. The overview depicted in Table 1.1 distinguishes the three approaches based on whether and to what extent the categories of interest (and their assignments to survey responses) are predefined by the researcher.

	I.) **Fully manual**	**II.)** **Semi-automated**	**III.)** **Fully automated**
Description	Each survey response is assigned a category manually by human operators without automation.	Detection of categories and their mapping to survey responses includes both automated methods and some human involvement.	Detection of categories and their mapping to survey responses is entirely performed without human involvement.
Methods	Quantitative content analysis, no machine involvement for automated categorization	Computational methods, supervised machine learning, fine-tuning of pre-trained models, prompt-based learning	Computational methods, unsupervised machine learning, e.g., word co-occurrence analysis and other clustering methods (e.g., topic models)

Table 1.1: Overview of methods for classifying open-ended survey responses. *Note*: The overview provided in this table is the author's own contribution and suggests a broad typology consisting of three main approaches. For instance, Macanovic (2022) offers a perspective outlining five major method groups within computational text analysis (e.g., with dictionary approaches representing a distinct group of methods). Mosca et al. (2022) also distinguish dichotomously between closed-vocabulary methods and open-vocabulary methods.

While fully manual approaches (I.) and semi-automated approaches (II.) both pursue the goal of assigning manually predefined categories, they differ in two regards. First, they differ with regard to how much of the data is supposed to be manually classified. For I.) this is the full dataset, meaning that each survey response is manually mapped to one of the predefined codes, whereas for II.) only a subset of survey responses is manually assigned a code and the remainder is classified in an automated way. Second, they differ in the way in which this human input is then used for the final classifications. In I.) the manually derived classifications correspond to the final model, but in II.) the manually coded survey responses serve as training or fine-tuning data to create a model that predicts final classifications using machine learning techniques.

Irrespective of the size of the data to be labeled, the manual mapping of codes to documents requires the work of multiple independent "coders". In its smallest variant, this is referred to as double-coding (He & Schonlau, 2020). The involvement of multiple coders aims at minimizing the risk of making systematic and nonsystematic errors throughout the annotation procedure and thus to increase the overall quality of annotations (Kurasaki, 2000). Ultimately, this collaborative approach allows for the computation of an intercoder reliability metric that captures the degree of agreement between coders (Kurasaki, 2000). Manual coding, especially with multiple coders, is expensive and one of its main challenges revolves around maintaining consistency and objectivity during the coding process. Due to such insecurities with regards to data quality as well as the slow and expensive nature of this endeavor, an increasing number of studies nowadays make use of automated approaches.

Automated approaches include semi-automated (II.) and fully automated methods (III.). Semi-automated approaches are particularly useful in scenarios where a classification task demands specialized knowledge, which can be incorporated into the model through labeled examples or a prompt. This data consists of a subset of the original data manually coded with corresponding target labels. The inclusion of such training data for semi-automated classifiers can be achieved in two ways: a model (e.g., decision trees, neural networks, support vector machines etc., for an overview see Sebastiani, 2001) can be fully trained, or an already pre-trained model can be fine-tuned. Full training, i.e. training from scratch, can

involve extensive resource requests and might only be useful for research projects with medium- to large-sized surveys since it requires a sufficiently large set of manually coded documents (Giorgetti & Sebastiani, 2003).[3] In projects with survey data collected from small samples, "hand-coding the training set may coincide with hand-coding the entire set" (Giorgetti & Sebastiani, 2003, p. 1272). A much more efficient alternative is fine-tuning. Fine-tuning is the process of refining a pre-trained machine learning model on a specific task using a smaller, task-specific dataset. For example, Study 1 in this thesis leverages the capabilities of a pre-trained language model (BERT, i.e., Bidirectional Encoder Representations from Transformers) and fine-tunes it for detecting specific codes and sentiment categories in open-ended text responses. In contrast, random forest classifiers that we trained from scratch in Study 1 (Appendix A.6, Study 1) demonstrate lower performance for our sample (however, note that there might be a tradeoff, for example, between accuracy and explainability). Fine-tuning comes with at least two advantages: efficiency and performance. First, pretrained models such as BERT already include a lot of information, hence it takes less time to fine-tune a model compared to a full training. Second, because these pretrained models were trained with large amounts of text, fine-tuning for a specific task can be achieved with smaller datasets. Both training and fine-tuning in semi-automated approaches fall under the broader category of supervised learning.

In many instances, such as when researchers are interested in a specific annotation scheme, fine-tuning is a common procedure. Today, the availability of state-of-the-art technologies such as language models or word embeddings, coupled with platforms like Hugging Face and GitHub that distribute such models, has made "off-the-shelf" fine-tuned models widely accessible.[4] For example, Study 3

[3] Schonlau and Cooper (2016) for instance show that for multinomial boosting, 500 observations are required for training the task of categorizing open-ended survey answers and that additional time savings could be attained by reducing the training data to 200 or 300 observations, but only for less complex problems (e.g., a binary and not multinomial classification problem). In general, automated categorization is shown to result in meaningful time savings as opposed to manual classification as soon as the data to be classified exceeds 1,500 documents (Schonlau & Couper, 2016).

[4] Hugging Face transformers (Wolf et al., 2019) is an open-source library that provides state-of-the-art transformer based language models, facilitating the implementation and experimentation of various NLP models. Before the creation of Hugging Face, researchers often had to depend on paid APIs from companies or use outdated or unavailable models.

in this thesis applies an "off-the-shelf" BERT model for sentiment classification called "pysentimiento" (Pérez et al., 2023). Whereas in Study 1, we fine-tuned the BERT model ourselves (for example due to the more complex and unique annotation codes), sentiment is a common task where we were able to successfully rely on readily available fine-tuned models. Nowadays, several pre-trained BERT models are available (Chiorrini et al., 2021).

Table 1.1 includes one more possibility that can be understood as a semi-automated approach: prompt-based learning, which is a type of model that only requires to be provided with a so-called prompt given in natural language to perform a classification or predictions. Prompting is a modern classification approach that gained popularity through the emergence of large language models such as GPT-3. Prompting can be further dissected into zero-shot or few-shot prompting. Zero-shot prompting "requires no training data and minimal programming to implement" (Burnham, 2023, p. 2), however, since it requires a prompt (which often includes the desired outcome categories), we can categorize this approach as semi-automated.

Lastly, in scenarios where predefined codes or data structure knowledge is absent, fully automated approaches achieved with unsupervised machine learning can be highly valuable. Fully automated methods differ from manual as well as semi-automated methods insofar as they do not require providing any categories beforehand. For example, Study 2 in this thesis employs a topic model, an unsupervised method, to detect previously unknown topics in our corpus. Similarly to full training of a supervised learner, for a successful application of unsupervised learning approaches it is crucial to have a large number of observations that result in a sufficient number of word co-occurrences.

Short Text Classification

For a successful classification of open-ended text answers from surveys, it is essential to consider the unique characteristics of survey answers. In many cases, they are likely to be short, concise and low in context.[5] This sets them apart

[5] Often (except for respondents that repeat the question wording in their answer) the context is only included in the survey question. Also, in open-ended survey answers, content is "related to a theme more specific to a certain field than politics, finance or society in newspapers" (Inui et al.,

from other text sources that are frequently used in the social sciences. The social sphere, the main objective of sociology, is rich with unstructured data, including texts, transcripts, and documents. A similar variety of texts, such as speech transcripts, discussions, and news articles can be found in the political sciences. Many of these text resources exhibit greater length compared to survey answers, as well as complete and well-structured sentences and formats. As a result, many standard machine learning methods for text classification (for an overview see Sebastiani, 2001) have been employed for such data structures but not all of them are suitable for short survey answers.

For example, descriptive methods such as the analysis of the most frequently appearing words in a given text, which can offer valuable insights for longer texts (e.g., news articles), may prove less useful in the context of short texts where many tokens (i.e., words) occur very rarely. Also dictionary methods (i.e., methods that involve assigning fixed codes to words and using these codes to classify documents) can result in dissatisfactory results due to their inflexible architecture (Giorgetti & Sebastiani, 2003). For example, some survey answers might be simply too short to provide a meaningful dictionary score or variance. Additionally, if a particular term does not exist in the pre-designed dictionary, the respective word cannot be classified at all (see Appendix Study 3 for an example). Bag-of-words methods (e.g., dictionary approaches) in general suffer from the disadvantage that new words encountered in application texts are treated as a nuisance (Rudkowsky et al., 2018).

Apart from descriptive methods, also certain machine learning approaches are not necessarily suitable for short text data. Many standard machine learning techniques rely on word co-occurrences (for example LDA, one of the most popular topic model algorithms) and these methods can underperform on short text documents due to sparsity in their co-occurrence matrices (Liang et al., 2018; Zhu et al., 2022).

Fortunately, the issue of short text length is not unique to survey data in the realm of social sciences. Social scientists can draw upon previous research in this area to address this challenge (Laureate et al., 2023; Macanovic & Przepiorka, 2022; Pietsch & Lessmann, 2018). Social media data, for instance, is similarly sparse,

2001, p. 2).

and consequently, there are numerous classifiers suitable for such data. For example, Study 3 in this thesis in the Appendix applies a dictionary approach called VADER that is specifically attuned to sentiments expressed in social media (Hutto & Gilbert, 2014). Study 2 in this thesis applies a structural topic model – a topic model algorithm that is especially suited for short text answers (Roberts et al., 2014).

Furthermore, recent developments such as word embeddings have contributed to significant advancements in the field of natural language processing. Word embeddings, which are usually trained on large corpora of text, "represent (or embed) words in a continuous vector space in which words with similar meanings are mapped closer to each other" (Rudkowsky et al., 2018, p. 1). The embeddings contain "general semantic and syntactic information of words" (Liang et al., 2018), and hence they can be leveraged to guide clustering approaches, such as topic modeling "for short text collections as supplementary information for sparse co-occurrence patterns" (Liang et al., 2018, p. 43612).

Word embeddings can be grouped into traditional word embeddings (e.g., Word2Vec or GloVe) as well as contextual word embeddings. While traditional embeddings consist of fixed vector representations for words, contextual ones are part of the Transformer Architecture (Vaswani et al., 2017) and consist of representations of a word dependent on its left and right semantic context in a document, eventually extracting richer information from shorter texts. Studies 1 and 3 of this thesis leverage language models (BERT and GPT) that both include powerful contextual word embeddings.

Previous research has shown that word embeddings can improve results obtained from bag-of-words classifiers (Rudkowsky et al., 2018) and that they can be used together with a variety of approaches, such as topic models (Jipeng et al., 2019; Liang et al., 2018; Qiang et al., 2017, 2020; Yan et al., 2013) or random forests (Bouaziz et al., 2014; Vora et al., 2017). Lastly, word embeddings from models with the Transformer Architecture can be leveraged for prompt-based learning (Chae & Davidson, 2023; Mayer et al., 2023; Rytting et al., 2023; Zhu et al., 2022; Ziems et al., 2023). Again, such procedures represent impressive examples of semi-automated approaches (cf. Table 1.1) because they only require minimal manual input in the form of a prompt or question text to guide the classification.

Summary

In conclusion, a variety of methods for the automated analysis of open-ended survey answers exists. Of course, the final choice of a modeling approach might depend on different circumstances: the size of the available dataset, the structure of the open-ended text answers (e.g., length, amount of context), as well as the available resources. These issues can all be important factors in deciding whether to pursue a fully manual, semi-automated or fully automated approach. For example, in fully manual or semi-automated approaches, the determination of the required number of coders (also with regards to possible learning effects) can be a crucial part of an efficient workflow (Gummer et al., 2019). A further general guideline concerning the choice of a modeling strategy could be to consider the tradeoff between accuracy and explainability (Lundberg & Lee, 2017).

The classification of content is an inherently complex problem and it is hardly surprising that, in the past, this task was predominantly performed by humans (Giorgetti & Sebastiani, 2003). The challenge with this fully manual approach is not necessarily the added effort.[6] The larger challenge lies in the fact that human codings may not always yield better classifications. Human codings have their limitations, as they can be biased (Mosca et al., 2022), lack objectivity (Inui et al., 2001), introduce errors when coders misinterpret answers or annotation codes (Giorgetti & Sebastiani, 2003), or face transparency issues related to unitization and intercoder reliability (Campbell et al., 2013). Automated coding can help in these various challenges as it can achieve higher reliability and an overall higher transparency because these methods are characterized by objectivity and systematicness (Zhang et al., 2022). However, it is important to acknowledge that certain challenges, such as maintaining complete objectivity, may persist in automated workflows, as human decisions are inevitably involved (e.g., preprocessing of textual data), including in automated workflows. Consequently, the social sciences require extensive research to test and evaluate various methods for classifying survey data to ensure their applicability.

[6] On the contrary, in a scenario where a manual classification could be worthwhile in terms of accuracy, social scientists should consider engaging in this effort.

1.3 Three Studies on the Classification of Open-Ended Survey Data

The following three studies are included into this thesis in the following order:

- Study 1: “How valid are trust survey measures? New insights from open-ended probing data and supervised machine learning“
- Study 2: “Open-ended survey questions: A comparison of information content in text and audio response formats”
- Study 3: “Asking Why: Is there an Affective Component of Political Trust Ratings in Surveys?”

In the following sections, I will summarize each of the three studies. Subsequently, I will discuss commonalities of the studies before presenting the full studies in the following three sections.

1.3.1 Summary of Study 1

Theoretical Background

This study examines the survey question wordings traditionally used in empirical trust research. Over the last few decades, various debates have revolved around these, including debates about scale length, the number of required items, or the “equivalence” debate (Bauer & Freitag, 2018) which discusses whether different respondents understand and interpret the concepts that are part of the question (e.g., “trust”, “most people”) in similar ways.

Research Question

The last in this list of debates is the research objective of this study. In particular, the research question concerns which frames of reference, also called associations, respondents have with different traditionally used items of social trust.

Previous Research and State-of-the Art

The number of studies on the associations respondents have with trust questions is limited, which might be attributed to the limited number of possibilities to collect and analyze such data. Previous approaches (Sturgis & Smith, 2010; Uslaner, 2002) have advanced the field by manually classifying open-ended answers of respondents that were asked to describe their thoughts while answering trust questions in surveys. Other, more recent, research was able to successfully apply the deep-learning model BERT (Devlin et al., 2018) to classify data from open-ended questions (Gweon & Schonlau, 2023; Schonlau et al., 2023).

Data and Methodology

A quota-based sample of 1,500 respondents from the United States was recruited to participate in our web survey. We first asked respondents to answer traditional trust questions, and afterwards probed respondents using an open-ended specific probe with the following (exemplary) wording: "In answering the previous question, who came to your mind when you were thinking about 'most people'? Please describe".

The classifications of the associations were achieved using a supervised classification approach for which we first sampled a random subset of text answers (n=1,000/1,500) that we manually labeled (using elaborated coding schemes and multiple coders).

Afterwards, only the remainder of open-ended answers (n=6,500/6,000) were automatically classified with BERT models that we fine-tuned using the manually classified data. The Appendix of this study includes alternative classification results from random forest models.

Results

Open-ended survey answers to probing questions about associations with standard trust items in surveys were successfully classified using fine-tuned BERT models (accuracy: 87% for the known-unknown dimension and 95% for the sentiment dimension). One of our central findings suggests that a notable proportion of

respondents (ranging from 13% to 31%) incorporated thoughts of known individuals in their responses while answering classic trust questions. Put differently, this represents a share of respondents that answered the survey question based on associations that do not resemble the researcher's purpose of using these questions.

1.3.2 Summary of Study 2

Theoretical Background

Open-ended survey questions are a valuable source of data in addition to closed-ended questions but they pose various challenges, for example the increased response burden they impose on respondents. High response burden can result in phenomena such as unsatisfactory survey experience for respondents, survey break-off, and answers of otherwise low response quality (e.g., uninformative answers), all of which detract from the potential of OEQs.

Research Question

A key question that arises in this context is which response format survey researchers should use to collect open-ended answers in order to maximize the number of informative answers. In particular, this study examines the effect of asking respondents to answer questions via voice input compared to text input. Additionally, we investigate whether the two response formats differ in their usefulness for different types of respondents.

Previous Research and State-of-the Art

Spoken in comparison to written answers are assumed to facilitate the answer process in surveys since they evoke an open narration and produce more intuitive and spontaneous answers (Gavras & Höhne, 2022; Gavras et al., 2022). In the context of mobile web surveys, speaking is assumed to require less effort than typing (Revilla et al., 2020).

For example, previous research found that spoken answers are longer than written ones (Gavras & Höhne, 2022; Gavras et al., 2022) as well as more elaborate and detailed (Lütters et al., 2018; Revilla et al., 2020). Gavras et al. (2022), advanced

the field by using NLP methods, and for instance used unsupervised topic models to compare text and audio answers.

Data and Methodology

We use a U.S. quota-based sample (n=1,500) and questions adapted from popular social survey programs. By experimentally varying the response format, we examine which format elicits answers with a higher amount of information. The amount of information was operationalized by utilizing three measures of information content: answer length, the number of topics, and response entropy.
Answer length is a simple descriptive measure that benefits from its easy implementation, yet provides a very insightful and easy to interpret measure. It also provides us with a benchmark that we can unequivocally compare to the findings of other studies. Response entropy is a measure from information theory that can be used to capture the additional or "unexpected" information in a given text. Response entropy represents a more advanced method than response length or other previously used measures (e.g., Type-Token Ratio). Lastly, in line with Gavras et al. (2022), we apply unsupervised topic models (i.e., structural topic models) to receive another measure of the range of information given in the text answers.

Results

The main findings of this Study indicate that, for the majority of our questions, spoken responses tend to be significantly longer, and also slightly more informative than their written counterparts. Moreover, we found that higher-educated respondents exhibit longer answer lengths in the audio condition. The presence of other individuals during survey participation, in our sample, had a negative effect on response length in audio formats.

1.3.3 Summary of Study 3

Theoretical Background

Previous investigations on trust are characterized by an understanding that trust is rooted in informed, rational, and consequential judgments. Recent research

however has outlined that additionally to a "cognitive" route there is an "affective" route to trust judgements (Grimmelikhuijsen, 2012). The role of affect in trust judgments can be investigated with the help of "mood models" (Dunn & Schweitzer, 2005) according to which people attribute their current mood to the judgment they are evaluating.

Research Question

This study investigates the nature of political trust ratings in surveys, with a specific focus on so-called affective rationales, for example emotions. We ask to what extent individual responses to a question about political trust in surveys are driven by affective rationales. Additionally, we investigate whether the presence of affective responses is related to the strength of trust scores.

Previous Research and State-of-the Art

Mossholder et al. (1995) is one of the first accounts in which emotions were measured with a dictionary approach. Other modern approaches tackle the task of classifying sentiment as well as emotions in Twitter data using BERT (Chiorrini et al., 2021). Currently, for investigating the influence of emotions in political trust judgments, we can only rely on evidence from applications to interpersonal trust showing how incidental moods with positive valence increase trust and how moods of negative valence decrease it (Dunn & Schweitzer, 2005; Myers & Tingley, 2017). Furthermore, previous research has indicated that the cognitive nature of an emotion (e.g., other person control in anger compared to sadness) impacts the influence of affect on trust.

Data and Methodology

Empirical evidence was derived from a web survey conducted among a sample of approximately n=1,500 respondents from the United States. We asked one of the standard political trust questions and subsequently collected open-ended data on the response process using response probing (category-selection probe).
The resulting answers are analyzed in terms of sentiment (negative, positive and neutral) and emotions (anger, sadness, happiness, neutrality). Additionally, we

make a distinction between analyzing the original audio files and the transcribed text answers. The sentiment analysis is achieved by two distinct sentiment classifiers that are based on deep learning: a fine-tuned BERT model called pysentimento (Pérez et al., 2023) and zero-shot prompting with GPT-3.5-turbo. The emotion analysis is carried out with an "off-the-shelf", fine-tuned wav2vec model from SpeechBrain (Ravanelli et al., 2021). The Appendix of this study includes various alternatives, such as a comparison of results from standard dictionary approaches (i.e., AFINN and VADER).

Results

We find that asking for trust in the government results in a significant share of non-neutral sentiment which is predominantly negative (59%-62% dependent on classifiers). Furthermore, we found that the valence of these open-ended survey responses (e.g., positive, negative) have a strong influence on a 5-point trust scale. In terms of emotions, we found very small shares of emotional language. The positive emotion of happiness has a significant and positive effect on political trust in our sample.

1.3.4 Commonalities of the Studies

To conclude, each of the three studies yields applications of methods for analyzing and classifying open text answers in surveys using machine learning. The studies vary in their analytical approaches and taken together they provide a selection of automated methods available for text classification tasks (cf. Table 1.1).
The three studies are not solely linked by their use of machine learning methods. In terms of content, they are all located in the research field of generalized, interpersonal, as well as political trust.[7]
Finally, a third commonality of the studies included in this thesis is a focus on researching how to collect open-ended data in the first place. In particular, the

[7]This is because these studies were conducted as part of a DFG-funded project that concerns questions on how to measure trust and how to explain differences in trust. The aim of this project is to develop better and more differentiated measures of trust, which includes the use of novel and innovative techniques in this field. The studies included in this thesis were part of this research project.

studies use traditional open-ended questions as well as response probing questions. The method of response probing is included in each of the three studies, as they all aimed at capturing the thought process of individuals when answering survey questions. Each of the three studies tried to capture cognitive processes that take place when individuals form decisions, answers and judgements in surveys. While research on such cognitive thought processes stem from psychology, the studies presented in this dissertation try to explore their functionality in sociological and survey contexts. Furthermore, the studies contribute to research on how to collect open-ended data in that they vary in whether the answers are collected using a more traditional text entry or a more innovative audio response mode.

References

Bailey, K. D. (1994). *Methods of social research*. Simon; Schuster.

Bauer, P. C., Barberá, P., Ackermann, K., & Venetz, A. (2017). Is the Left-Right scale a valid measure of ideology? *Political Behavior*, *39*(3), 553–583.

Bauer, P. C., & Freitag, M. (2018, March). Measuring trust. In E. M. Uslaner (Ed.), *The oxford handbook of social and political trust* (pp. 1–27). Oxford University Press.

Behr, D., Kaczmirek, L., Bandilla, W., & Braun, M. (2012). Asking probing questions in web surveys: Which factors have an impact on the quality of responses? *Soc. Sci. Comput. Rev.*, *30*(4), 487–498.

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). Web probing – implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions. *Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines)*.

Bouaziz, A., Dartigues-Pallez, C., da Costa Pereira, C., Precioso, F., & Lloret, P. (2014). Short text classification using semantic random forest. *Data Warehousing and Knowledge Discovery*, 288–299.

Burnham, M. (2023). Stance detection with supervised, Zero-Shot, and Few-Shot applications. *arXiv preprint:2305.01723*.

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociol. Methods Res.*, *42*(3), 294–320.

Chae, Y., & Davidson, T. (2023). Large language models for text classification: From Zero-Shot learning to Fine-Tuning. *OSF, osf.io/5t6xz*.

Chiorrini, A., Diamantini, C., Mircoli, A., & Potena, D. (2021). Emotion and sentiment analysis of tweets using BERT. *EDBT/ICDT Workshops*.

Converse, J. M. (1984). Strong arguments and weak evidence: The Open/Closed questioning controversy of the 1940s. *Public Opin. Q.*, *48*(1B), 267–282.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint: 1810.04805*.

Dunn, J. R., & Schweitzer, M. E. (2005). Feeling and believing: The influence of emotion on trust. *J. Pers. Soc. Psychol.*, *88*(5), 736–748.

Gavras, K., & Höhne, J. K. (2022). Evaluating political parties: Criterion validity of open questions with requests for text and voice answers. *Int. J. Soc. Res. Methodol.*, *25*(1), 135–141.

Gavras, K., Höhne, J. K., Blom, A. G., & Schoen, H. (2022). Innovating the collection of open-ended answers: The linguistic and content characteristics of written and oral answers to political attitude questions. *J. R. Stat. Soc. Ser. A Stat. Soc.*, *tba*(tba), 1–19.

Geer, J. G. (1991). DO OPEN-ENDED QUESTIONS MEASURE "SALIENT" ISSUES? *Public Opin. Q.*, *55*(3), 360–370.

Giorgetti, D., & Sebastiani, F. (2003). Automating survey coding by multiclass text categorization techniques. *J. Am. Soc. Inf. Sci. Technol.*, *54*(14), 1269–1277.

Grimmelikhuijsen, S. (2012). Linking transparency, knowledge and citizen trust in government: An experiment. *International Review of Administrative Sciences*, *78*(1), 50–73.

Groves, R. M. (2011). Three eras of survey research. *Public Opin. Q.*, *75*(5), 861–871.

Gummer, T., Blumenberg, M. S., & Roßmann, J. (2019). Learning effects in coders and their implications for managing content analyses. *Int. J. Soc. Res. Methodol.*, *22*(2), 139–152.

Gweon, H., & Schonlau, M. (2023). Automated classification for open-ended questions with bert. *Journal of Survey Statistics and Methodology*.

He, Z., & Schonlau, M. (2020). Automatic coding of text answers to Open-Ended questions: Should you double code the training data? *Soc. Sci. Comput. Rev.*, *38*(6), 754–765.

Heffington, C., Park, B. B., & Williams, L. K. (2019). The "most important problem" dataset (MIPD): A new dataset on american issue importance. *Conflict Management and Peace Science*, *36*(3), 312–335.

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious Rule-Based model for sentiment analysis of social media text. *ICWSM*, *8*(1), 216–225.

Inui, H., Murata, M., Uchimoto, K., & Isahara, H. (2001). Classification of Open-Ended questionnaires based on surface information in sentence structure. *NLPRS*, 315–322.

Iyengar, S. (1996). Framing responsibility for political issues. *Ann. Am. Acad. Pol. Soc. Sci.*, *546*, 59–70.

Jipeng, Q., Zhenyu, Q., Yun, L., Yunhao, Y., & Xindong, W. (2019). Short text topic modeling techniques, applications, and performance: A survey. *arXiv preprint:1904.07695*.

Keusch, F. (2014). The influence of answer box format on response behavior on List-Style Open-Ended questions. *J Surv Stat Methodol*, *2*(3), 305–322.

Kurasaki, K. S. (2000). Intercoder reliability for validating conclusions drawn from Open-Ended interview data. *Field methods*, *12*(3), 179–194.

Laureate, C. D. P., Buntine, W., & Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis. *Artif Intell Rev*, 1–33.

Lazarsfeld, P. F. (1935). The art of asking WHY in marketing research: Three principles underlying the formulation of questionnaires. *National Marketing Review*, *1*(1), 26–38.

Lebart, L., Salem, A., & Berry, L. (1997, December). *Exploring textual data*. Springer Science & Business Media.

Liang, W., Feng, R., Liu, X., Li, Y., & Zhang, X. (2018). GLTM: A global and local word Embedding-Based topic model for short texts. *IEEE Access*, *6*, 43612–43621.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint: 1705.07874*.

Lütters, H., Friedrich-Freksa, M., & Egger, M. (2018). Effects of speech assistance in online questionnaires. *General Online Research Conference, Vol. 18*.

Macanovic, A. (2022). Text mining for social science - the state and the future of computational text analysis in sociology. *Soc. Sci. Res.*, *108*, 102784.

Macanovic, A., & Przepiorka, W. (2022). A systematic evaluation of text mining methods for short texts: Mapping individuals' internal states from feedback texts and tweets. *OSF, osf.io/m6n9t*.

Mayer, C. W. F., Ludwig, S., & Brandt, S. (2023). Prompt text classifications with transformer models! an exemplary introduction to prompt-based learning with large language models. *Journal of Research on Technology in Education*, *55*(1), 125–141.

Meitinger, K., & Kunz, T. (2022). Visual design and cognition in List-Style Open-Ended questions in web probing. *Sociol. Methods Res.*, 00491241221077241.

Mosca, E., Harmann, K., Eder, T., & Groh, G. (2022). Explaining neural NLP models for the joint analysis of Open-and-Closed-Ended survey answers. In A. Verma, Y. Pruksachatkun, K.-W. Chang, A. Galstyan, J. Dhamala, & Y. T. Cao (Eds.), *Proceedings of the 2nd workshop on trustworthy natural language processing (TrustNLP 2022)* (pp. 49–63). Association for Computational Linguistics.

Mossholder, K. W., Settoon, R. P., Harris, S. G., & Armenakis, A. A. (1995). Measuring emotion in open-ended survey responses: An application of textual data analysis. *J. Manage.*, *21*(2), 335–355.

Myers, C. D., & Tingley, D. (2017). The influence of emotion on trust. *Polit. Anal.*, *24*(4), 492–500.

Neuert, C., Meitinger, K., & Behr, D. (2021). Open-ended versus closed probes: Assessing different formats of web probing. *Sociol. Methods Res.*, *52*(4), 1981–2015.

Neuert, C., Meitinger, K., Behr, D., & Schonlau, M. (2021). The use of open-ended questions in surveys. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, *15*(1), 3–6.

Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alemany, L. A., & Martínez, M. V. (2023). Pysentimiento: A python toolkit for opinion mining and social NLP tasks.

Pietsch, A.-S., & Lessmann, S. (2018). Topic modeling for analyzing open-ended survey responses. *Journal of Business Analytics*, *1*(2), 93–116.

Popping, R. (2015). Analyzing open-ended questions by means of text analysis procedures. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *128*(1), 23–39.

Qiang, J., Chen, P., Wang, T., & Wu, X. (2017). Topic modeling over short texts by incorporating word embeddings. *Pacific-Asia Conference on Knowledge*.

Qiang, J., Qian, Z., Li, Y., Yuan, Y., et al. (2020). Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on*.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., … Bengio, Y. (2021). SpeechBrain: A General-Purpose speech toolkit.

Revilla, M., Couper, M. P., Bosch, O. J., & Asensio, M. (2020). Testing the use of voice input in a smartphone web survey. *Soc. Sci. Comput. Rev.*, *38*(2), 207–224.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *Am. J. Pol. Sci.*, *58*(4), 1064–1082.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Commun. Methods Meas.*, *12*(2-3), 140–157.

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., & Wingate, D. (2023). Towards coding social science datasets with language models. *arXiv preprint: 2306.02177*.

Scanlon, P. J. (2019). The effects of embedding closed-ended cognitive probes in a web survey on survey response. *Field methods*, *31*(4), 328–343.

Scholz, E., & Zuell, C. (2012). Item non-response in open-ended questions: Who does not answer on the meaning of left and right? *Soc. Sci. Res.*, *41*(6), 1415–1428.

Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods*, *10*, 143–152.

Schonlau, M., Weiß, J., & Marquardt, J. (2023). Multi-label classification of open-ended questions with BERT. *arXiv preprint: 2304.02945*.

Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. *Am. Sociol. Rev.*, *31*(2), 218–222.

Schuman, H., & Presser, S. (1979). The open and closed question. *Am. Sociol. Rev.*, *44*(5), 692–712.

Sebastiani, F. (2001). Machine learning in automated text categorization. *arXiv preprint:cs/0110053*.

Singer, E., & Couper, M. P. (2017). Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys. *methods, data, analyses*, *11*(2), 20.

Singer, M. (2011). Who says "it's the economy"? Cross-National and Cross-Individual variation in the salience of economic performance. *Comp. Polit. Stud.*, *44*(3), 284–312.

Stoneman, P., Sturgis, P., & Allum, N. (2013). Exploring public discourses about emerging technologies through statistical clustering of open-ended survey questions. *Public Underst. Sci.*, *22*(7), 850–868.

Sturgis, P., Brunton-Smith, I., & Jackson, J. (2019). Regression-Based response probing for assessing the validity of survey questions. In Paul Beatty, Debbie Collins, Lyn Kaye, Jose Luis Padilla, Gordon Willis, Amanda Wilmot (Ed.), *Advances in questionnaire design, development, evaluation and testing* (pp. 573–591). unknown.

Sturgis, P., & Smith, P. (2010). Assessing the validity of generalized trust questions: What kind of trust are we measuring? *Int J Public Opin Res*, *22*(1), 74–92.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychol. Bull.*, *103*(3), 299–314.

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

Uslaner, E. M. (2002). *The moral foundations of trust*. Cambridge University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint: 1706.03762*.

Vora, P., Khara, M., & Kelkar, K. (2017). Classification of tweets based on emotions using word embedding and random forest classifiers. *Int. J. Comput. Appl. Technol.*, *178*(3), 1–7.

Willis, G. B. (2004, September). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE Publications.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. M. (2019). HuggingFace's transformers: State-of-the-art natural language processing.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *Proceedings of the 22nd international conference on World Wide Web - WWW '13*.

Zhang, R., Gong, J., Ma, S., Firdaus, A., & Xu, J. (2022). Automatic coding mechanisms for Open-Ended questions in journalism surveys: An application guide. *Digital Journalism*, 1–22.

Zhu, Y., Zhou, X., Qiang, J., Li, Y., Yuan, Y., & Wu, X. (2022). Prompt-Learning for short text classification. *arXiv preprint: 2202.11345*.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can large language models transform computational social science? *arXiv preprint: 2305.03514*.

Züll, C. (2016). *Open-Ended questions (version 2.0)*, GESIS - Leibniz-Institut für Sozialwissenschaften.

2 How valid are trust survey measures? New insights from open-ended probing data and supervised machine learning

Abstract

Trust is a foundational concept of contemporary sociological theory. Still, empirical research on trust relies on a relatively small set of measures. These are increasingly debated, potentially undermining large swathes of empirical evidence. Drawing on a combination of open-ended probing data, supervised machine learning, and a U.S. representative quota sample, our study compares the validity of standard measures of generalized social trust with more recent, situation-specific measures of trust. We find that survey measures that refer to 'strangers' in their question wording best reflect the concept of generalized trust, also known as trust in unknown others. While situation-specific measures should have the desirable property of further reducing variation in associations, i.e., producing more similar frames of reference across respondents, they also seem to increase associations with known others, which is undesirable. In addition, we explore to what extent trust survey questions may evoke negative associations. We find that there is indeed variation across measures, which calls for more research.

Keywords

social trust, generalized trust, survey experiment, open-ended survey questions, text analysis, sentiment analysis, BERT

2.1 Introduction

Generalized social trust is one of the fundamental concepts in contemporary social theory (Coleman, 1994; Herreros, 2004; Putnam et al., 1994; Schilke et al., 2021; Smith, 2010; Sztompka, 1999; Uslaner, 2002) and scholarly interest in this concept has grown alongside the increasing number of studies on social capital and social cohesion, as trust is considered a main indicator of these concepts (Larsen, 2013; Portes & Vickstrom, 2011; Van Deth, 2003). Consequently, empirical research investigating the causes and consequences of trust has multiplied (Buskens & Weesie, 2000; Cook & Cooper, 2003; Dinesen, 2012; Dinesen & Sønderskov, 2015; Dinesen et al., 2014; Sønderskov, 2011). At the same time, the underlying empirical research program relies on a relatively small set of established survey measures, some of which date back to the 1940s. In recent years, we have seen a growing debate about the validity of these measures, particularly regarding their ability to capture the same concept across all individuals (Bauer & Freitag, 2018; Delhey & Newton, 2005; Delhey et al., 2011; Ermisch & Gambetta, 2010; Nannestad, 2008; Robbins, 2022; Sturgis & Smith, 2010; Torpe & Lolle, 2011).
Our study aims to address this debate by investigating the validity of survey measures of generalized social trust. In doing so, we make several contributions to current research.
First, we evaluate three classic trust measures in a U.S. sample, thus extending previous work that examined fewer measures using data from the UK (Sturgis & Smith, 2010; Sturgis et al., 2019). All three measures have been used to measure generalized social trust, specifically trust in unknown others (Sønderskov, 2011; Uslaner, 2002). The first measure is known as the "most people question" (Rosenberg, 1956), which poses the query "Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?". The second measure, referred to as the "people first time question" (e.g., Torpe & Lolle, 2011), asks respondents about their level of trust in people they meet for the first time. Both of these measures have been established and utilized in numerous large-scale surveys. In contrast, what we call the "stranger question" (Robbins, 2021, 2022), which is "Imagine meeting a total stranger for the first time. Please identify how much you would trust this stranger.", is a more recent alternative

and hopeful contender, expected to alleviate some of the problems that appear to characterize the former two. Our study revolves around exploring the validity of these three measures and scrutinizing whether they genuinely measure trust in unknown others, thus identifying possible measurement errors that might influence estimates of trust levels. To achieve this, we designed a survey experiment in which the different measures were randomly assigned to respondents. Our main findings are derived from using open-ended questions that ask about respondents' frames of reference, what we call associations, underlying their response.

Second, we contrast classic measures of generalized social trust with situative measures of trust. Such measures differ from the classical ones in that they specify a more refined trustee category (e.g., "most people" is replaced with "stranger") as well as some behavior at which the expectation is directed (e.g., "keeping a secret"). Ideally, such measures are able to provide a higher degree of interpersonal comparability since they leave less room for different interpretations by the survey respondents. We are the first to probe such measures and provide evidence on whether validity and comparability increases when these measures are used.

Third, we explore the sentiment of associations, a dimension that has been neglected so far in trust research. Theory assumes that trust in known others is higher due to effects of in-group bias and reciprocity (Vollan, 2011), which is supported by empirical evidence (e.g., Bauer & Freitag, 2018; Sturgis & Smith, 2010). However, independently of whether respondents refer to known or unknown others, associations may also vary in terms of their sentiment, for example whether they are positive or negative.

Fourth, we extend the methodological toolbox that is used to evaluate the validity of survey measures, using a combination of open-ended probing questions (e.g., Behr et al., 2012, 2017; Meitinger & Kunz, 2022; Neuert et al., 2021) and automated text analysis (e.g., Schonlau & Couper, 2016). The data we labeled and the resulting supervised classifiers we built are suitable for future applications.

2.2 Theory, hypotheses, and previous research

2.2.1 Associations with known and unknown others

Generalized social trust is often referred to as trust in the generalized other and can be described as trust in individuals who are unfamiliar or unknown (Sønderskov, 2011; Stolle, 2015; Sturgis & Smith, 2010; Uslaner, 2002). Stolle (2015) for example emphasizes the need to distinguish the scope of generalized trust from trust toward people one personally knows (Stolle, 2015, p.398). Notably, other accounts have chosen to expand the concept of generalized or social trust trust to encompass a wider range of trustees, such as trust "in people in general" (Yamagishi & Yamagishi, 1994, p.146), or as trust in the “average person [one] meets” (Coleman, 1994, p.104). Our study, however, uses the understanding of generalized trust that stresses the difference between generalized and particularized trust. Particularized trust is defined as "[...] trust found in close social proximity and extended toward people the individual knows from everyday interactions" (Freitag & Traunmüller, 2009, p.784), including family members, friends, neighbours and co-workers (Freitag & Traunmüller, 2009, p.784) (i.e., known others), whereas generalized trust encompasses "[...] those beyond immediate familiarity, including strangers" (Freitag & Traunmüller, 2009, p.784) (i.e., unknown others). In this study, we argue that when conceptualizing generalized trust, it should ideally be measured as trust towards unknown others.

Currently, the measurement of trust primarily relies on survey questions, although behavioral measures and their combination with survey measures have gained popularity (Barr, 2003; Ermisch & Gambetta, 2010; Fehr et al., 2002; Naef & Schupp, 2009). Various different questions are used in different large-scale surveys. Undoubtedly, the standard measure is the so-called "most people question" which inquires whether most people can be trusted. Different versions of this question were used in thousands of influential studies and underlying surveys, such as the General Social Survey, the World Values Survey or the European Social Survey.

However, the measurement of trust using the most people question has been subject of many debates (cf. Bauer & Freitag, 2018) regarding various aspects, such as scale length or balance (Lundmark et al., 2016), and the frames of reference em-

ployed by respondents when answering it (Delhey et al., 2014; Nannestad, 2008; Sturgis & Smith, 2010). These frames of reference, what we call associations, are important as they are linked to the conceptual validity of a measure. Conceptual validity increases when the respective survey questions capture generalized trust without specification or measurement error. Figure 2.1 depicts our main argument regarding these associations.

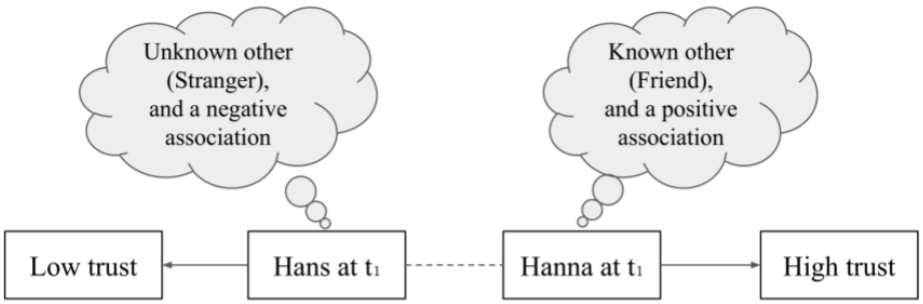

Figure 2.1: Variation in associations and trust measurement values.

When employing trustee categories such as "most people" in standard trust measures, it is probable that distinct associations may arise among different respondents. For instance, in the illustrated example presented in Figure 2.1, respondent Hanna envisions a friend, while Hans envisions a stranger when answering the corresponding survey question. This scenario highlights the ongoing debate on equivalence and whether the concepts in the questions are uniformly interpreted by all respondents (Bauer & Freitag, 2018). Consequently, due to these varying associations, Hanna's response reflects particularized trust, resulting in a specification error, while Hans's response more closely aligns with the notion of the generalized other. These differences in associations can lead to divergent responses on the trust scale between two individuals (e.g., Hans and Hanna) or even within the same individual at different points in time (depicted by the dashed line in Fig-

ure 2.1).

Given that the conceptual definition of generalized (and particularized) trust refers to the distinction between known and unknown others, our study aims to identify the associations arising from the specific wording of survey questions. Empirical evidence in that direction is given by Sturgis and Smith (2010). In examining the most people question using think-aloud probing, they describe 6 higher-order topics they found respondents to associate with the term "most people". The two largest categories they found by manually classifying responses to their probing question were "known others" (42%) and "unknown others" (22%).[8] In a similar approach, Bauer and Freitag (2018) surveys student samples from Switzerland using a probe that asks respondents who they had in mind when answering the most people question. The open-ended text answers reveal that “respondents do not necessarily tend to think of strangers or people that are unknown to them. Many think of situations (e.g., meeting someone in the train/street) or of people they know (e.g., friends, family members, etc.)” (Bauer & Freitag, 2018, p.9). Lastly, Uslaner (2002, p.72-74), as part of the 2000 ANES Pilot Survey, investigated the most people question via think-aloud techniques and showed that 58% of the respondents referred to a “general worldview” while 23% mentioned “personal experiences". While personal experiences do not necessarily involve known others, the 2002 ANES data was also coded into more fine-grained categories by Johnson (cf. ANES, 2000): 8% of respondents referred to family members, 11% to co-workers and 12% to neighbors.

The present study compares three established measures of generalized social trust, the "most people question" (M1), the "people first time question" (M2) and the "stranger question" (M3). Next to M1, M2 is the second most common generalized trust measure used in many large-scale surveys, such as the World Values Survey or the Socio-Economic Panel in Germany. M3 is a more recent measurement approach, which is not yet part of larger surveys, and was developed with the aim that respondents imagine strangers in their answer (Robbins, 2021, 2022). Our particular interest for each of these measures lies in the proportion of

[8] Smaller categories they found refer to "local community" (e.g., people in their town) (3%), "job/profession" (e.g., politicians, salesmen) (4%), "other" (e.g., "trusting is naive") (5%) and "don’t know/no answer" (6%).

respondents who think of personally known others (short: known others), when answering expressed as $p_k = \frac{1}{n}\sum_{i=1}^{n} Y_i$, where Y_i is a dummy that indicates whether individual i thought of known others (1) or unknown others (0) in their response. Importantly, across the three measures M1–M3, the trustee category is gradually refined. M1 is fairly vague and only refers to most people. M2 already specifies that respondents should think of first-time encounters. M3 further specifies the trustee category by clarifying that the trustee category encompasses strangers. We expect that explicitly referring to "people you meet for the first time" (M2) or "a total stranger you meet for the first time" (M3) as compared to "most people" (M1) may increase the proportion of respondents thinking of others they do not know ($1 - p_k$). Furthermore, we expect that using the stranger-wording (M3) should increase this share even more than using the people-wording (M2). In our view, the people-wording is more likely to produce associations of situations where the respondent has had first-time encounters with persons that are well-known by now. For instance, respondents may think of a first-time encounter with friends, work colleagues or relatives or first-time encounters with persons who are already connected (e.g. first time meeting the new partner of a sibling). In contrast, the stranger-wording should make it more likely that respondents think about situations in which they really don't have (or haven't had) any information about the trustee (e.g., encounters in the street). Eventually, we hypothesize that a refinement of the trustee category (most people $\rightarrow$ people you meet for the first time $\rightarrow$ a total stranger you meet for the first time), decreases the proportion of respondents in whom the association with known people (p_k) is evoked (H1). Evidence for H1 would be provided by statistically significant differences between those proportions: $p_{k,M1} > p_{k,M2}$; $p_{k,M1} > p_{k,M3}$; $p_{k,M2} > p_{k,M3}$.

Additionally, following Sturgis and Smith (2010), we also expect that individual associations with known others positively influence trust scores (H2) across all three measures. For instance, when calculating the aggregate mean level of trust, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, where y_i is an individual i's reported trust score, we could expect a positive difference in trust between the subset of respondents who think of known others and respondents who think of unknown others. Estimating such differences could help us identify the measurement error that is included in common aggregate estimates of trust scores.

2.2.2 Negative associations

While trust research regularly discusses the impact of experiences on trust (Brehm & Rahn, 1997; Cao et al., 2014; Dinesen, 2010; Freitag & Traunmüller, 2009; Glanville & Paxton, 2007; Glanville et al., 2013; Uslaner, 2002), studies about trust measurement have neglected this dimension. On average, trust in known others is higher (Bauer & Freitag, 2018; Sturgis & Smith, 2010; Vollan, 2011) – as is also evidenced by measures that directly gauge trust in family members, neighbors, etc. (Freitag & Traunmüller, 2009; Nannestad, 2008). Theoretically, however, this does not always have to be the case. In fact, some of the more important betrayals of trust in our lives may happen through people we know. For instance, a close friend may spill our secrets or a family member may fail to return a loan. Referring to Figure 2.1, Hans's response may be based on a negative association as opposed to Hanna's response. Put differently, we may collect negative (or positive) experiences with known others just as we may collect negative (or positive) experiences with unknown others, i.e., strangers. Independently from whether a trustee is known or unknown, individual associations that emerge when answering survey questions may vary in terms of their sentiment. Hence, we also want to measure the proportion of respondents who have negative associations, expressed as $p_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$, where Y_i is a dummy that indicates whether individual i's association can be classified as negative (1) or not (0).[9]

Again, the share of negative associations may depend on the measure we use. Since M2 (in contrast to M1) explicitly asks respondents to think of first-time encounters ("people you meet for the first time"), we expect that this question wording may evoke more negative associations than the most people question. This could be either because respondents remember past first-time interactions that turned out to be negative and/or because we are generally taught to be careful in first-time encounters. M3, then, explicitly specifies the trustee as a stranger. The term "stranger" has a rather negative connotation in English compared to the more neutral terms "people" or "person". "Stranger danger" describes the idea that all strangers can potentially be dangerous. In countries such as Great Britain, stranger-danger education often conducted by local police force has the objective

[9] Where the latter —0— category comprises both neutral and positive associations.

to teach children to refuse offers from strangers (Ellen et al., 1999, p.11). Postulating H1, we assume that M2 and M3 result in higher conceptual validity (i.e., lower share of associations of known others) which is desirable. However, finding that M3 or M2 in comparison to M1 result in more negative sentiment would be undesirable as it could indicate that using concepts such as "stranger" in M3 affects respondents' mindset.

We hypothesize that changing trustee categories (most people $\rightarrow$ people you meet for the first time $\rightarrow$ a total stranger you meet for the first time) increases the proportion of respondents who have negative associations (p_n) (H3). Again, evidence for H3 would be provided by statistically significant differences between those proportions: $p_{n,M1} > p_{n,M2}$; $p_{n,M1} > p_{n,M3}$; $p_{n,M2} > p_{n,M3}$. We also expect that negative associations should negatively influence trust scores (H4) across all three measures. Thus, when calculating the mean level of trust $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, where y_i is an individual i's trust score, we expect a negative difference between the subset of respondents who have negative associations and those who do not have negative associations with M1, M2 and M3.

2.2.3 Situative trust measures

Empirical operationalizations of generalized trust, for example M1–M3, depict trust as a "one-part relationship, where neither B [the trustee] nor x [expected behavior] enters explicitly" (Nannestad, 2008, p.415). In contrast, conceptual work argues that trust is a three-part relationship, in which A (truster) trusts B (trustee) with respect to some behavior X (Cook et al., 2005; Schilke et al., 2021). Ermisch et al. (2010) criticize common survey measures of generalized trust to be too generic since the "[...] answers do not reveal either the reference group or the types of action or the stakes that respondents have in mind when making such an assessment" (2010, p.750). Their notion of trust includes a situative character, because they describe a trust situation to be characterized by "trust that someone will do X" (Ermisch & Gambetta, 2010; Ermisch et al., 2009, p.4, p.751).

The measures we investigate (M4.1–4.4) follow this conceptual work and include the context in which a trust decision takes place. This context entails two components, the trustee category, and the trustee's expected behavior in a certain sit-

uation. Importantly, the decision to trust in situation A may not carry over to situation B (Ermisch & Gambetta, 2010, p.4) even though both situations involve the same trustee. We argue that situative trust measures may be able to solve some of the problems that characterize the more vague standard measures of generalized trust. Since the latter do not specify either of the two components of context, respondents may simply fill in such specifications themselves.

Our study investigates situative trust measures introduced by Robbins (2021). These novel measures are based on the stranger question (M3) because they specify the trustee to be a stranger (cf. M3) (see Buskens & Weesie, 2000; Yamagishi & Yamagishi, 1994; Yuki et al., 2005, for similar approaches). Further, they specify the expected behavior of the trustee, namely keeping a secret (M4.1), repaying a loan (M4.2), providing advice on managing money (M4.3), and looking after a child/family member/loved one (M4.4). Unlike the stranger question (M3) that allows for varying interpretations by respondents, these situative measures provide a more specific context, leaving less room for ambiguity. This avoids situations where different respondents envision different scenarios, potentially leading to varying trust values (cf. Figure 2.1). Analogous to H1, we hypothesize that by specifying the trustee as a total stranger, as opposed to most people or people you meet for the first time, the proportion of respondents associating trust with known people (p_k) will decrease (H5). As these situative measures are relatively new, we do not have specific expectations regarding the negativity of associations they may evoke or how they compare to each other. It is plausible that questions concerning money lending or money advice could elicit negative associations or memories. The question is, however, whether they do so systematically. Therefore, the empirical insights we present below are exploratory in nature.

2.3 Data, experimental design, and methods

2.3.1 Sample

Our target population are U.S. citizens. Data was collected using a two-stage non-probability sample recruited by Prolific, a participant recruitment and payment software to conduct online surveys and experiments (Palan & Schitter, 2018). First, respondents were identified to be eligible according to quotas on self-reported

gender, age, and ethnicity in accordance with the U.S. Census Bureau population group estimates from 2015.[10] Second, out of 43,131 panelists that were considered eligible, we continued to collect data until our target and final sample size of n=1,500 was reached. Respondents who did not complete the questionnaire (n=87, overall response rate=95%) were excluded and replaced with other panelists who would fit the quotas. Summary Statistics for all variables and their comparison to population estimates can be found in Appendix A.1. The survey was fielded between July 14, 2021 and July 21, 2021. For each completed survey, we paid a wage of 9.60 USD/hr on average while the mean duration was 6.8 minutes.

2.3.2 Experimental design and measures

Our questionnaire design is depicted in Table 2.1. Respondents provided their data via an online self-administered survey (created using formR, cf. Arslan et al., 2020). The survey started with information on its objective and a consent form. Subsequently, respondents received two blocks of questions. Block #1 included the standard generalized trust measures with respective probing questions and Block #2 included situative trust measures with respective probing questions. Since we wanted to avoid priming effects (meaning subsequent answers might be influenced by previous questions) we used an experimental design in which the order of questions is randomized. Specifically, the order of Block #1 and #2 as well as the question order within these blocks was randomized. This design allows us to conclude that the differences we find between the trust measures for the outcomes we examined (i.e., the proportion of associations that refer to known individuals or are negative) are actually due to the wording of the question and not to the order of the questions.

Furthermore, data collected with this questionnaire allows for within- and between-person comparisons for each variable because each respondent received all available trust questions in Block #1 and #2 in a randomized order. To allow further examination of the role of question order despite the introduction of random question order, we can consider two data subsets: Subset 1 only includes respondents'

[10] Gender: two groups, namely males and females; Age: five groups in 10-year brackets; Ethnicity: five groups, namely White, Mixed, Asian, Black, and Other.

	——— Survey direction ———>		
	Order of Blocks #1 and #2 is randomized		
Intro	**Block #1: Generalized trust measures Randomized question order and probe after all three questions**	**Block #2: Situative trust measures Randomized question order and probe after question #1 and #4**	**Additional questions**
Information consent form	M1: Most people question M2: People first time question M3: Stranger question	M4.1: Keep secret M4.2: Repay loan M4.3: Money advice M4.4: Look after child	Socio-demographics (see Online Appendix A.2)

Table 2.1: Experimental Design.

responses to the first trust question they received (ignoring the order of the blocks) and is called "first question only" below; Subset 2 includes respondents' responses to the first trust question from the first block only and is called "first question and first block only" below. While there might still be priming from the preceeding block for Subset 1, this possibility should be excluded for Subset 2.

Block #1: Generalized trust measures and probing questions

In Block #1, we assessed generalized trust using three established measures: trust towards "most people" (M1), "people you meet for the first time" (M2), and "a total stranger you meet for the first time" (M3). These measures had different response categories: 7-, 4-, and 4-point scales for M1, M2, and M3, respectively. To ensure comparability, we employed min-max normalization, which rescales the responses to a range between 0 and 1 while preserving the original distribution. We treat the resulting variable as continuous for all our analyses.[11] The

[11] By introducing this assumption, an ordinal-level measure becomes an interval-level measure with discrete categories (Blaikie, 2003). Carifio and Perla (2007) and Glass, Peckham, and Sanders

specific phrasing as well as summary statistics of these questions can be found in Appendix A.2. Directly after respondents answered these closed-ended questions, each was followed by an open-ended probing question using the following wording (exemplary for M1): "In answering the previous question, who came to your mind when you were thinking about 'most people'? Please describe". Our specific interest here is to elicit who respondents had in mind when they were exposed to the three different trustee categories.[12]

Block #2: Situative trust measures and probing questions

Block #2 included four situative measures that represent the Imaginary Stranger Trust Scale (IST) developed by Robbins (Robbins, 2021, 2022, 2023). These measures specify the trustee category as well as the content of the trust relationship, overall aiming to reduce the vagueness we argued to find for the standard generalized trust measures from Block #1. The four items elicit trust in a total stranger met for the first time to[13], (1) "keep a secret that is damaging to your reputation" (M4.1), (2) "repay a loan of one thousand dollars" (M4.2), (3) "provide advice about how best to manage your money" (M4.3) and to (4) "look after a child, family member, or loved one while you are away" (M4.4). Each of these items was rated on a 4-point scale. We applied min-max normalization to rescale these items to a range between 0 and 1.

Again, the question order was randomized. Analogous to Block #1, the situative measures were also probed using the following wording: "In answering the previous question, who came to your mind when you were thinking about 'a total

(1972) describe how Monte Carlo Simulations have shown that parametric tests, such as a F-Test in a linear regression, are strongly robust to the interval data assumption (as well as moderate skewing) when data was collected using a 5 to 7 point Likert response format (preferably 7) with no resulting bias.

[12] In crafting the above wording, we deliberately chose to repeat the closed-ended question. This decision was based on pretesting the questionnaire with independent testers, considering their feedback, and being guided by relevant literature on probing techniques (e.g., Behr et al., 2012). Research has shown that repeating the wording can lead to more informative answers compared to presenting the probe without context (Behr et al., 2012). In principle, repetitions of question wording in probing questions could create demand effects and further research using appropriate randomized designs to study such effects are necessary.

[13] A randomly selected share of respondents was assigned an alternative wording to the one describing the trustee as a stranger met for the first time, namely which describes the trustee as a person met for the first time (question wordings can be found in Appendix A.2).

stranger you meet for the first time'? Please describe.". To avoid memory effects as well as errors due to response fatigue, we only probed the situative measures that were randomly assigned to come first and fourth.

2.3.3 Methods

Table 2.2 illustrates the structure of our data. Due to the intra-person design, there are multiple measures of trust (i.e., 7) (indicated by the column 'Measure') for each respondent alongside their respective trust score (column 'Trust'). Overall, we collected open-ended responses using five open-ended probing questions and received 7,497 out of potentially 7,500 text answers (column 'Probing Answer').[14] Appendix A.3 provides a detailed description of the open-ended text answers. Table 2.2 also displays the results for our classification of the open-ended responses (columns 'Associations (known–unknown others)' and 'Associations (sentiment)'). Both approaches are described in detail below.

[14]Each respondent was probed for each generalized trust measure (M1 – M3), resulting in $3x1,500$ entries, as well as for two out of four situative trust measures (M4.1 – M4.4), resulting in additional $2x1,500$ entries. Out of 10,500 answers to trust questions, 3,000 responses were not probed.

ID	Measure	Trust	Probing Answer	Associations (known-unknown)	Associations (sentiment)
123	Most people	0.33	I was thinking of people I don't know personally.	0 (No)	0 (neutral/positive)
3139	Most people	0.17	Tourists that come to our little village. I tend to be very wary of them.	0 (No)	1 (negative)
7214	People first time	0.33	My friends back in high school.	1 (Yes)	0 (neutral/positive)
7304	People first time	0.67	No specific person	0 (No)	0 (neutral/positive)
1365	Stranger	0.67	A person sitting next to me at a game	0 (No)	0 (neutral/positive)
2980	Stranger	0	No one in particular, but I don't think I could trust anyone ever again.	0 (No)	1 (negative)
1289	Keeping a secret	0	An anonymous, faceless man was my first thought, perhaps someone in a train or bus station.	0 (No)	0 (neutral/positive)
1487	Repaying a loan	0	White man, about 60, good looking, widower	0 (No)	0 (neutral/positive)
4286	Watching a loved one	0	A former neighbor of mine who was a single father with a son close to my son's age.	1 (Yes)	0 (neutral/positive)
1	Money advice	0	Just a random stranger.	0 (No)	0 (neutral/positive)
...	...	...	...	...	...

Table 2.2: Illustration of exemplary data. *Note*: The table displays different exemplary respondents. Note that in the actual dataset each respondent/ID (cf column 1) appears seven times because each respondent received all 7 trust items (for 5 of these questions the respondents received a respective probing question).

Both classifications (i.e., known–unknown and sentiment) were achieved using automated text analysis, which in survey data research has become a popular alternative to manual coding (Esuli & Sebastiani, 2010; Giorgetti & Sebastiani, 2003; Gweon & Schonlau, 2023). In particular, we pursued a supervised classification approach in which randomly sampled subsets of text answers were manually labeled and only the remainder were automatically classified using fine-tuned BERT models.

For the known–unknown classification, we manually labeled a sample of n=1,000 text answers, while for the sentiment classification, we increased[15] this number to n=1,500. Both samples were a random selection of text answers from the generalized trust measures (see Appendix A.5 for further details). Based on previous implementations in the literature, we argue that these sample sizes are sufficiently large.[16]

Both manual classification tasks were achieved using a hand-crafted coding scheme. For both schemes the main distinction lies between two categories. In the known–unknown classification, Category 0 was assigned when respondents mentioned individuals or groups of individuals that can be identified as "unknown others" in their text answer. Importantly, our primary focus was on identifying respondents' personal unfamiliarity with these individuals or groups, and not on the specific characteristics of these individuals/groups. For example, an answer that describes personally unknown others that have rather specific characteristics (i.e., tourists in ID 3139 in Table 2.2 falls into category 0.[17] Code 1 on the other hand subsumes all statements that made mentions of "others known" to the respondent. Survey answers that had no references to either known or unknown others (e.g., "just people

[15]Detecting sentiment proves more complex than spotting mentions of known and unknown others due to several factors, such as ambiguous word meanings.

[16]Schonlau and Cooper (2016) for instance show that 500 observations suffice for training the task of categorizing open-ended survey answers and that additional time savings could be attained by reducing the training data to even 300 or 200 observations, but only for less complex problems. Not only but also because Schonlau and Cooper (2016) are concerned with a multinomial rather than a binary classification problem (i.e., the latter is a less complex task), our training data of n=1,000/1,500 should be large enough. In general, automated categorization is shown to result in meaningful time savings as opposed to manual classification as soon as the data to be classified exceeds 1,500 documents (Schonlau & Couper, 2016).

[17]Coding of the n=1,000 training data observations shows that circa 9% of the answers include mentions of "groups of people", these instances were all coded as "unknown others".

as a whole") were coded as 0, and survey answers with mixed references to both known and unknown others (e.g., "People I may run into everyday.") were coded as 1. To label sentiment, the main distinction lies between "negative sentiment" (Code 1) and "neutral or positive sentiment" (Code 0). Appendix A.4 provides an overview of the coding schemes with examples and descriptions of all available codes.

The manual classification was carried out by three independent coders. All three coders assigned codes to the same 1,000/1,500 text answers, and conflicts were resolved by finding consensus between the coders or using majority vote.

For the remainder of text answers (i.e., n=6,500/6,000), we fine-tuned the weights of two BERT models (BERT base model in its uncased version), using the manually coded data (n=1,000/1,500) as training data. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) is an empirically powerful machine learning technique that can be used for various natural language processing tasks (Devlin et al., 2018, p.1). BERT comes with two attributes that are of special importance here: first, it is able to model contextual representations by incorporating both the left and right context of a document (i.e., bidirectional). Second, BERT provides pre-trained vector representations for words by using a deep, pre-trained neural network. These so-called embeddings suggest a representation for each term based on its context by using information from the entire input sequence. For our data, this could mean, for example, that terms that appear in the (pre-trained) context of “family”, e.g. brother and sister, are likely to be predicted as “known other”. Last but not least, by using BERT, we aim at addressing the class imbalance that is present in our sentiment data insofar as few respondents (8.7%) have negative associations. BERT achieves higher class-wise accuracy in the presence of class imbalance than other ngram-based machine learning techniques (Gweon & Schonlau, 2023), and is further demonstrated to remove the need to use data augmentation techniques to mitigate problems of imbalanced data (Madabushi et al., 2020).[18] Importantly, the imbalanced data structure and its consequences does not call into question the effects we found but may have

[18]Still, we attempted oversampling (see e.g., Gosain & Sardana, 2017) the minority class to address the problem of class imbalance. This however did not lead to any further significant improvements. Results are available upon request.

resulted in their slight underestimation. Appendix A.5 shows our findings when using the manually classified data only.
A detailed evaluation of the two classifiers in terms of accuracy, precision, recall and F1-Score can be found in Table 2.3.

Associations (known-unknown)				Associations (sentiment)			
	Precision	Recall	F1 Score		Precision	Recall	F1 Score
0	0.87	0.95	0.91	0	0.97	0.97	0.97
1	0.86	0.71	0.78	1	0.68	0.72	0.70
accuracy			0.87	accuracy			0.95
macro avg	0.87	0.83	0.84	macro avg	0.83	0.84	0.84
weighted avg	0.87	0.87	0.87	weighted avg	0.95	0.95	0.95

Table 2.3: Accuracy, precision, recall, and F1-score.

Alternative approaches with which we classified our data (i.e, regular expressions, Random Forest) can be found in Appendix A.6.

2.4 Results

2.4.1 Trust scores across standard and situative measures

We begin by assessing the variations in trust scores obtained from our seven trust measures across different sample specifications (Figure 2.2). Regardless of the subsample, there is a gradual decline in trust from Measure 1 (most people question) to Measure 2 (people first time question), and finally to Measure 3 (stranger question).

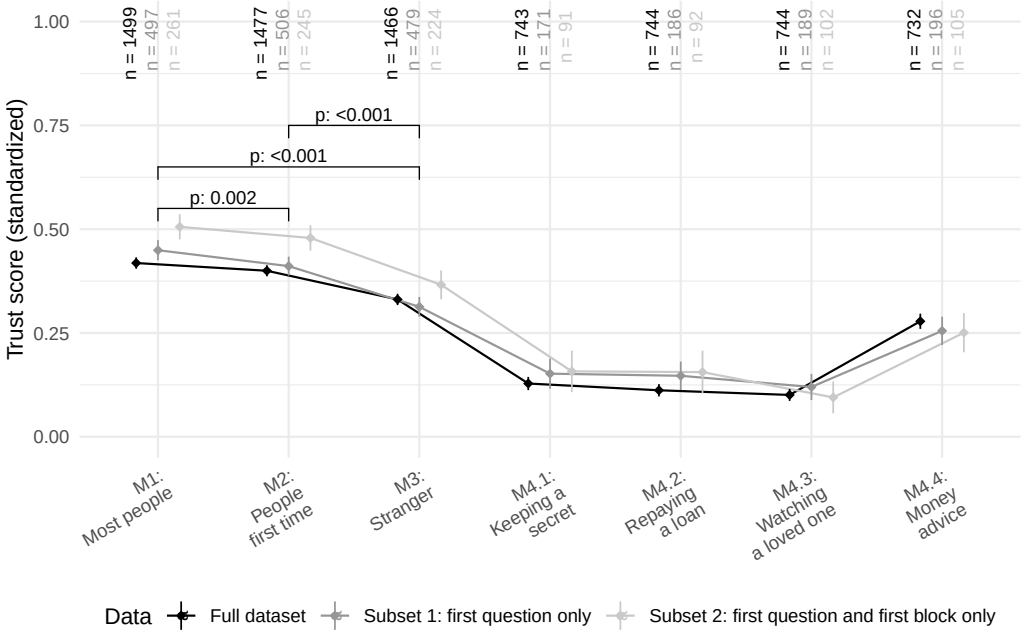

Figure 2.2: Standardized trust scores across different trust measures and respondent subsets. *Note*: The figure shows point estimates for average trust scores and 95% confidence intervals. Estimates for the full long-format dataset are colored in black, those for Subset 1 in dark gray, those for Subset 2 in light gray. Details on the respondent subsets are provided in the Methods Section. P-values are derived from t-tests for the Full dataset, for details see footnote 19. Data for M4.1-4.4 include the 'stranger' wording only (see Footnote 13).

Within-subjects ANOVA reveals that the generalized trust scores differed statistically significantly for the same individual for the three question wordings (F(1.7, 2,505)=129, $p < 0.001$).[19]

Additionally, situative trust measures M4.1 to M4.4 consistently exhibit lower trust levels likely owing to their emphasis on trust decisions where the truster has a lot to lose.[20] It is crucial to note that Figure 2.2 provides a descriptive overview

[19]Moreover, we investigated the full dataset via paired sample t-tests with a Bonferroni adjusted alpha level of .016 per test (.05/3): on average, the trust score for M1 (M = 0.42, SD = 0.27) was significantly higher than the trust score for M3 (M = 0.33, SD = 0.27), t(1,464) = 13.81, p<0.001. Furthermore, but to a lesser extent (as is also depicted in Figure 2.2), M1, on average, results in higher trust scores than M2 (M = 0.4, SD = 0.26), t(1,475) = 3.11, p<0.01. Also, the differences in trust scores for M2 and M3 are statistically significant, t(1,455) = 15.15, p < 0.001.

[20]To address potential outliers in individual situations, we propose exploring the concept of "cross-situational trust" (Bauer and Freitag 2018) and computing an average across measures (see

of the seven measures concerning their sample means. The observed differences may be influenced by various factors, such as question interpretation, demand effects, and scale effects. In our subsequent analysis, we focus on examining one specific factor: the associations formed by respondents when answering our trust survey questions.

2.4.2 Associations across standard and situative measures

We start by examining the known–unknown dimension. Figure 2.3 displays the share of respondents who described associations of either known or unknown others across our seven measures.[21] In line with our expectation (H1), the share of respondents referring to a known other statistically significantly decreases for M3 (i.e., 13%) while shares for M1 and M2 are similar (31% and 30%, respectively). The share of respondents referring to a known other again increases for our situative measures M4.1 – 4.4, however, none of these differences are statistically significant. Nevertheless, it could indicate that referring to specific situations and behaviors in those survey questions could increase the number of respondents who think of known others. This is undesirable from a conceptual perspective.

---

our detailed idea and discussion on this in the conclusion). This approach could help mitigate the impact of strong outliers from specific situations.

[21] Appendix A.5 shows these results using data from the manually coded share of data only (n=1,000/1,500). Appendix A.5 shows these results using data for Subset 2 only (n=1,500).

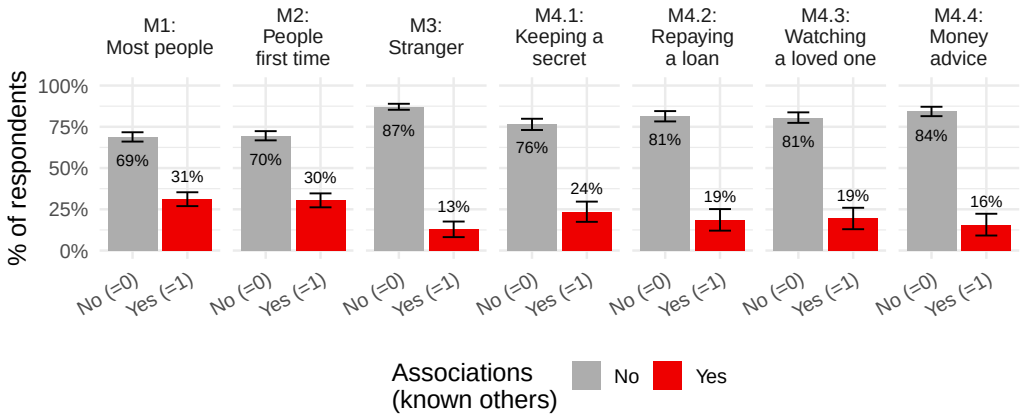

Figure 2.3: Distribution of associations with known people across trust measures. *Note*: Error bars represent 95% confidence intervals (lower cutoff at 0). Data is the full dataset irrespective of the question or block randomization (details are provided in the Methods Section). Results for different Subsets of the data can be found in Appendix A.5.

With regards to the sentiment dimension, we expected to find different shares of negative sentiment for each question wording (see Figure 2.4). In line with our expectations (H3), the share of negative associations is higher for M3 (i.e., 8.7%) compared to M2 (7%). Not in line with our hypothesis, the share for M1 is higher (10%). However, none of these differences are statistically significant. Moreover, the share of negative associations remains similarly low for the situative measures, which is in accordance with the findings for M3 since the situative measures also describe the trustee category to be a "stranger".

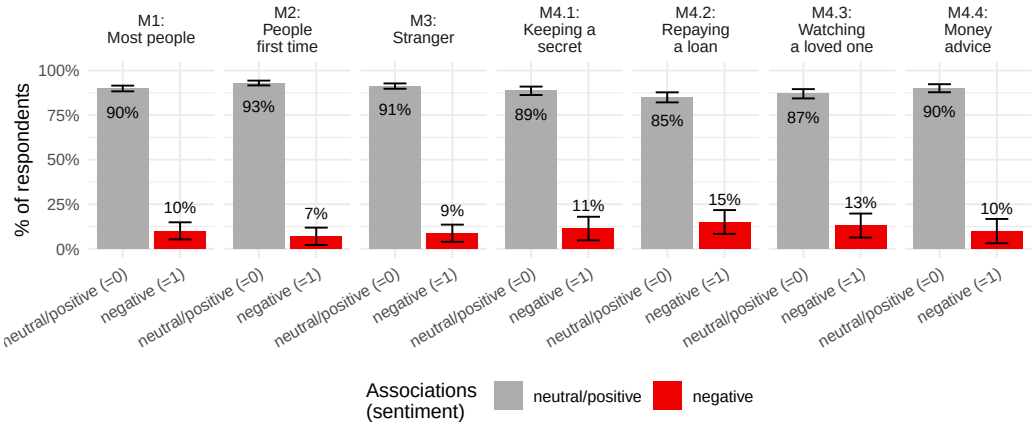

Figure 2.4: Distribution of associations and their sentiment across trust measures. *Note*: Error bars represent 95% confidence intervals (lower cutoff at 0). Data is the full dataset irrespective of the question or block randomization (details are provided in the Methods Section).

In sum, we find that, across all seven measures, there are respondents who have associations with known others as well as associations of negative sentiment. However, strong differences between measures in terms of associations can only be found for the known–unknown dimension. The sentiment dimension seems less relevant. The two classification dummies only correlate weakly (r(7,490) = -0.08, $p =< 0.001$).

2.4.3 Associations and trust scores

Above we demonstrated that there is variation in associations across individuals. Next, we examine whether different associations affect the measurement values. Figure 2.5 visualizes the coefficients for a series of regression models (see Appendix A.9 for detailed regression tables). We estimated five models for each of our seven trust measures which are indicated on the left side. Two models are bivariate and only include one of the association dummies (e.g., Model #1 and #2 in Figure 2.5). We subsequently add covariates to these bivariate regressions (e.g., Model #3 and #4 in Figure 2.5).[22] Finally, the fifth model includes both dummies in one model and adds covariates.

[22] Age (catgeorical), sex, ethnicity, socioeconomic status, income, and education.

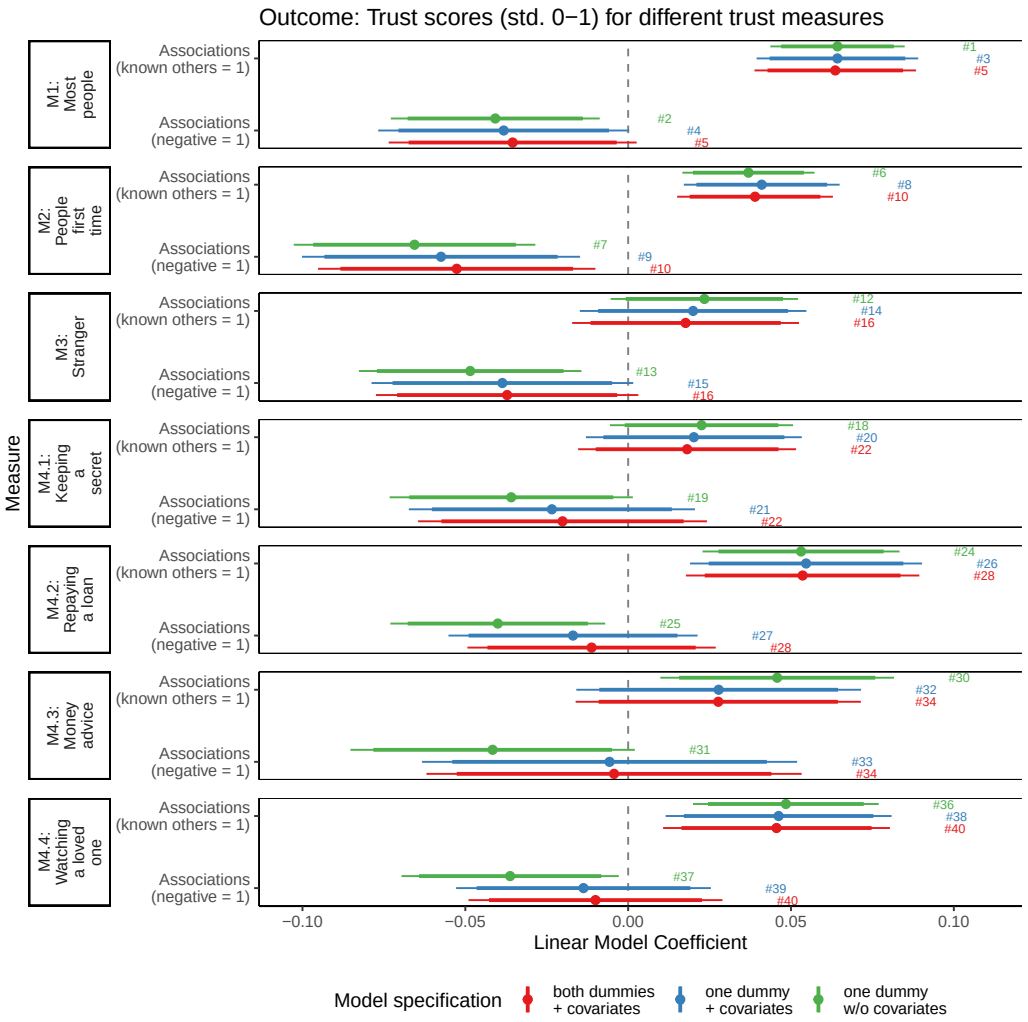

Figure 2.5: Associations and trust scores across different measures. *Note*: The figure shows point estimates for coefficients of our dummy variables of interest namely having associations with known others or negative associations. Bars represent 90% (thicker) and 95% (thinner) confidence intervals. Data is the full dataset irrespective of the question or block randomization (details are provided in the Methods Section).

In accordance with our expectations (H2), we observe that associations with known others have a positive effect on trust for all three of our generalized trust measures, M1, M2, and M3 ($\beta_{\#1} = 0.064$, $\beta_{\#6} = 0.037$, and $\beta_{\#12} = 0.023$, respectively). While this effect is especially pronounced for M1 and M2 in terms of effect size

and statistical significance (p<0.001), it becomes smaller and less robust for M3. This may be due to the fact that M3 evokes associations with known people in fewer respondents than M1 an M2 do (see Figure 2.3), thus resulting in a smaller sample size of that subgroup, increasing the uncertainty of the corresponding estimate. In addition, adding the sentiment dummy as a control variable in Models #5, #10 and #16 (see Figure 2.5) does not mitigate the effect of the known-unknown dummy on trust.

In line with our expectation (H4), we find that negative associations have a negative effect on trust for all of our three generalized trust measures M1, M2, and M3 regardless of the control set specifications ($\beta_{\#2} = -0.041$, p<0.01; $\beta_{\#7} = -0.066$, p<0.001 and $\beta_{\#13} = -0.049$, p=0.059, respectively). While the different generalized trust measures are not affected differently, we suggest that the role of negative associations for trust measurement requires future research.

Also for the four situative measures, the effects are in line with H2. Associations with known people have a positive effect on for example M4.4, trusting someone to watched a loved one ($\beta_{\#36} = 0.053$, p<0.001), or on M4.2, i.e., trusting someone to repay a loan ($\beta_{\#24} = 0.053$, p<0.001). For the situative measures, however, while consistent with (H4), we find smaller and less robust effects for our dummy capturing negative associations.

In sum, for the generalized trust measures, we find statistically significant effects in our hypothesized directions, namely that associations with known others (in contrast to unknown others) influences trust scores positively and that negative sentiment (in contrast to neutral/positive sentiment) influences trust scores negatively. Especially the effect of the dummy capturing the known–unknown dimension is undesirable from a conceptual point and its effect varies across measures of generalized trust. We can conclude that estimates based on the three classic measures – M1, M2 or M3 – overestimate trust scores because they do not measure generalized trust for a significant share of the respondents. Without these respondents, our estimated trust averages would differ (namely by the coefficients we depict in Figure 2.5 for the bivariate models). The bias is smallest for the stranger measure M3 and all four of the situative measures seem to be characterized by the same problem.

2.5 Discussion and conclusion

Generalized social trust is a foundational concept in the social sciences. However, there have been doubts about the validity of commonly used measures (Delhey et al., 2011; Ermisch & Gambetta, 2010; Nannestad, 2008; Robbins, 2022; Sturgis & Smith, 2010). In our study, we examined various trust survey measures in a U.S. sample and explored how respondents answered those questions. To eliminate interviewer effects, we used a web probing approach (Behr et al., 2012, 2017; Meitinger & Kunz, 2022). Open-ended probing (Neuert et al., 2021) is still a novelty in trust research, and similar data has so far only been collected in interviewer-administered settings (Sturgis & Smith, 2010; Uslaner, 2002). The data collected through open-ended probing was analyzed using a supervised machine learning approach. Our findings can be categorized into four key aspects. First, our study revealed significant variations in overall and intra-individual reported trust levels across different question formats, and the question employing the phrase "most people" yielded the highest average trust score (cf. Figure 2.2). This finding suggests that the different question formats should not be considered interchangeable measures of generalized trust. However, it is important to note that Figure 2.2 provides only a descriptive overview, and our subsequent analysis centered on exploring the associations formed by respondents while answering the trust survey questions.

Second, we delved into the associations respondents made when responding to the questions. We described generalized trust as trust in unknown others, and argued that it should ideally be measured accordingly. Remarkably, a notable proportion of respondents (ranging from 13% to 31%, cf. Figure 2.3) incorporated thoughts of known individuals in their responses while answering classic trust questions, which is in line with previous research (e.g., Sturgis & Smith, 2010). Hence, for this particular group of respondents, classic trust measures actually do seem to capture what is commonly known as particularized trust (cf. Freitag & Traunmüller, 2009). In other words, for these respondents, our measures suffer from construct invalidity. However, the proportion of mentions of known individuals in responses decreased for the "stranger" question (M3), suggesting a higher degree of construct validity for this measure (in line with Robbins, 2022, 2023).

Interestingly, compared to M3, the situative measures (M4.1 - M4.4) showed an increase in respondents thinking about known individuals (but still considerably smaller than in M1 and M2) (cf. Figure 2.3), despite being instructed to consider the trustee as a stranger. This outcome may be attributed to respondents drawing upon their past experiences to contextualize and anchor the given situations.

Thirdly, we conducted an examination of the influence of associations on trust levels. If confirmed, this would imply that trust estimates produced by specific measures (e.g., the "most people" wording) could be biased, potentially leading to an overestimation of generalized trust in diverse populations. Indeed, we found that respondents who reported thinking about known others displayed higher levels of trust across all three generalized trust measures (cf. Figure 2.5). The effects were less robust for the stranger question (M3), which might be due to the smaller share of respondents having known others in mind when answering. This is a desirable feature of the latter measure.[23] Overall, this finding demonstrates that differences in trust between individuals and over time may not be solely reflective of variation in the substantive dimension of trust. Instead, they might be influenced by specification errors and differences in how respondents interpret the question due to inter-individual differences in frames of reference.

Fourth, we also explored a hitherto neglected dimension – the sentiment of association. We found a relatively low proportion of respondents reporting negative associations which remained consistent across measures (cf. Figure 2.4). Against our expectations, M3, the stranger-question (without situations) does not seem to evoke more negative associations than the most people and people first time question. While negative associations did influence trust scores negatively, the effect was not uniform across measures and models (cf. Figure 2.5). These findings offer encouraging insights into measurement, yet we call for further research to explore whether specific question formats trigger more emotional responses or negative memories. Our study yields several key findings that not only allow us to draw valuable conclusions but also pave the way for future research directions.

Firstly, among the trust questions we investigated, our various "stranger" questions

[23] Analogous to Sturgis and Smith (2010), we randomized respondents to trust measures in Block #1 and #2; hence, we can conclude that the differences in the distribution of associations are the result of divergent frames evoked by the questions in respondents' minds.

(M3, and M4.1 to M4.4) demonstrated the highest level of construct validity, as evidenced by the lower share of respondents thinking of known individuals. However, from an empirical perspective, we may question how many trust situations actually take place among total strangers. For example, the four situations in our study are more likely to take place among individuals who have some knowledge about each other (e.g., acquaintances). Certainly it can be challenging to pinpoint situations that entirely lack associations to known others, but we think that further theoretical work is necessary to classify based on whether a trust measure primarily pertains to strangers or also encompasses acquaintances.[24]

Secondly, researchers should carefully consider various factors when selecting measures for their studies, aligning with their specific definition of generalized trust. Our findings indicate that M3 best captures generalized trust when defined as trust towards unknown others (cf. Figure 2.3). However, for those interested in interpersonal comparability, situative measures like the Imaginary Stranger Trust Scale (IST) offer a viable alternative, since they explicitly define the concrete situation in which trust has to be placed and thus leave less room for different interpretations. Nonetheless, they demand additional questionnaire space due to longer item descriptions.[25] Generally, future studies could make use of additional, situative measuresby using vignette designs. The resulting data could be analyzed in such a way, that one caclulates the average trust across a set of situative trust measures, yielding a score of what we call cross-situational trust (Bauer & Freitag, 2018; Robbins, 2023).[26] However, we would also like to emphasize that the use of traditional measures such as M1 and M2 may be justified if the main objective is comparability with previous studies using these measures or corresponding panel studies.

Thirdly, our study focused on a U.S. sample, expanding on prior evidence from

[24]It may be beneficial to explore the semantic meaning of the term "stranger" and consider situations where individuals might perceive acquaintances as strangers for specific trust decisions, such as lending money. This highlights the situative nature of trust, where perceptions may vary depending on the context of the interaction (Hardin, 2002, p.9).

[25]For considerations between short and long versions of IST, see Robbins (2023).

[26]This approach could extract an individual specific general personal component of trust while acknowledging trust to be inherently situational, mitigate the effects of non-valid associations in single items and provide a more robust assessment of trust across diverse situations. A high-truster would then be someone who has a high-level of trust across a large set of situations that involve trust.

the UK (Sturgis & Smith, 2010). While we expect similar findings in other populations, we lack direct evidence to support this claim. The lack of interpersonal comparability within a "homogeneous" sample of U.S. citizens may be amplified when comparing individuals from different cultures, countries, and languages. Nevertheless, we must exercise caution in generalizing our conclusions to other samples.

Fourthly, the main aim of this study was to examine established measures as they have been used for decades. This implied that we use original wordings characterized by answer scales of different length (e.g., 4pt and 7pt). Although we assume scale length does not significantly affect our main variable of interest (i.e., shares of associations), a potential full-factorial design (7x2) where all seven items are measured with both scales, could explore any subtle differences in greater detail. Also, we used a particular set of emerging measures (i.e., IST Robbins, 2021, 2022), and considering other emerging measures, such as the Risk Aversion question in the GSOEP and the UK Household Longitudinal Study[27], could provide valuable insights.

Fifth, we employed a probing technique (see Experimental Design Section) that restated the trustee category originally presented (e.g., "In answering the previous question, who came to your mind when you were thinking about 'most people'?"). Repeating this category could be regarded as a form of priming potentially creating demand effects (cf. Fn 6). For future research, exploring various probing strategies and utilizing designs that provide respondents with as little information as possible, and thereby avoiding any priming, could be a valuable avenue to pursue.

Finally, an open question emerges concerning whether frames of reference are systematically linked to respondents' demographic characteristics. Preliminary correlational evidence (see Appendix A.7) seems to show that this is not the case. This is encouraging and could mean that associations are predominantly random. However, to gain further clarity, future studies could extend the set of covariates considered and potentially employ a randomized design that attempts to induce associations of a particular kind to avoid post-hoc rationalization.

[27] "Are you generally a person who is fully prepared to take risks in trusting strangers or do you try to avoid taking such risks?".

References

Anandarajan, M., Hill, C., & Nolan, T. (2019). Term-Document representation. In M. Anandarajan, C. Hill, & T. Nolan (Eds.), *Practical text analytics: Maximizing the value of text data* (pp. 61–73). Springer International Publishing.

ANES. (2000). Anes Pilot Study (codebook variable documentation).

Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). Formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behav. Res. Methods*, *52*(1), 376–387.

Barr, A. (2003). Trust and expected trustworthiness: Experimental evidence from zimbabwean villages. *Econ J*, *113*(489), 614–630.

Bauer, P. C., & Freitag, M. (2018, March). Measuring trust. In E. M. Uslaner (Ed.), *The oxford handbook of social and political trust* (pp. 1–27). Oxford University Press.

Behr, D., Kaczmirek, L., Bandilla, W., & Braun, M. (2012). Asking probing questions in web surveys: Which factors have an impact on the quality of responses? *Soc. Sci. Comput. Rev.*, *30*(4), 487–498.

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). Web probing – implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions. *Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines)*.

Blaikie, N. (2003). *Analyzing quantitative data*. SAGE Publications Ltd.

Brehm, J., & Rahn, W. (1997). Individual-level evidence for the causes and consequences of social capital. *Am. J. Pol. Sci.*, *41*(3), 999.

Breiman, L. (2001). Random forests. *Mach. Learn.*, *45*(1), 5–32.

Buskens, V., & Weesie, J. (2000). An experiment on the effects of emebeddedness in trust situations: Buying a used car. *Ration. Soc.*, *12*(2), 227–253.

Cao, J., Galinsky, A. D., & Maddux, W. W. (2014). Does travel broaden the mind? breadth of foreign experiences increases generalized trust. *Soc. Psychol. Personal. Sci.*, *5*(5), 517–525.

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. *J. Soc. Sci.*, *3*(3), 106–116.

Coleman, J. S. (1994). *Foundations of social theory*. Harvard University Press.

Cook, K. S., & Cooper, R. M. (2003). Experimental studies of cooperation, trust, and social exchange. *Trust and reciprocity: Interdisciplinary lessons from experimental research.*, *409*, 209–244.

Cook, K. S., Hardin, R., & Levi, M. (2005, June). *Cooperation without trust?* Russell Sage Foundation.

Delhey, J., & Newton, K. (2005). Predicting Cross-National levels of social trust: Global pattern or nordic exceptionalism? *Eur. Sociol. Rev.*, *21*(4), 311–327.

Delhey, J., Newton, K., & Welzel, C. (2011). How general is trust in "most people"? solving the radius of trust problem. *Am. Sociol. Rev.*, *76*(5), 786–807.

Delhey, J., Newton, K., & Welzel, C. (2014). The radius of trust problem remains resolved. *Am. Sociol. Rev.*, *79*(6), 1260–1265.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint: 1810.04805*.

Dinesen, P. T. (2010). Upbringing, early experiences of discrimination and social identity: Explaining generalised trust among immigrants in denmark. *Scan. Polit. Stud.*, *33*(1), 93–111.

Dinesen, P. T. (2012). Does generalized (dis)trust travel? examining the impact of cultural heritage and destination-country environment on trust of immigrants: Does generalized (dis)trust travel? *Polit. Psychol.*, *33*(4), 495–511.

Dinesen, P. T., Nørgaard, A. S., & Klemmensen, R. (2014). The civic personality: Personality and democratic citizenship. *Polit. Stud.*, *62*, 134–152.

Dinesen, P. T., & Sønderskov, K. M. (2015). Ethnic diversity and social trust: Evidence from the Micro-Context. *Am. Sociol. Rev.*, *80*(3), 550–573.

Ellen, M., David, W., Lindsey, M., Gillian, M., & John, G. (1999). Stranger-Danger: What do children know? *Child Abuse Review*, *6*(1), 11–23.

Ermisch, J., & Gambetta, D. (2010). Do strong family ties inhibit trust? *J. Econ. Behav. Organ.*, *75*(3), 365–376.

Ermisch, J., Gambetta, D., Laurie, H., Siedler, T., & Noah Uhrig, S. C. (2009). Measuring people's trust. *J. R. Stat. Soc. Ser. A Stat. Soc.*, *172*(4), 749–769.

Esuli, A., & Sebastiani, F. (2010). Machines that learn how to code Open-Ended survey data. *International Journal of Market Research*, *52*(6), 775–800.

Fehr, E., Fischbacher, U., Rosenbladt, B. v., Schupp, J., & Wagner, G. G. (2002). A Nation-Wide laboratory. examining trust and trustworthiness by integrating behavioral experiments into representative surveys. *J. Context. Econ.*, *122*(4), 519–542.

Freitag, M., & Traunmüller, R. (2009). Spheres of trust: An empirical analysis of the foundations of particularised and generalised trust. *Eur. J. Polit. Res.*, *48*(6), 782–803.

Giorgetti, D., & Sebastiani, F. (2003). Automating survey coding by multiclass text categorization techniques. *J. Am. Soc. Inf. Sci. Technol.*, *54*(14), 1269–1277.

Glanville, J. L., Andersson, M. A., & Paxton, P. (2013). Do social connections create trust? an examination using new longitudinal data. *Soc. Forces*, *92*(2), 545–562.

Glanville, J. L., & Paxton, P. (2007). How do we learn to trust? a confirmatory tetrad analysis of the sources of generalized trust. *Soc. Psychol. Q.*, *70*(3), 230–242.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Rev. Educ. Res.*, *42*(3), 237–288.

Gosain, A., & Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: A review. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 79–85.

Gweon, H., & Schonlau, M. (2023). Automated classification for open-ended questions with bert. *Journal of Survey Statistics and Methodology*.

Hardin, R. (2002, March). *Trust and trustworthiness*. Russell Sage Foundation.

Herreros, F. (2004, July). *The problem of forming social capital: Why trust?* Springer.

Kathuria, A., Gupta, A., & Singla, R. K. (2021). A review of tools and techniques for preprocessing of textual data. *Computational Methods and Data Engineering*, 407–422.

Larsen, C. A. (2013). *The rise and fall of social cohesion. the construction and deconstruction of social trust in the US, UK, sweden and denmark*. Oxford University Press.

Lundmark, S., Gilljam, M., & Dahlberg, S. (2016). Measuring generalized trust: An examination of question wording and the number of scale points. *Public Opin. Q.*, *80*(1), 26–43.

Madabushi, H. T., Kochkina, E., & Castelle, M. (2020). Cost-Sensitive BERT for generalisable sentence classification with imbalanced data. *arXiv preprint:2003.11563*.

Meitinger, K., & Kunz, T. (2022). Visual design and cognition in List-Style Open-Ended questions in web probing. *Sociol. Methods Res.*, 00491241221077241.

Naef, M., & Schupp, J. (2009). Measuring trust: Experiments and surveys in contrast and combination. *SOEPpaper No. 167*.

Nannestad, P. (2008). What have we learned about generalized trust, if anything? *Annu. Rev. Polit. Sci.*, *11*(1), 413–436.

Neuert, C., Meitinger, K., & Behr, D. (2021). Open-ended versus closed probes: Assessing different formats of web probing. *Sociol. Methods Res.*, *52*(4), 1981–2015.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Portes, A., & Vickstrom, E. (2011). Diversity, social capital, and cohesion. *Annu. Rev. Sociol.*, *37*(1), 461–479.

Putnam, R. D., Leonardi, R., & Nanetti, R. Y. (1994, May). *Making democracy work*. Princeton University Press.

Robbins, B. G. (2021). An empirical comparison of four generalized trust scales: Test–Retest reliability, measurement invariance, predictive validity, and replicability. *Sociol. Methods Res.*, 00491241211055765.

Robbins, B. G. (2022). Measuring generalized trust: Two new approaches. *Sociol. Methods Res.*, *51*(1), 305–356.

Robbins, B. G. (2023). Valid and reliable measures of generalized trust: Evidence from a nationally representative survey and behavioral experiment. *Socius*, *9*.

Rosenberg, M. (1956). Misanthropy and political ideology. *Am. Sociol. Rev.*, *21*(6), 690–695.

Schilke, O., Reimann, M., & Cook, K. S. (2021). Trust in social relations. *Annu. Rev. Sociol.*, *47*(1), 239–259.

Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods*, *10*, 143–152.

Smith, S. S. (2010). Race and trust. *Annu. Rev. Sociol.*, *36*(1), 453–475.

Sønderskov, K. M. (2011). Explaining large-n cooperation: Generalized social trust and the social exchange heuristic. *Ration. Soc.*, *23*(1), 51–74.

Stolle, D. (2015). Trusting strangers – the concept of generalized trust in perspective. *Österr. Z. Polit.*, *31*(4), 397–412.

Sturgis, P., Brunton-Smith, I., & Jackson, J. (2019). Regression-Based response probing for assessing the validity of survey questions. In Paul Beatty, Debbie Collins, Lyn Kaye, Jose Luis Padilla, Gordon Willis, Amanda Wilmot (Ed.), *Advances in questionnaire design, development, evaluation and testing* (pp. 573–591). unknown.

Sturgis, P., & Smith, P. (2010). Assessing the validity of generalized trust questions: What kind of trust are we measuring? *Int J Public Opin Res*, *22*(1), 74–92.

Sztompka, P. (1999). *Trust: A sociological theory*. Cambridge University Press.

Torpe, L., & Lolle, H. (2011). Identifying social trust in Cross-Country analysis: Do we really measure the same? *Soc. Indic. Res.*, *103*(3), 481–500.

Uslaner, E. M. (2002). *The moral foundations of trust*. Cambridge University Press.

Van Deth, J. W. (2003). Measuring social capital: Orthodoxies and continuing controversies. *Int. J. Soc. Res. Methodol.*, *6*(1), 79–92.

Vollan, B. (2011). The difference between kinship and friendship: (field-) experimental evidence on trust and punishment. *J. Socio Econ.*, *40*(1), 14–25.

Wang, M.-w. (2006). Automatic text classification model based on random forest. *Journal of Shandong University(Natural Science)*.

Xu, B., Guo, X., Ye, Y., & Cheng, J. (2012). An improved random forest classifier for text categorization. *J. Comput.*, *7*(12).

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the united states and japan. *Motiv. Emot.*, *18*(2), 129–166.

Yuki, M., Maddux, W. W., Brewer, M. B., & Takemura, K. (2005). Cross-cultural differences in relationship- and group-based trust. *Pers. Soc. Psychol. Bull.*, *31*(1), 48–62.

A.1 Summary statistics

Below we provide summary statistics for our sample. Our main, long-format dataset has 10,500 rows because we repeatedly observe our 1,500 respondents across 7 trust measures (1,500*7).

Table A2.1 provides summary statistics for our trust measures which have been standardized to range from 0 to 1. Unique (#) describes the number of unique values the variable assumes (including the missing category "NA"). Missing (%) describes the percentage of missing values on that variable.[28] The corresponding means are also displayed in Figure 2.2 (cf. Full dataset).

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
M1: Trust most people (std.)	8	0	0.42	0.27	0.00	0.50	1.00
M2: Trust people first time (std.)	5	2	0.40	0.26	0.00	0.33	1.00
M3: Trust stranger (std.)	5	2	0.33	0.27	0.00	0.33	1.00
M4.1: Trust stranger secret (std.)	5	1	0.14	0.23	0.00	0.00	1.00
M4.1: Trust stranger loan (std.)	5	1	0.13	0.22	0.00	0.00	1.00
M4.3: Trust stranger child (std.)	5	1	0.13	0.24	0.00	0.00	1.00
M4.4: Trust stranger advice (std.)	5	2	0.30	0.26	0.00	0.33	1.00

Table A2.1: Summary statistics across (standardized) trust scales.

Table A2.2 and Table A2.3 present summary statistics for numeric and categorical variables (excluding trust measures), along with population estimates where applicable. For the socio-demographic variables, which remain constant across

[28] The difference in missing values for M1 (n=1) and M2 (n=23), as well as M1 (n=1) and M3 (n=34) is statistically significant (p<0.001).

our various trust measures, we utilized the first slice of our long-format dataset, encompassing all 1,500 respondents, to generate these statistics.

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
Socio-economic status (numeric)	11	9	5.36	1.70	1.00	5.00	10.00
Income (numeric)	13	17	3.97	2.92	1.00	3.00	12.00
Education (numeric)	8	21	4.48	1.53	1.00	5.00	7.00

Table A2.2: Summary statistics: Numeric covariates.

		N	%	N (U.S. Census)	% (U.S. Census)
Age (factor)	18-27	296	19.7	43 355 638	17.9
	28-37	278	18.5	42 085 420	17.4
	38-47	248	16.5	39 974 287	16.5
	48-57	258	17.2	43 370 543	17.9
	58-80+	420	28.0	73 462 149	30.3
Sex (factor)	Female	768	51.2	125 196 929	51.7
	Male	731	48.7	117 051 108	48.3
Ethnicity (factor)	Asian	95	6.3	14 040 646	5.8
	Black	197	13.1	30 097 066	12.4
	Mixed	38	2.5	3 893 117	1.6
	Other	31	2.1	3 601 403	1.5
	White	1138	75.9	190 615 805	78.7

Table A2.3: Summary statistics: Categorical covariates.

A.2 Question wording

Table A2.4 outlines the wording of our different survey measures.

Measure	Question wording	Response scale recoding
M1: Most people	Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people? Please tell me on a score of 0 to 6, where 0 means you can't be too careful and 6 means that most people can be trusted.	Original scale: 0 - You can't be too careful; 1; 2; 3; 4; 5; 6 - Most people can be trusted; Don't know; Recoded scale: Don't know = NA and values 0-6 standardized to 0-1.
M2: People first time	How much do you trust people you meet for the first time?	Original scale: Do not trust at all; Trust not very much: Trust somewhat; Trust completely; Don't know; Recoded scale: Don't know = NA and values 1-4 standardized to 0-1.
M4.1: Keep a secret	...keep a secret that is damaging to your reputation?	See above.
M4.2: Repay a loan	...repay a loan of one thousand dollars?	See above.
M4.3: Look after child	...look after a child, family member, or loved one while you are away?	See above.
M4.4: Money advice	...provide advice about how best to manage your money?	See above.

Continued on next page

(continued)

Measure	Question wording	Response scale recoding
M4.1-M4.4: Probe	In answering the previous question, who came to your mind when you were thinking about a 'person you meet for the first time'/ 'total stranger you meet for the first time'? Please describe.	open textbox
Age (factor)	What is your current age in years?	Original scale: Simple numeric entry; Recoded scale: Recoded to factor with four levels (1) 17-29, (2) 30-43, (3) 44-59 and (4) 59-93.
Sex (factor)	What sex were you assigned at birth, such as on an original birth certificate?	Original scale: Two answers options 'Male' and 'Female'; Recoded scale: Recoded to factor with two levels (1) Female and (2) Male.
Ethnicity (factor)	What ethnic group do you belong to?	Original scale: Five answer options 'White', 'Black', 'Asian', 'Mixed' and 'Other'; Recoded scale: Recoded to factor with corresponding levels. Reference category is 'Asian'.

Continued on next page

(continued)

Measure	Question wording	Response scale recoding
Socioeconomic status (numeric)	Think of a ladder (see image) as representing where people stand in society. At the top of the ladder are the people who are best off—those who have the most money, most education and the best jobs. At the bottom are the people who are worst off—who have the least money, least education and the worst jobs or no job. The higher up you are on this ladder, the closer you are to people at the very top and the lower you are, the closer you are to the bottom. Where would you put yourself on the ladder? Choose the number whose position best represents where you would be on this ladder.	Original scale: Ten answer options; Recoded scale: Numeric with 10 values.

Continued on next page

(continued)

Measure	**Question wording**	**Response scale recoding**
Income (numeric) (in GBP)	What is your personal income per year (after tax) in GBP? If you need to convert from another currency you can find a converter [here]	Original scale: Answer options 1 - Less than £10,000; 2 - £10,000 - £19,999; 3 - £20,000 - £29,999; 4 - £30,000 - £39,999; 5 - £40,000 - £49,999; 6 - £50,000 - £59,999; 7 - £60,000 - £69,999; 8 - £70,000 - £79,999; 9 - £80,000 - £89,999; 10 - £90,000 - £99,999; 11 - £100,000 - £149,999; 12 - More than £150,000; Rather not say; Recoded scale: Numeric with 13 values, Don't know = NA.
Education (numeric)	Which of these is the highest level of education you have completed?	Original scale: Answer options 1 - No formal qualifications; 2 - High school diploma/A-levels

Table A2.4: Question wording.

A.3 Open-ended text answers

Figure A2.1 displays the distribution of answer lengths for our 7,497 open-ended probing answers. On average, respondents used 11 words (min = 1, max = 257, sd = 13) for each probe.

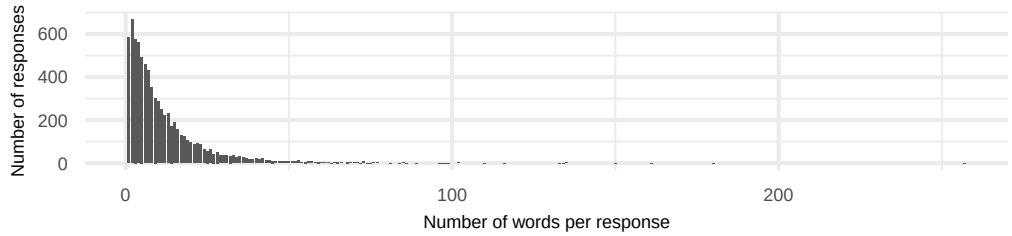

Figure A2.1: Length of open-ended responses.

Figure A2.2 displays the 15 most frequent words by probing question. Besides the overview on which words are commonly used, the side-by-side barplot also depicts which frequent words do not appear for all three measures. For instance, only among answers to the question about most people, the term "family" appears quite frequently.

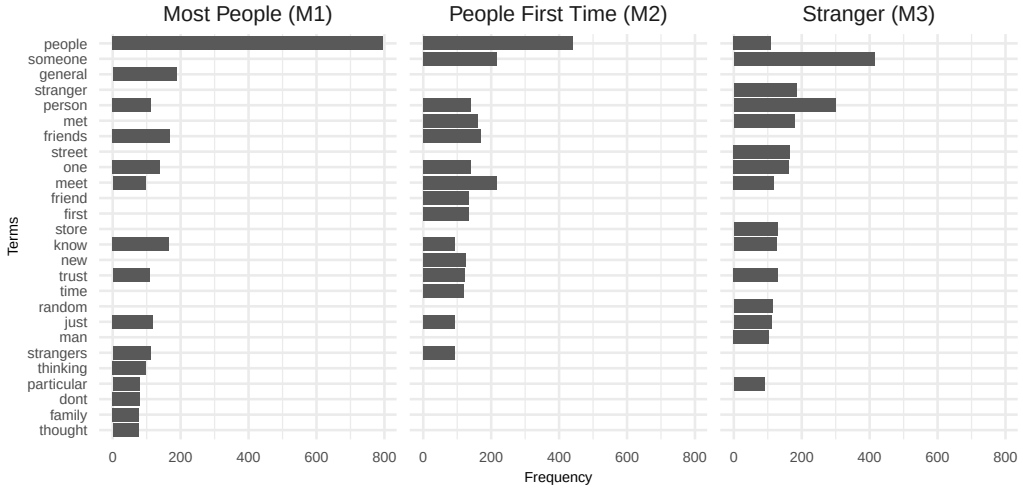

Figure A2.2: Most frequent words in open-ended answers to generalized trust questions.

A.4 Manual classification: Coding schemes

Tables A2.5 and A2.6 show descriptions and examples for the different codes. To classify associations referring to known/unknown others, documents of Code 8 were subsumed under Code 0 and documents of Code 9 under Code 1. For the sentiment analysis, Code -1 and 0 were combined into one category (0). Also, we added Code 8 to this neutral/positive category (0). Code 9 was applicable to only very few documents (3 of 1,500) and thus was excluded. These manipulations allow us to examine our hypotheses using a dichotomous classifier (negative vs. neutral/positive sentiment) while at the same time reducing complexity for the classifier.

Code	Description	Examples
1	known others: includes all mentions of persons the respondent personally knows. This also includes persons the respondent has only met once before (i.e., no stranger no more).	Everyone I know people I interacted with the sum total of all people you know and meet A person I met a week ago
0	unknown others: includes all mentions of persons the respondent does not know. This also includes descriptions of groups, where the respondent might know some of the persons, however certainly not all.	No one in particular just people as a whole a random bunch of people people that are young like myself people in my town
8	not applicable: includes all answers that do not refer to Code 1 or Code 0, including non-sense or irrelevant answers (indicators of low response quality).	no one/nothing everyone / everybody / anyone myself don't know

Continued on next page

(continued)

Code	Description	Examples
9	mixed: not always did respondents decide to only describe known or unknown others but rather made mentions of both: includes all statements that make mentions of both known and unknown persons.	those I meet in my everyday activities People I interact with on a daily basis; people at work, people at the grocery store, etc. People I may run into everyday.

Table A2.5: Coding scheme for associations (known-unknown others).

Code	Description	Examples
1	negative: includes all documents that make use of explicit negative sentiment.	Chloe, met her at the gym, asked her to help watch my stuffs while I use the restroom. When I came back, she was gone. Someone doing something behind my back that will jeopardize my well-being, my place of residence, or blame me or start stories about me that aren't true. This has happened a few times to me before.
-1	positive: includes all documents that make use of explicit positive sentiment.	I guess because I live in a city where the population is more dense, the chance of dealing with a wider spectrum of people increases. I can see most encounters would be of a kind person with good intentions, so just about anyone would and could be kind. Generally, someone that I might have contact with for the first time and might not ever have contact with again. Someone stopping to give help on the side of the road, for example. I just thought of general strangers and how I approach them. In general, as long as I don't need to trust them with anything in particular, I start with a little trust

Continued on next page

(continued)

Code	Description	Examples
0	neutral: includes all documents that make use of explicit neutral sentiment. Importantly, there has to be enough text to assess that some kind of sentiment is given.	No particular person came to mind. For me when first meeting someone I have to see how the conversation flows. Trust is earned. I wouldn't have any reason to trust them completely but would give them the benefit of doubt It depends on what you are trusting the individual for. In general, you would trust that the stranger means no harm to you.
8	not applicable: includes all documents in which no sentiment is mentioned or in which it is unclear which sentiment is being associated. This also means that documents that make only implicit (some kind of interpretation is needed) use of sentiment. Also, all documents that do not make use of the previous codes, including non-sense answers. These documents could be too short to make an assessment.	myself don't know friends/family/coworker OMG

Continued on next page

(continued)

Code	Description	Examples
9	mixed: all documents that make explicit use of negative and positive sentiment.	Most people can't be trusted because people have different thoughts from one another. Some people want the other people to succeed while some people want the other people to fail or harm them By most people, I was thinking about the extremes of people between those who hold themselves to strict high, moral standards regularly, and those who live on impulse with aggression issues and mental instability.

Table A2.6: Coding scheme for associations (sentiment).

A.5 Automated classification: Evaluation

Below, we assess our automatic classification approach by comparing its results to different subsets of the data: manually coded data only and a data subset that eliminates question order effects.

Manual vs. automated classification

Figure A2.3 displays distributions of the codes by coding procedure, i.e., manual (n=1,000/1,500) and automated classification (n=6,500/6,000).

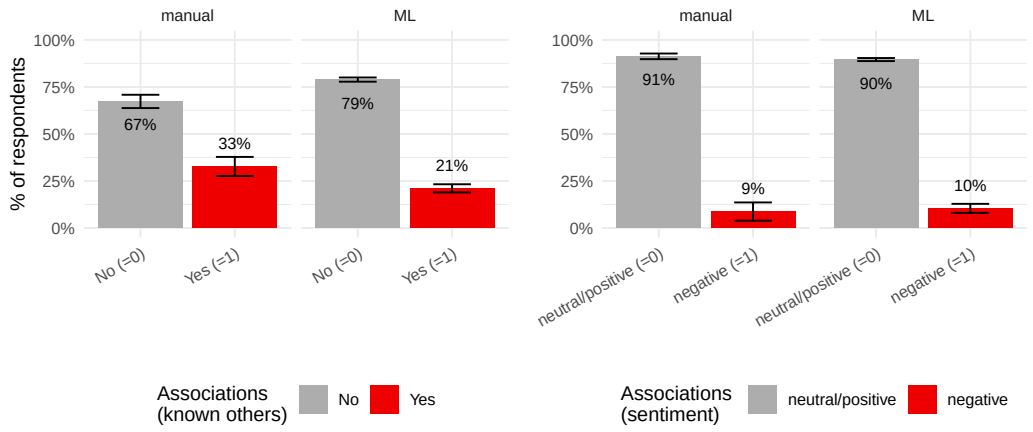

Figure A2.3: Shares of codes by coding procedure (manual vs. ML). *Note*: Error bars represent 95% confidence intervals (lower cutoff at 0).

Generally, we observe an imbalance where for both types of association the code of particular theoretical interest (known people and negative sentiment, respectively) was assigned less often than the reference code (unknown people, neutral/positive sentiment). In the case of the known–unknown classification, using the the BERT classifier results in a even smaller share of the known other code (e.g., classification error).

Subset analysis: Manual classification

To additionally examine the robustness of our main findings, Figure A2.4 shows findings for a dataset in which only our manually coded ("gold standard") data is

included (n=1,000/n=1,500). Note that for manually coding the known–unknown dimension we drew a sample of answers to M1, while for manually coding the sentiment dimension we drew a sample of answers to M1, M2 and M3 (see Methods Section). Again, both figures show that the codes of substantial interest (i.e., known people and negative experiences) appear more often in our manually coded data while maintaining the same pattern across measures as was shown in the main paper.

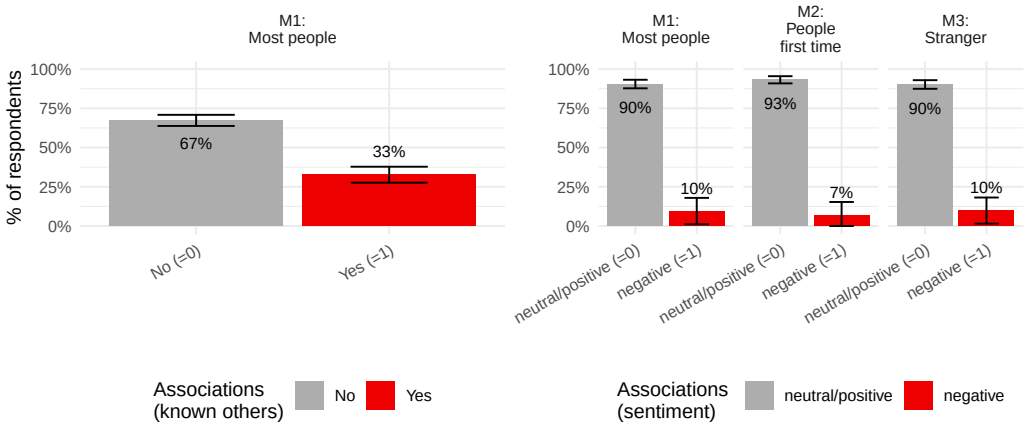

Figure A2.4: Distribution of associations with known people across trust measures. *Note*: Error bars represent 95% confidence intervals (lower cutoff at 0).

Overall, we can assume that in our main analysis we underestimated the prevalence[29] and effects[30] of associations of known people and negative experiences, which only strengthens our overall findings.

[29]The possibility that we manually coded a subset of "special" documents (e.g., relatively high share of negative experiences and known others) by chance is ruled out due to the random sampling.

[30]Regression analyses (results available upon request) using the manually coded data only (n=1,000/1,500) yield similar findings as when using the overall data (see Figure 5; Tables 12 - 18). First, mentioning known others statistically significant increases reported trust scores ($\beta = 0.089$ p<0.001; model includes covariates). Second, sentiment in the form of negative experiences statistically significantly decreases reported trust scores ($\beta = -0.11$, p<0.001; model includes covariates).

Subset analysis: Subset 2

Figure A2.5 displays the share of respondents who described associations of either known or unknown others across our seven measures but only for Subset 2, i.e., where n=1,500 and for each question data is used from only those respondents that got the respective question in the very first position of the questionnaire (details are provided in the Methods Section). Findings strongly support the findings we found in the main paper.

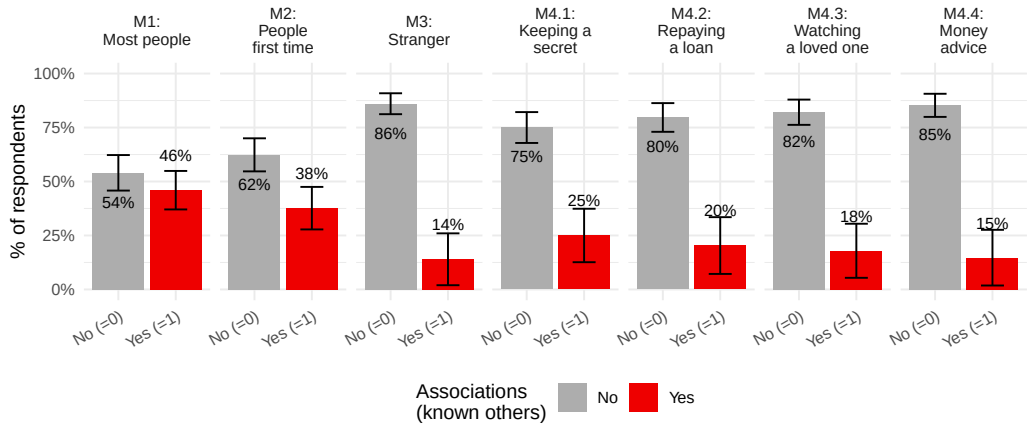

Figure A2.5: Distribution of associations with known people across trust measures. *Note*: Error bars represent 95% confidence intervals (lower cutoff at 0).

A.6 Automated classification: Alternatives

Below are alternative approaches we utilized to classify the content (known–unknown) and sentiment of the open-ended answers.

Classification of associations (known–unknown others) with regular expressions

To identify responses that mentioned known others, we additionally automatically detected open-ended responses that contained the following terms: friend, family, coworker, co-worker, neighbor, relative, boyfriend, girlfriend, husband, wife, father, mother, sister, brother. Figure A2.6 shows findings for the known–unknown categories across our seven trust measures. The emerging pattern mimics the one from our main analysis (cf. Figure 2.3).

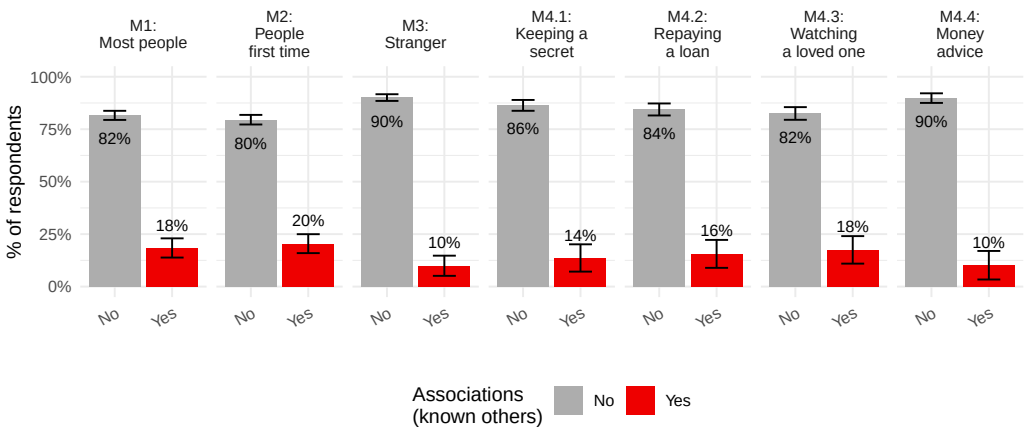

Figure A2.6: Distribution of associations with known people across trust measures. *Note*: Error bars represent 95% confidence intervals (lower cutoff at 0).

Classification of associations (known–unknown others) and sentiment using random forests

Random forest and document-term matrix

Furthermore, we trained two random forest classifiers. To make the text data processable to this machine learning algorithm, we first transformed it into numerical

data via tokenization, where each unigram (i.e., each unique term used in the open-ended text answer) is one-hot encoded into a separate variable indicating whether or not the respective document (i.e., a certain text answer) contains the unigram of interest. These binary indicators are stored in a Document-Term Matrix, short DTM.[31] A glimpse into this representation of text data is given in Table A2.7.

ID	**Probing Answer**	**Associations (known-unknown)**	**Associations (senti-ment)**	**dont**	**know**	**think**	**littl**	**tourist**	**..**
123	I was thinking of people I don't know personally.	0 (No)	0 (neu-tral/positive)	1	1	1	0	0	..
3139	Tourists that come to our little village. I tend to be very wary of them.	0 (No)	1 (negative)	0	0	0	1	1	..
7214	My friends back in high school.	1 (Yes)	0 (neu-tral/positive)	0	0	0	0	0	..
..	..	..	..	..	..	..	..	..	..

Table A2.7: Illustration of exemplary document-term matrix.

We pursued several common steps in pre-processing text data including stemming, transformation to lowercase, removal of punctuation, numbers and common stop-words (e.g., Kathuria et al., 2021). Also, before we started training the classifier,

[31] In text mining, a DTM is a specific type of a matrix used to represent the frequencies of terms in documents. Typically, a DTM will have m rows and n columns, where m represents the total number of documents and n represents the total number of terms. Each entry aij contains the frequency with which term i occurs in document j (Anandarajan et al., 2019).

we removed rare terms and only kept terms that appear in at least 0.5% of the documents. Random forests are commonly used for classifying text because they are algorithmically simple and at the same time provide high levels of performance even for multidimensional data (e.g., Wang, 2006; Xu et al., 2012). Briefly, the intuition of a random forest classifier is that a large number of simple decision trees (here 500) are fitted to the data. This is achieved through bootstrapping, where new training datasets are created by random sampling from the original data with replacement. Each decision tree is grown using random feature selection.[32] Importantly, sampling with replacement (i.e., bootstrapping) ensures that approximately one-third of the documents will be out-of-bag (OOB) data (Breiman, 2001, p.11). This OOB data serves as a built-in validation set, eliminating the need for additional splitting of the data into test and training sets.

The task of classifying new data is done by bagging methods. More explicitly, each new datapoint d (i.e., document) is passed down each tree following the logic of a simple decision tree. Results from doing this for all trees are aggregated and d is assigned its prediction via majority vote.

Using the above representation of data, we trained two classifiers. For evaluation, the OOB error rate (averaged over all boostrapped datasets) provides an unbiased measure of accuracy (Breiman, 2001, p.11). Classifying the known–unknown dimension achieves an overall OOB accuracy rate of 0.83. The classifier for sentiment achieves an overall OOB accuracy rate of 0.92.

Table A2.8 shows a glimpse into exemplary documents that were classified with the Random Forest classifiers (cf. Table 2.2 in the main paper).

[32] In one of its most popular variants (Breiman, 2001), the single trees in the forest are constructed by randomly selecting a subspace of features (e.g., 2) at each node of a tree to grow further branches. For clarification, features in the case of text data are terms (see Table A2.7).

ID	Measure	Trust	Probe	Associations (known-unknown others)	Associations (sentiment)
123	Most people	0.33	I was thinking of people I don't know personally.	0 (No)	0 (neutral/positive)
3139	Most people	0.17	Tourists that come to our little village. I tend to be very wary of them.	0 (No)	1 (negative)
7214	People first time	0.33	My friends back in high school.	1 (Yes)	0 (neutral/positive)
7304	People first time	0.67	No specific person	0 (No)	0 (neutral/positive)
1365	Stranger	0.67	A person sitting next to me at a game	0 (No)	0 (neutral/positive)
2980	Stranger	0	No one in particular, but I don't think I could trust anyone ever again.	0 (No)	1 (negative)
1289	Keeping a secret	0	An anonymous, faceless man was my first thought, perhaps someone in a train or bus station.	0 (No)	0 (neutral/positive)
1487	Repaying a loan	0	White man, about 60, good looking, widower	0 (No)	0 (neutral/positive)
1756	Watching a loved one	0	A friend named Cecil, I don't trust anybody after I Immediately meet them and they have to not do anything horrible to earn my trust.	1 (Yes)	0 (neutral/positive)
1	Money advice	0	Just a random stranger.	0 (No)	0 (neutral/positive)
...	...	...	...	...	...

Table A2.8: Illustration of exemplary data.

Figure A2.7 shows findings for the share of known and unknown others for our seven measures. The emerging pattern mimics the one from our main analysis (cf. Figure 2.3).

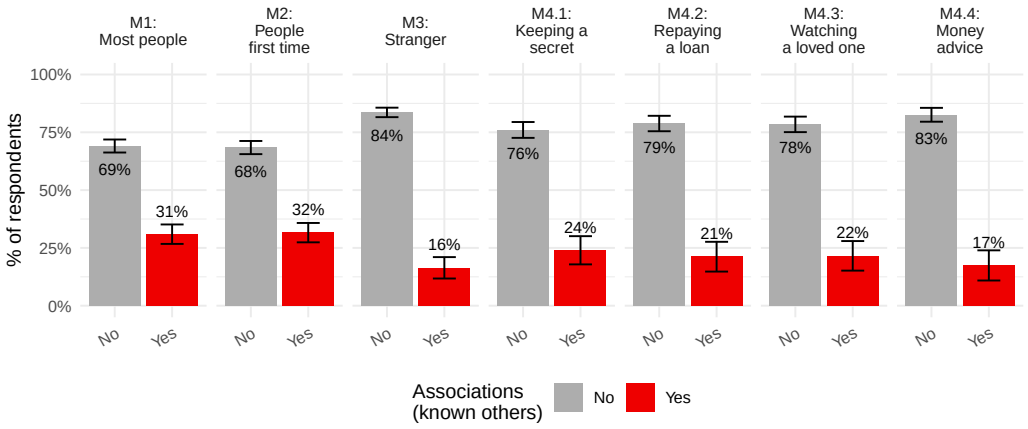

Figure A2.7: Distribution of associations with known people across trust measures. *Note*: Error bars represent 95% confidence intervals (lower cutoff at 0). Data is the full dataset irrespective of the question orblock randomization (details are provided in the Methods Section).

Figure A2.8 shows findings for the share of negative and neutral/positive for our seven measures. The emerging pattern mimics the one from our main analysis (cf. Figure 2.4).

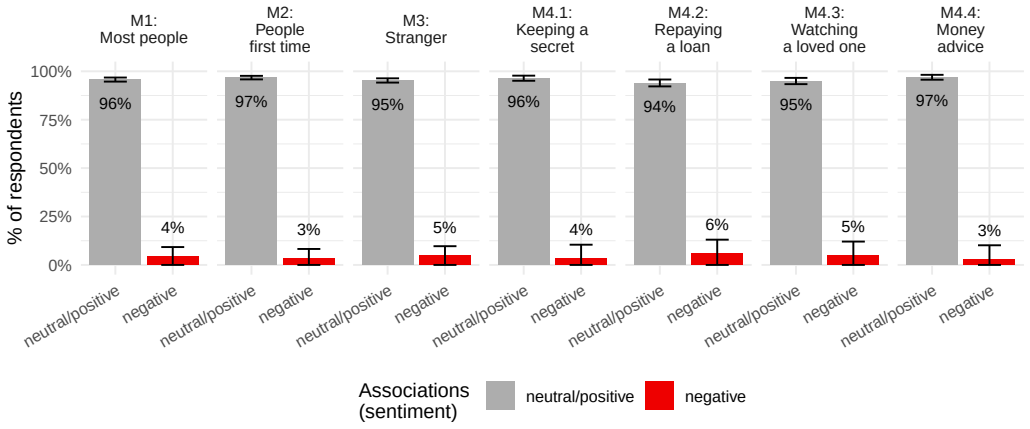

Figure A2.8: Distribution of associations with known people across trust measures. *Note*: Error bars represent 95% confidence intervals (lower cutoff at 0). Data is the full dataset irrespective of the question orblock randomization (details are provided in the Methods Section).

Figure A2.9 shows findings for the regression analysis of associations on trust scores. The emerging pattern mimics the one from our main analysis (cf. Figure 2.5).

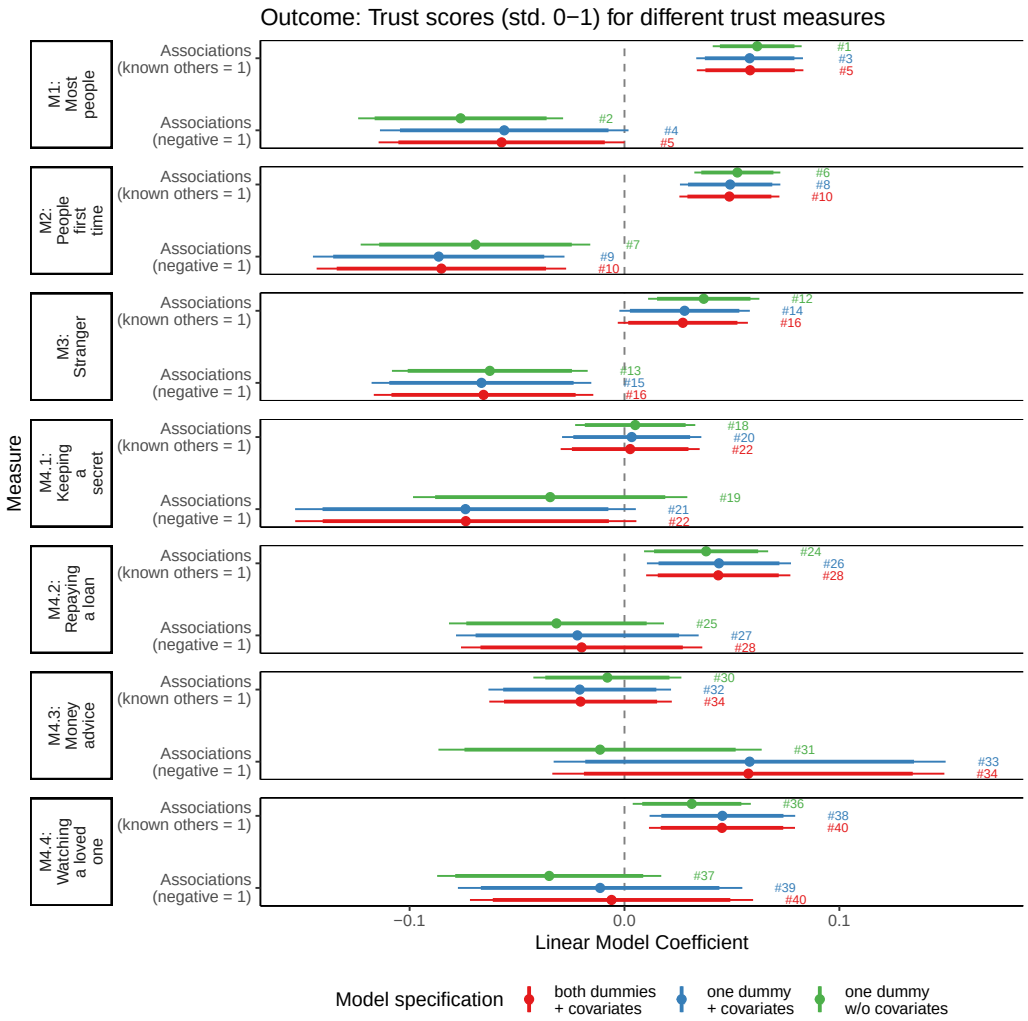

Figure A2.9: Distribution of associations with known people across trust measures. *Note*: The figure shows point estimates for coefficients of our dummy variables of interest, namely having associations with known others or negative associations. Bars represent 90% (thicker) and 95% (thinner) confidence intervals. Data include the full dataset, irrespective of the question or block randomization (details are provided in the Methods Section).

A.7 Systematic bias of trust scores

As mentioned in the conclusion, we were interested in examining whether associations differ according to respondents' characteristics, e.g., their education, income, etc. Figure A2.10 displays pearsons r correlation coefficients for a set of potentially interesting variables and our two binary association variables.

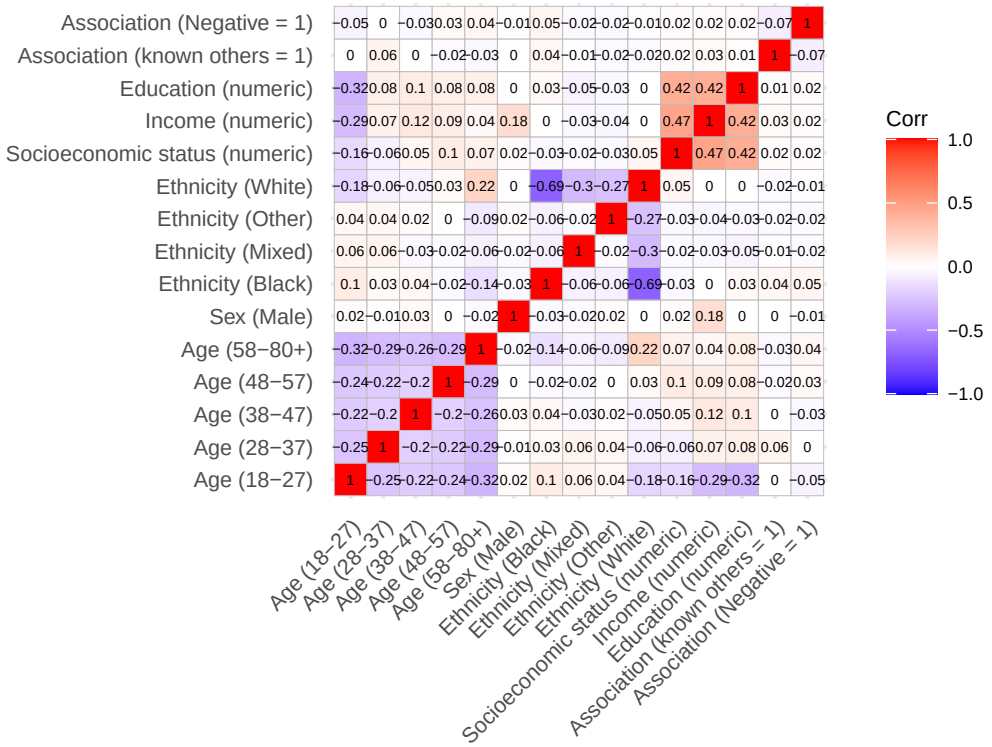

Figure A2.10: Correlation matrix: Socio-demographic variables and association dummies. *Note*: The figure shows point estimates for coefficients of our dummy variables of interest namely having associationswith known others or negative associations. Bars represent 90% (thicker) and 95% (thinner) confidence intervals. Data is the full dataset irrespective of the question or block randomization (details are provided in the Methods Section).

Overall, we find that none of the sociodemographic variables have a meaningful correlation with our binary association measures. While this analysis is prelimi-

nary, not finding any systematic relationship might be a good indicator: although sentiment and content (known others) of associations statistically significantly influence the trust score, these associations do not emerge from specific (socio-demographic) covariates in the first place. In other words, it is unlikely that differential associations may introduce bias when we study the impact of different socio-demographics on trust scores.

A.8 Cross-situational trust

Figure A2.11 depicts a correlation matrix for our trust measures. As we might have expected, all the different trust measures correlate positively. At the same time, these correlations do not seem high enough to argue that the different trust measures tap into one single concept. One possible approach would then be to take an average across the different situational trust measures to obtain an estimate of cross-situational trust (Bauer & Freitag, 2018).

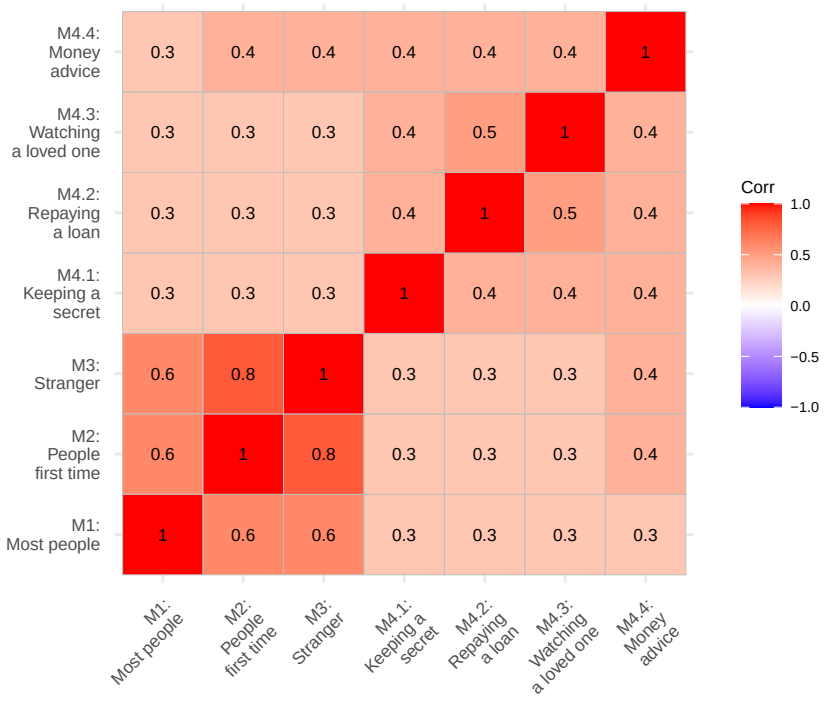

Figure A2.11: Correlation matrix: trust measures.

A.9 Regression models

Belowe we show regression model estimates (cf. Figure 2.5) in the main paper.

	M1: Most people				
	#1	#2	#3	#4	#5
(Intercept)	0.436***	0.395***	0.248***	0.191***	0.225***
	(0.007)	(0.012)	(0.047)	(0.049)	(0.048)
Associations (known others = 1)	0.064***		0.064***		0.064***
	(0.011)		(0.013)		(0.013)
Associations (negative = 1)		-0.041*		-0.038+	-0.035+
		(0.016)		(0.020)	(0.019)
Age (28-37)			-0.014	-0.008	-0.012
			(0.028)	(0.028)	(0.028)
Age (38-47)			0.011	0.009	0.012
			(0.029)	(0.029)	(0.029)
Age (48-57)			-0.002	-0.003	0.000
			(0.028)	(0.028)	(0.028)
Age (58-80+)			0.085***	0.085**	0.087***
			(0.025)	(0.026)	(0.025)
Sex (Male)			0.028+	0.028+	0.027
			(0.017)	(0.017)	(0.017)
Ethnicity (Black)			-0.060	-0.048	-0.058
			(0.041)	(0.042)	(0.041)
Ethnicity (Mixed)			-0.092	-0.093	-0.094
			(0.062)	(0.063)	(0.062)
Ethnicity (Other)			-0.057	-0.049	-0.059
			(0.066)	(0.066)	(0.066)
Ethnicity (White)			0.021	0.030	0.021
			(0.036)	(0.036)	(0.036)
Socioeconomic status (numeric)			0.015**	0.016**	0.015**
			(0.006)	(0.006)	(0.006)
Income (numeric)			0.003	0.002	0.002
			(0.003)	(0.003)	(0.003)
Education (numeric)			0.011+	0.011+	0.011+
			(0.006)	(0.006)	(0.006)
Num.Obs.	1499	1498	985	985	985
R2	0.024	0.004	0.099	0.079	0.102
RMSE	0.27	0.27	0.25	0.26	0.25

Table A2.9: Linear regression of trust scores (Y) on associations (Xs). *Note*: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

	M2: People first time				
	#6	#7	#8	#9	#10
(Intercept)	0.410***	0.360***	0.223***	0.183***	0.192***
	(0.007)	(0.013)	(0.045)	(0.046)	(0.046)
Associations (known others = 1)	0.037***		0.041***		0.039**
	(0.010)		(0.012)		(0.012)
Associations (negative = 1)		-0.066***		-0.057**	-0.053*
		(0.019)		(0.022)	(0.022)
Age (28-37)			-0.036	-0.035	-0.037
			(0.027)	(0.027)	(0.026)
Age (38-47)			-0.018	-0.023	-0.019
			(0.028)	(0.028)	(0.028)
Age (48-57)			0.004	-0.002	0.004
			(0.027)	(0.027)	(0.027)
Age (58-80+)			0.083***	0.078**	0.084***
			(0.025)	(0.025)	(0.025)
Sex (Male)			0.016	0.012	0.014
			(0.016)	(0.016)	(0.016)
Ethnicity (Black)			-0.106**	-0.100*	-0.102**
			(0.040)	(0.040)	(0.040)
Ethnicity (Mixed)			-0.062	-0.067	-0.066
			(0.059)	(0.059)	(0.059)
Ethnicity (Other)			0.014	0.000	0.010
			(0.063)	(0.063)	(0.063)
Ethnicity (White)			0.011	0.013	0.013
			(0.034)	(0.034)	(0.034)
Socioeconomic status (numeric)			0.013*	0.013*	0.013*
			(0.006)	(0.006)	(0.006)
Income (numeric)			0.000	0.001	0.000
			(0.003)	(0.003)	(0.003)
Education (numeric)			0.022***	0.020***	0.021***
			(0.006)	(0.006)	(0.006)
Num.Obs.	1476	1476	973	973	973
R2	0.009	0.008	0.104	0.100	0.110
RMSE	0.26	0.26	0.24	0.24	0.24

Table A2.10: Linear regression of trust scores (Y) on associations (Xs). *Note*: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

	M3: Stranger				
	#12	#13	#14	#15	#16
(Intercept)	0.343***	0.303***	0.153**	0.115*	0.128**
	(0.010)	(0.012)	(0.047)	(0.047)	(0.049)
Associations (known others = 1)	0.023		0.020		0.018
	(0.015)		(0.018)		(0.018)
Associations (negative = 1)		-0.049**		-0.039+	-0.037+
		(0.017)		(0.020)	(0.021)
Age (28-37)			-0.008	-0.004	-0.006
			(0.027)	(0.027)	(0.027)
Age (38-47)			0.014	0.014	0.014
			(0.029)	(0.029)	(0.029)
Age (48-57)			0.007	0.010	0.009
			(0.028)	(0.028)	(0.028)
Age (58-80+)			0.101***	0.105***	0.105***
			(0.025)	(0.025)	(0.025)
Sex (Male)			0.035*	0.036*	0.036*
			(0.017)	(0.017)	(0.017)
Ethnicity (Black)			-0.088*	-0.080+	-0.083*
			(0.041)	(0.041)	(0.041)
Ethnicity (Mixed)			-0.066	-0.068	-0.069
			(0.061)	(0.061)	(0.061)
Ethnicity (Other)			0.030	0.031	0.030
			(0.065)	(0.065)	(0.065)
Ethnicity (White)			0.027	0.028	0.027
			(0.035)	(0.035)	(0.035)
Socioeconomic status (numeric)			0.014*	0.014*	0.014*
			(0.006)	(0.006)	(0.006)
Income (numeric)			-0.001	-0.001	-0.001
			(0.003)	(0.003)	(0.003)
Education (numeric)			0.014*	0.014*	0.014*
			(0.006)	(0.006)	(0.006)
Num.Obs.	1466	1466	968	968	968
R2	0.002	0.005	0.089	0.091	0.092
RMSE	0.27	0.27	0.25	0.25	0.25

Table A2.11: Linear regression of trust scores (Y) on associations (Xs). *Note*: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

	M4.1: Keeping a secret				
	#18	#19	#20	#21	#22
(Intercept)	0.151***	0.123***	0.029	0.002	0.011
	(0.010)	(0.013)	(0.055)	(0.058)	(0.058)
Associations (known others = 1)	0.023		0.020		0.018
	(0.014)		(0.017)		(0.017)
Associations (negative = 1)		-0.036+		-0.023	-0.020
		(0.019)		(0.022)	(0.023)
Age (28-37)			-0.034	-0.031	-0.032
			(0.033)	(0.033)	(0.033)
Age (38-47)			-0.023	-0.019	-0.021
			(0.035)	(0.035)	(0.035)
Age (48-57)			-0.058+	-0.054	-0.055
			(0.034)	(0.034)	(0.034)
Age (58-80+)			-0.022	-0.019	-0.019
			(0.031)	(0.031)	(0.031)
Sex (Male)			0.009	0.011	0.011
			(0.020)	(0.021)	(0.021)
Ethnicity (Black)			0.057	0.062	0.061
			(0.048)	(0.048)	(0.048)
Ethnicity (Mixed)			-0.010	-0.011	-0.008
			(0.069)	(0.069)	(0.069)
Ethnicity (Other)			-0.022	-0.029	-0.024
			(0.089)	(0.089)	(0.089)
Ethnicity (White)			0.060	0.059	0.061
			(0.042)	(0.042)	(0.042)
Socioeconomic status (numeric)			0.001	0.002	0.001
			(0.007)	(0.007)	(0.007)
Income (numeric)			0.000	0.000	0.000
			(0.004)	(0.004)	(0.004)
Education (numeric)			0.017*	0.017*	0.017*
			(0.008)	(0.008)	(0.008)
Num.Obs.	773	773	509	509	509
R2	0.003	0.005	0.028	0.027	0.029
RMSE	0.24	0.24	0.22	0.22	0.22

Table A2.12: Linear regression of trust scores (Y) on associations (Xs). *Note*: + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	M4.2: Repaying a loan				
	#24	#25	#26	#27	#28
(Intercept)	0.166***	0.123***	0.010	-0.027	0.004
	(0.011)	(0.012)	(0.054)	(0.054)	(0.055)
Associations (known others = 1)	0.053***		0.055**		0.054**
	(0.015)		(0.018)		(0.018)
Associations (negative = 1)		-0.040*		-0.017	-0.011
		(0.017)		(0.019)	(0.019)
Age (28-37)			0.022	0.029	0.023
			(0.033)	(0.033)	(0.033)
Age (38-47)			0.047	0.049	0.045
			(0.036)	(0.036)	(0.036)
Age (48-57)			0.000	0.005	0.001
			(0.033)	(0.033)	(0.033)
Age (58-80+)			0.009	0.010	0.010
			(0.030)	(0.031)	(0.030)
Sex (Male)			0.039+	0.038+	0.039+
			(0.020)	(0.020)	(0.020)
Ethnicity (Black)			-0.005	-0.003	-0.005
			(0.047)	(0.047)	(0.047)
Ethnicity (Mixed)			-0.027	-0.031	-0.026
			(0.081)	(0.081)	(0.081)
Ethnicity (Other)			0.139+	0.132+	0.139+
			(0.075)	(0.075)	(0.075)
Ethnicity (White)			0.029	0.029	0.029
			(0.040)	(0.040)	(0.040)
Socioeconomic status (numeric)			0.000	0.000	0.000
			(0.007)	(0.007)	(0.007)
Income (numeric)			0.003	0.003	0.003
			(0.004)	(0.004)	(0.004)
Education (numeric)			0.018*	0.018*	0.018*
			(0.008)	(0.008)	(0.008)
Num.Obs.	721	721	471	471	471
R2	0.016	0.008	0.070	0.053	0.071
RMSE	0.23	0.23	0.21	0.21	0.21

Table A2.13: Linear regression of trust scores (Y) on associations (Xs). *Note*: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

	M4.3: Money advice				
	#30	#31	#32	#33	#34
(Intercept)	0.313***	0.267***	0.274***	0.254***	0.271***
	(0.013)	(0.016)	(0.068)	(0.069)	(0.070)
Associations (known others = 1)	0.046*		0.028		0.028
	(0.018)		(0.022)		(0.022)
Associations (negative = 1)		-0.042+		-0.006	-0.004
		(0.022)		(0.029)	(0.029)
Age (28-37)			-0.021	-0.022	-0.021
			(0.042)	(0.042)	(0.042)
Age (38-47)			-0.008	-0.009	-0.008
			(0.042)	(0.042)	(0.042)
Age (48-57)			-0.066	-0.069+	-0.066
			(0.041)	(0.041)	(0.041)
Age (58-80+)			0.001	-0.002	0.002
			(0.037)	(0.037)	(0.037)
Sex (Male)			0.003	0.002	0.003
			(0.024)	(0.024)	(0.024)
Ethnicity (Black)			-0.019	-0.015	-0.019
			(0.063)	(0.063)	(0.063)
Ethnicity (Mixed)			-0.064	-0.066	-0.064
			(0.090)	(0.090)	(0.090)
Ethnicity (Other)			-0.102	-0.099	-0.102
			(0.093)	(0.093)	(0.093)
Ethnicity (White)			-0.036	-0.033	-0.036
			(0.054)	(0.054)	(0.054)
Socioeconomic status (numeric)			0.011	0.011	0.011
			(0.008)	(0.008)	(0.008)
Income (numeric)			-0.005	-0.005	-0.005
			(0.005)	(0.005)	(0.005)
Education (numeric)			0.008	0.009	0.008
			(0.009)	(0.009)	(0.009)
Num.Obs.	730	730	489	489	489
R2	0.009	0.005	0.024	0.020	0.024
RMSE	0.25	0.25	0.25	0.25	0.25

Table A2.14: Linear regression of trust scores (Y) on associations (Xs). *Note*: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

	M4.4: Watching a loved one				
	#36	#37	#38	#39	#40
(Intercept)	0.142***	0.102***	-0.032	-0.052	-0.039
	(0.010)	(0.012)	(0.059)	(0.061)	(0.060)
Associations (known others = 1)	0.048***		0.046**		0.046*
	(0.015)		(0.018)		(0.018)
Associations (negative = 1)		-0.036*		-0.014	-0.010
		(0.017)		(0.020)	(0.020)
Age (28-37)			0.053+	0.059+	0.054+
			(0.032)	(0.032)	(0.032)
Age (38-47)			0.031	0.031	0.031
			(0.034)	(0.034)	(0.034)
Age (48-57)			-0.016	-0.014	-0.015
			(0.032)	(0.032)	(0.032)
Age (58-80+)			0.038	0.045	0.038
			(0.031)	(0.031)	(0.031)
Sex (Male)			0.047*	0.045*	0.047*
			(0.020)	(0.020)	(0.020)
Ethnicity (Black)			0.001	-0.002	0.003
			(0.051)	(0.051)	(0.051)
Ethnicity (Mixed)			0.060	0.050	0.061
			(0.073)	(0.073)	(0.073)
Ethnicity (Other)			0.087	0.081	0.089
			(0.075)	(0.076)	(0.075)
Ethnicity (White)			0.033	0.026	0.034
			(0.045)	(0.046)	(0.045)
Socioeconomic status (numeric)			0.007	0.007	0.007
			(0.007)	(0.007)	(0.007)
Income (numeric)			0.000	0.000	-0.001
			(0.004)	(0.004)	(0.004)
Education (numeric)			0.012	0.010	0.012
			(0.008)	(0.008)	(0.008)
Num.Obs.	728	728	476	476	476
R2	0.015	0.006	0.062	0.050	0.063
RMSE	0.22	0.22	0.21	0.21	0.21

Table A2.15: Linear regression of trust scores (Y) on associations (Xs). *Note*: + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

3 Open-ended survey questions: A comparison of information content in text and audio response formats

Abstract

Open-ended survey questions are a valuable source of data in addition to closed-ended questions, but they can be challenging for respondents, leading to decreases in response quality. Our study aims at examining how survey researchers can design open-ended questions so that their answers contain a high degree of informative answers. In particular, we examine the effect of requesting respondents to answer questions via voice input compared to text input. We use a U.S. sample (n=1,500) and questions adapted from popular social survey programs. By experimentally varying the response format, we examine which format elicits answers with a higher amount of information where the latter is operationalized by answer length, the number of topics, and response entropy. Our findings show that spoken responses tend to be longer, and also slightly more informative than written responses. We also identify sociodemographic and interview context-related factors that may influence the effectiveness of audio formats in certain settings. Our study contributes to the design of open-ended survey questions and provides insights into the potential benefits of spoken responses for specific groups of respondents.

Keywords

survey experiment, open-ended questions, voice responses, automatic speech recognition, natural language processing, topic models, entropy

3.1 Introduction

The introduction of standardized, closed-ended survey questions in the social sciences dates back to the 1960s. This period, also known as the "second era" of survey research (Groves, 2011), saw a rapid growth in quantitative social sciences as researchers recognized the benefits of using standardized survey formats. These benefits included ease of comparison and analysis of the survey data. Despite different attempts on (re-)introducing[33] open-ended questions into surveys (e.g., Schuman's (1966) "Random Probes"), the use of open-ended questions has varied over time, but generally received little attention due to the challenges involved in data analysis.

Recent advances in data analysis techniques have sparked a renewed interest in open-ended questions. For example, one area where progress has been made is in the analysis of short text data, which has made it more feasible to extract insights from open-ended responses. According to Singer & Couper (2017), OEQs are particularly useful for understanding reasons for reluctance or refusal, testing methodological theories and hypotheses, encouraging truthful answers, providing an opportunity for feedback, and improving response quality. Moreover, OEQs offer non-reactive responses due to their lack of predefined closed response categories, making them a powerful tool for gathering rich and nuanced data (Iyengar, 1996). One of the most well-known examples for OEQs can be found in the American National Election Studies (ANES, e.g., 2008, 2012, 2016, 2020), which routinely asks respondents to provide the "most important problems of their country of residence". Generally, OEQs in comparison to CEQs are preferred when surveying information that is too diverse to be stored in pre-coded ways (e.g., job descriptions, most important problem) (Barth & Schmitz, 2021) or when the aim is to elicit subjective meanings, argumentations, descriptions or associations with concepts (Bauer et al., 2017; Heffington et al., 2019; Scholz & Zuell, 2012; M. Singer, 2011). Other studies use open-ended questions as part of (web) probing, a methodology derived from cognitive interviewing with the goal of collecting information on how respondents perceive and understand survey questions or single

[33] Cf. Lazarsfeld's (1935) descriptions of open-ended follow-up questions which he also calls "why questions".

expressions within questions (Willis, 2004).

Still, despite their many benefits, OEQs are often not included in survey designs. Instead, CEQs are used, even in cases where OEQs could provide more elaborated answers. Reasons for this behavior are diverse, including concerns about data analysis or protecting respondents' privacy, and importantly, the increased response burden that OEQs come along with. OEQs can make the survey response process more difficult (Tourangeau & Rasinski, 1988; Tourangeau et al., 2000), leading to reduced response quality and high non-response rates (Roberts et al., 2014). In fact, researchers have found that there is often large variation in response quality for open-ended answers (Barth & Schmitz, 2021). Therefore, the potential of OEQs can only be fully realized when the answers are substantial, interpretable, and of consistent response quality.

A key question that arises is which response format survey researchers should use to elicit open-ended questions that yield a high degree of informative answers. The present study examines the effect of asking respondents to answer by audio compared to text entry.

To date, only few studies have examined the effect of text- and audio-based response formats on responses to OEQs (Chen et al., 2022; Gavras et al., 2022; Höhne & Gavras, 2022; Revilla et al., 2020). In recent years however, the use of voice-based response options in surveys has gained momentum, owing to the widespread use of mobile devices equipped with voice input technologies and the ongoing increase in smartphone surveys (Gummer et al., 2019; Peterson et al., 2017; Revilla et al., 2016). Additionally, advancements in speech-to-text technologies have enabled researchers to transcribe audio into a textual format, facilitating further analysis (Landesvatter et al., 2023). Spoken in comparison to written answers, are assumed to facilitate the answer process in surveys since they trigger an open narration and produce more intuitive and spontaneous answers (Gavras & Höhne, 2022; Gavras et al., 2022), so called "System 1" answers (Lütters et al., 2018). Written answers on the contrary were described to be characterized by intentional and conscious answering ("System 2"). Speaking is assumed to require less effort than typing (Revilla et al., 2020) and voice input options should ideally make answering survey questions easier and quicker. Such effects can have different implications for the resulting data quality. In particular, previous

research found that spoken answers are longer than written ones (Gavras et al., 2022; Höhne & Gavras, 2022) as well as more elaborate and detailed (Lütters et al., 2018; Revilla et al., 2020). Other relationships researched in the past include the positive influence of audio response formats on completion times (Lütters et al., 2018; Revilla et al., 2020), their negative influence on response rates (Lütters et al., 2018; Revilla & Couper, 2021; Revilla et al., 2020), and their negative influence on survey evaluation (Revilla et al., 2020).

Overall our study aims at examining two research questions. The first (RQ1) concerns the information content in open-ended answers: Are there differences in information content between responses given in audio and text formats? By investigating different measures of information content, we go beyond common and previously used measures, such as response length. Second (RQ2), we aim to make a contribution to examining whether the two formats differ in their usefulness for different types of respondents: Do sociodemographic and interview context-related characteristics moderate the effect of response format on information content (e.g., response length)? By exploring these questions, our study will shed light on the strengths and limitations of both text and audio responses and their potential implications for survey research.

3.2 Previous Research and Hypotheses

Previous research has approached the comparison of audio and text-based responses to open-ended questions through two perspectives. The first perspective focuses on outcome variables that measure various aspects of the respondent's answering process and experience, such as survey evaluation (Lenzner & Höhne, 2022; Revilla et al., 2020), participation rate (Berens & Hobert, 2022), user experience (Berens & Hobert, 2022), and completion time (Lütters et al., 2018; Revilla et al., 2020). The second perspective examines outcome variables related to response content and structure, such as response length, lexical diversity or elaboration (Chen et al., 2022; Gavras et al., 2022; Lütters et al., 2018; Revilla et al., 2020), item non-response (Chen et al., 2022; Höhne & Gavras, 2022; Lütters et al., 2018; Revilla et al., 2020), and other linguistic and content characteristics (Berens & Hobert, 2022; Gavras et al., 2022).

Our study aims to examine the second perspective, i.e. the content and structure of responses, with the aim to investigate how asking respondents to answer open-ended questions by audio compared to text entry influences response length and other measures of information content (RQ1).

Various previous studies have investigated the relationship between response format and response length and generally found audio response formats to positively influence answer length. Revilla et al. (2020) ask open-ended questions on different general-purpose topics and in comparing answers from a text entry to a voice input option, they find answers in the voice condition to be longer and more elaborated. Chen et al. (2022) ask open-ended questions about study behaviors in a student sample and experimentally vary three response format groups: text, voice or optional and find longer responses for the group that only had the option to respond by voice. Gavras et al. (2022) survey a German sample with open-ended questions about political attitudes and behavior and find that answers in the voice compared to the text condition are up to 40% longer. Höhne & Gavras (2022) were able to replicate this finding for sensitive items. Berens & Hobert (2022) study a sample of undergraduate social science students at a German university asking multiple open ended questions on topics of remote teaching at their university. In terms of answer length they found that the experimental groups that were asked to answer via audio-based formats produced far longer answers. Eventually, Lütters et al. (2018) for a German sample show that audio produces longer answers (2.9-4.66 words). However, this effect was also attributed to question ordering, as it held true especially for questions asked at later survey timepoints. In the beginning, responses in the text condition were equally long as those in the audio condition. It was only later, as the survey included more open-ended requests, that respondents in the text condition became fatigued and provided shorter responses, while those in the audio condition continued to provide longer responses.

One explanation for the positive relationship between spoken answers and response length is that spoken language is inherently different from written language. Previous research suggested that audio-based response formats are able to "[humanize] the communication process" (Lenzner & Höhne, 2022), which can elicit more open and intuitive responses (Gavras & Höhne, 2022; Gavras et al., 2022). Moreover, written responses tend to be intentional, more deliberate

and structured, while spoken responses tend to be more spontaneous, intuitive and narrative (Gavras & Höhne, 2022; Gavras et al., 2022), which may affect response length. Respondents create their responses as they speak in a more conversational manner. Other reasons for differences in response lengths could be of technical nature, e.g., typing difficulties on small screens (Chen et al., 2022; Lambert & Miller, 2015; Lugtig & Toepoel, 2016a; Mavletova, 2013; Peytchev & Hill, 2010), may leading to shorter text-based responses (Lambert & Miller, 2015; Lugtig & Toepoel, 2016b; Struminskaya et al., 2015). Overall, previous research indicates that requests for spoken or written answers can elicit different response styles, which could impact response length. Specifically, we hypothesize that responses to audio-based formats will be longer than those to text-based formats:

H1: On average, responses to audio-based formats are longer than those to text-based formats.

While response length can be treated as a measure of "amount of information", researcher have begun to explore other measures. We follow these accounts. Longer answers do not always contain more information. This is especially true when we take into account the presence of discourse particles (e.g., "like," "well," and "y'know") as well as repetitive or non-informative parts of speech (e.g., "as I said"). These elements may occur more frequently in spoken language, as typing them out requires extra effort.

One possibility to analyze the amount of information irrespective of answer length is the number of topics, usually computed by unsupervised machine learning methods, like topic models. Gavras et al. (2022) finds that respondents (for five out of six OEQs) mention significantly more topics in the oral condition than in the written condition. The authors conclude "[...] open-ended questions with requests for written answers seem to prevent respondents from mentioning all relevant aspects that they may have in mind" (Gavras et al., 2022, p.16). The same finding of more topics in a voice format holds when Höhne & Gavras (2022) asked sensitive questions. Similarly, Berens & Hobert (2022) also find audio-based formats to create answers on more topics than text-based format. Revilla et al. (2020) manually code the amount of information that is provided in each valid answer.[34] They

[34]This measure includes the number of characters, use of abbreviations, elaboration of answers, number of subquestions answered, number of topics mentioned and opt-out response.

find that the number of elaborated answers[35] is significantly higher for the voice condition as compared to the text group. However, they do not find significant differences for the number of topics and conclude that voice answers are characterized by more elaboration, but not necessarily new information.
In sum, previous research has shown differences in terms of number of topics for the two formats which may also be attributed to the different response styles as discussed above (e.g. more spontaneous and conversational versus deliberate and concise). Consequently, we hypothesize that the audio format produces answers with more topics:

H2: On average, responses to audio-based formats contain a higher number of topics compared to those in text-based formats.

In addition to the number of topics, previous research has utilized response entropy as a measure of information content in open-ended survey responses. For instance, Barth & Schmitz (2021) employed information entropy as a supplementary metric alongside response length to assess the quality of open-ended answers. Consistent with our argument presented in H2, we hypothesize that the audio format yields responses with higher information content, not only in terms of the number of topics but also in response entropy:

H3: On average, responses to audio-based formats have a higher response entropy compared to those in text-based formats.

Beyond these more general between-format comparisons, previous research has not yet explored how the two response formats may differ in their effects (e.g., be particularly helpful or attractive) depending on particular respondent characteristics (i.e., within-format comparison) (RQ2). Such analyses are fruitful in the context of survey design research, because incorporating sociodemographic variables into questionnaire design can help identify relevant subgroups that may benefit from different designs (Zuell et al., 2015). While there is research regarding which group of respondents are hypothetically or even actually willing to use audio formats (e.g., Chen et al., 2022; Höhne, 2021; Lenzner & Höhne, 2022), only Revilla et al. (2020) examined whether certain subgroups were more or less successful (i.e., answering at least one of the six questions for a voice condition)

[35]Elaboration is described to be coded whenever respondents provide "additional descriptive information or explanation about a theme without introducing a new theme" (Revilla et al., 2020)

in using voice input. They found no differences in success for sociodemographic variables such as age, education, gender, and mother tongue. In a similar manner, our study aims to compare the impact of audio and text formats on the information content outcome variables for various participant groups. As previous research on such effects has been limited, we have adopted an exploratory approach and have not pre-specified any hypotheses regarding the variation of the response format effect (H1, H2, H3) across subpopulations. We are interested in examining variables that capture participants' perceptions of convenience, usefulness, and familiarity with the respective format, as well as indicators that reflect their likelihood of reluctance, fear of disclosure, or experiencing technical issues.

3.3 Methods

3.3.1 Data

The survey was fielded between December 13, 2022 and January 17, 2023. The sample was selected using quotas based on gender, age, and ethnicity in alignment with the 2015 U.S. Census Bureau population group estimates. Notably, the 58-80+ age category did not have an exact quota match due to lower participation, resulting in slight deviations across all categories (see Appendix A.2). Regarding non-response rates, the audio format had an average of 0.47% (SD = 0.07, min = 0.41)[36] respondents who provided answers to the open-ended questions and the text format had an average response rate of 0.99% (SD = 0.01, min = 0.98). Appendix A.3 shows how respondents sociodemographic variables relate to these rates.

Participants were recruited through the recruitment/payment platform Prolific (Palan & Schitter, 2018). From 34,524 eligible participants, we collected data from 1,707 participants. 246 participants were screened-out because they did not finish the survey. The final sample size comprises 1,461[37] respondents who completed the

[36]Previous research shows similarly high item non-response rates for items asked in audio formats, for example 36% item non-response in a audio condition compared to 2% in the text condition (Höhne & Gavras, 2022).

[37]Prior to data collection, a power analysis was conducted indicating that a sample size of 1,100 is sufficient to detect a treatment effect of response format on response length of at least 2.937 percentage points (an effect size, which we, in light of previous research on response length

survey and thus yields a break-off rate of 14% (American Association for Public Opinion Research (AAPOR), 2016; Callegaro & DiSogra, 2008).
The average time to complete the questionnaire was 15 minutes (Mdn = 12.3) and was compensated with an average wage of 12$/hr. Participation in our survey required respondents to use a smartphone[38] and spoken responses were transcribed using Whisper, an automatic speech recognition (ASR) system developed by OpenAI.[39]

3.3.2 Study Design

The survey started with information on its objective and a consent form. Subsequently, respondents were randomly assigned into either the text or the audio condition. In the text condition, respondents were displayed a large answer box. In the audio condition, respondents were requested to record their voice with the open source tool "SurveyVoice (SVoice)" (Höhne et al., 2021) which allows recording via the built-in microphone of smartphones, irrespective of the operating system (e.g. Android and iOS). Figure A3.3 in Appendix A.5 show both response options for an exemplary questionnaire page.
The topic of the survey was "Life, Work and Politics". Respondents received multiple, randomly ordered, question blocks. The blocks contained a total of 9 open-ended questions covering some of the most common standardized self-report measures from popular social survey programs (e.g., most important problem) as well as follow-up ("probing") questions in an open-ended response format. Namely, the items were about satisfaction with life, social trust, most important problem facing the U.S., vote decision, attitude towards the U.S. president, political trust,

differences between text and audio (e.g. Gavras et al., 2022), argue to be conservative).

[38] Among other things, this decision should help to ensure that the responses are recorded under same conditions. Revilla & Ochoa (2016) have found answers to open-ended questions to differ between PC and smartphone respondents with regards to response time written per character, number of total characters and for the use of abbreviations.

[39] The machine generated transcripts derived with Whisper achieved a 0.01-0.16pp (percentage points) improvement rate in the word error rate (WER) over a variety of other ASR algorithms (i.e., Google's NLP API, Meta's wav2vec, Nvidia's NeMo). In order to compute the WER, 100 spoken answers were randomly selected from a subset of the data (including all data collected until the 9th January 2023). Human generated transcripts were created by 2 independent coders who in the case of disagreement agreed on one final version.

decision to (not) have children in the future[40], climate evaluation and educational decisions. Each question block consists of at least one sequence of substantial and open-ended follow-up questions. For example for the question on satisfaction with life we first asked:

"All things considered, how satisfied are you with your life as a whole these days?"

Extremely satisfied - Very satisfied - Moderately satisfied - Slightly satisfied

On the next questionnaire page, the previous substantial and closed-ended question was followed by the open-ended follow-up question where we also repeated the question wording and the respondents answer, for example:

"The previous question was: 'All things considered, how satisfied are you with your life as a whole these days?'. Your answer was: 'Extremely satisfied'. In your own words, please explain why you selected this answer".

For two of the ten questions, no follow-up questions were asked, but the actual question was directly asked as an open-ended question (i.e., most important problem facing the U.S., attitude towards the U.S. president). Appendix A.1 gives an overview of all question wordings as well as a description of considerations that were made in selecting the survey items.

In order to avoid priming effects, i.e., subsequent answers might be influenced by previous questions, we used an experimental design in which the order of questions is randomized. Respondents could skip the open-ended questions but they couldn't go back to previous questions.

3.3.3 Measures and Analytical Strategy

Our study aims to measure the information content of survey responses using multiple methods. First, we calculate the answer length by excluding punctuation and counting the number of words. However, we recognize that longer answers do not always indicate higher information content (as argued by Barth and Schmitz, 2021 and Schmidt et al., 2020). In particular, we investigate how additional measures of information content can provide insights. To address this, we compute the number

[40] Originally, respondents were also probed to explain their decision to (not) have children (right before they were asked for children in the future). Unfortunately, due to a filter error made by the survey company, we cannot analyze this item in this study.

of topics and in line with previous research (Gavras et al., 2022; Höhne & Gavras, 2022) using Structural Topic Models (STM) (Roberts et al., 2014). STM is a probabilistic and mixed-membership topic model that enables the incorporation of document-level metadata (e.g., assignment to text or audio condition) to model how topics vary across subgroups. Following recommendations by the STM documentation[41], we set the maximum number of topics to be 10. In line with Gavras et al. (2022) and Höhne & Gavras (2022) we count a topic to be represented within an individual answer only if the topic is represented in at least 10% of the answers (i.e., minimum topic size = 10%) and eventually calculate the average number of topics. We use the R package stm to fit the models (Roberts et al., 2019). In addition to the number of topics, we also compute response entropy as a measure of additional or unexpected information in each response. Entropy is a concept from information theory used to capture the information content in a variable (e.g., a text). In the survey context, entropy has been recently used to capture the linguistic complexity and to examine the qualitative variation in open-ended survey responses (Barth & Schmitz, 2021). Importantly, when using entropy, the amount of information is estimated not only based on the number of types (i.e., number of unique words) and tokens (i.e., total number of words irrespective of repetitions) in a text (e.g., this is the Type-Token Ratio[42]) but also including the frequencies and distribution of all tokens. The latter provides information on how many unique words have been used in a text and how frequently and evenly these words are distributed (Shi & Lei, 2022). For example, by taking the distribution of the tokens across a document into account, we can conclude that the entropy will be higher if the words are used in a more varied or unpredictable way throughout the text. This provides a measure of the amount of unexpected or new information that a document holds. Previous studies on the lexical richness and linguistic complexity of open-ended answers have relied on the Type-Token Ratio (TTR) (Gavras et al., 2022; Höhne & Gavras, 2022). However, researchers have pointed

[41] "For short corpora focused on very specific subject matter (such as survey experiments) 3-10 topics is a useful starting range" (Roberts et al., 2014). Setting k=10 is only an approximation to the optimal number of topics, but since we are only interested in the relative difference between text and audio, this is not of further importance.

[42] Type-token Ratio (TTR) is the ratio of unique words (types) to the total number of words (tokens) in a text (Johnson, 1944) and has been used in previous comparisons of lexical richness in text and audio answers (Gavras et al., 2022; Höhne & Gavras, 2022).

out its sensitivity to text length as a major concern with this method (Daller & Xue, 2007; Richards, 1987; Shi & Lei, 2022).[43] Entropy-based measures build upon these previous approaches by considering not only the number of tokens in a text but also their frequencies and distribution. This contributes to the call for applying additional measures when comparing lexical richness in audio and text data (Gavras et al., 2022). The application of entropy-based measures in the context of text-audio comparisons remains unexplored, and this study aims to address that gap.

To compute the entropy ($H(_X)$), we use the R package quanteda.textstats (Benoit et al., 2018), which includes a function to compute Shannon's Entropy with the logarithm to base 2 (Budescu & Budescu, 2012; Shannon, 1948):

$$H(_X) = -\sum_{i=1}^{n} P(_i) \log_2 P(_i)$$

where P_i is the probability of occurrence of the i^{th} token in the text response (relative frequency) and n is the total number of unique tokens in the text responses. The entropy $H(_X)$ is a value between 0 (i.e., only one token is used throughout all text answers) and the maximum entropy possible for the given number of unique tokens in the text response (i.e., wide variety of text answers with many words used just once). Higher entropy values indicate more unique words in a text, and a more evenly distribution of these words. Put differently, a larger entropy value of a text reflects greater uncertainties and hence higher information content of the text.

Lastly we investigate how the effects of response format on response length and information content differ for respondent characteristics. Here we investigate age, education (highest educational attainment[44]), socioeconomic status (10 cate-

[43] The Type-Token Ratio (TTR) is sensitive to text length, because an increasing length of texts decreases the occurrence of new words, leading to a biased measure of lexical richness (Shi & Lei, 2022). Shi & Lei (2022) have shown that there is a complex, non-linear relationship between lexical richness and text length, with rapid increases in lexical richness observed in shorter texts. Consequently, comparing the lexical richness of a shorter text to that of a longer text using TTR may not be a legitimate comparison.

[44] Less than a high school diploma, GED, High school diploma, Associates degree (AA, AS, etc., Bachelor's degree (BA, BS, etc.), Master's degree (MA, MS, MBA, etc.), Doctorate or professional degree (PhD, MD, JD, DDS, etc.) (cf. US Census).

gories), the location of survey participation (from home, from another place), and presence of others during survey participation. We measure these variables at the end of the survey and include them as predictor variables in a Linear Probability, interacting them with the response format variable. Detailed summary statistics for all variables are provided in Appendix A.2.

3.4 Results

3.4.1 Response Format and Information Content (RQ1)

We start by investigating the response length in the two response formats. Across all items the response length significantly differs for the two response formats with a mean response length of 27.9 (SD = 12.5) words for the audio format and 22.6 (SD = 14.5) words for the text format (t(2,1167.69) = -6.9, $p < 0.001$).[45] Additionally, we examine differences for the single items and Figure 3.1 shows the average response lengths by response format. For 5 out of 9 OEQs we observe significantly longer answers for the audio than for the text format (p<.05) which provides support for our first hypothesis (*H1*). Results from the two-sample t-tests can be found in Appendix A.6. For the items with significant differences, we observe the smallest difference for the "Most Important Problem" question (abs. difference: 3 words) and the largest difference for the "Future Children" question (abs. difference: 7 words). Previous research shows similar effect sizes, e.g., Gavras et al. (2022) show a maximum increase of 40%, Lütters et al. (2018) show increases of 2.9-4.7 words, for the voice condition respectively.

[45] Unpaired two-sample t-test with an alpha level of .05.

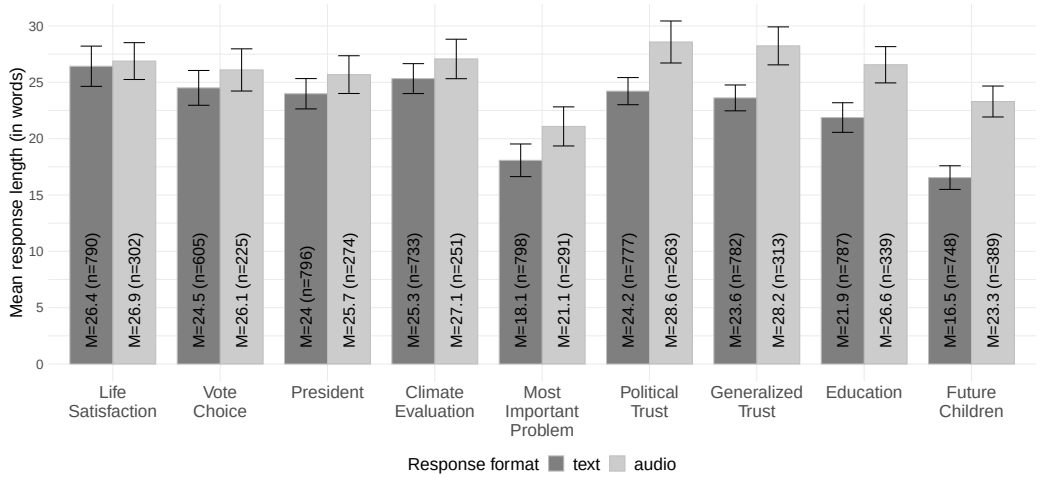

Figure 3.1: Response length by response format for 9 open-ended questions. *Note*: Error bars represent 95% confidence intervals. Results from the two-sample t-tests can be found in Appendix A.6.

Importantly, the length of an answer provides limited insights into the amount of information in a survey answer. We should use additional measures from information theory. We start with the number of topics and Figure 3.2 shows the number of topics by response format.

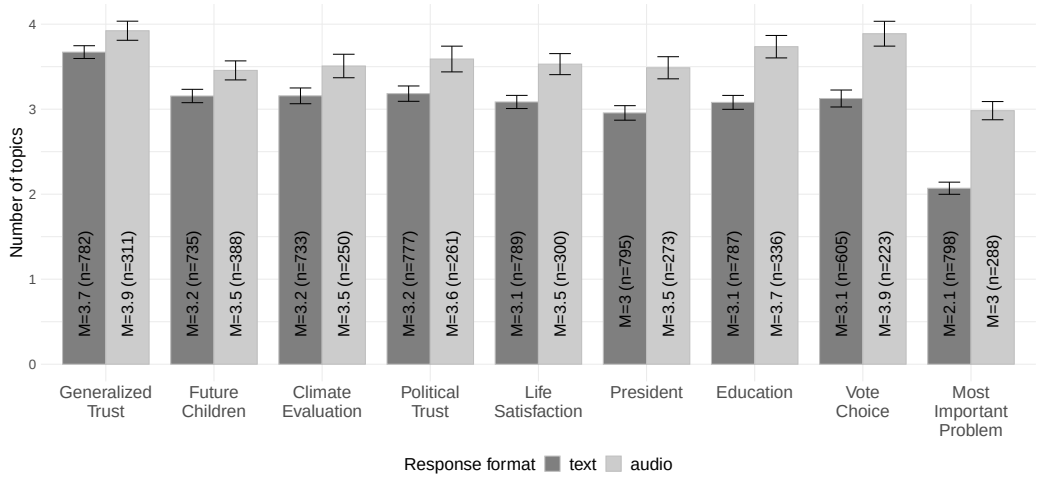

Figure 3.2: Number of topics by response format for 9 open-ended questions. *Note*: Error bars represent 95% confidence intervals. Results from the two-sample t-tests can be found in Appendix A.6.

Figure 3.2 displays the results of a structural topic model, revealing a consistent pattern of higher numbers of topics generated for the voice format compared to the text response format for 9 out of 9 OEQs ($p<.05$). Results from the two-sample t-tests can be found in Appendix A.6. This finding supports our second hypothesis (*H2*). We observe the smallest difference for the "Generalized Trust" question (abs. difference: 0.3 topics) and the largest difference for the "Most Important Problem" question (abs. difference: 1 topic). Previous research showed audio answers to contain up to 0.5 more topics than text answers (e.g., Gavras et al., 2022).

Additionally we examine the entropy of answers. Figure 3.3 shows that differences in entropy between the two response formats can be considered small but for 9 out of 9 OEQs entropy is higher in the audio format ($p<.05$). This finding is in line with previous research on lexical richness in spoken answers (e.g., Gavras et al., 2022) and lends support to our third hypothesis (*H3*).

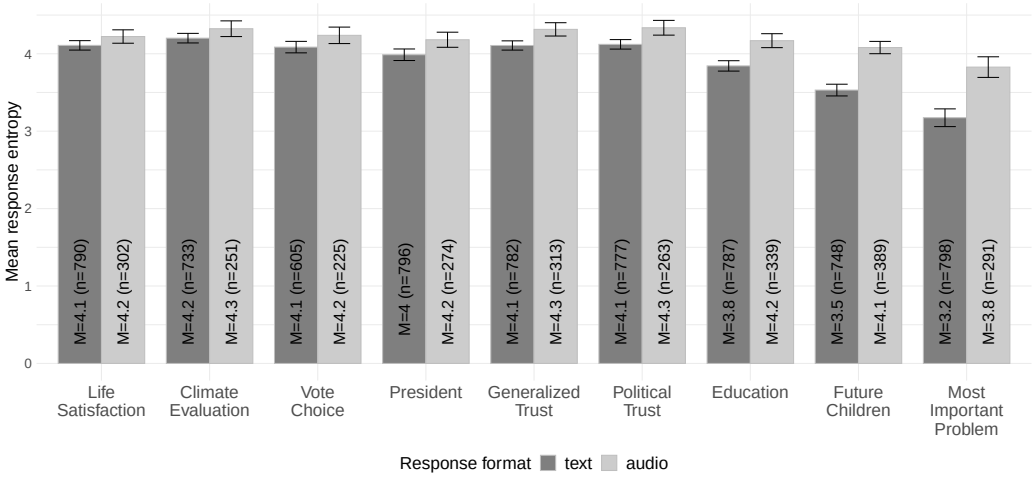

Figure 3.3: Response entropy by response format for 9 open-ended questions. *Note*: Error bars represent 95% confidence intervals. Results from the two-sample t-tests can be found in Appendix A.6.

3.4.2 Respondent and Interview Context-Related Characteristics (RQ2)

Lastly, we examine whether there are differences within the two formats for different respondent groups. In particular, in Figure 3.4 we investigate whether respon-

dent's sociodemographics and interview context-related characteristics are related to the findings for response length.[46]

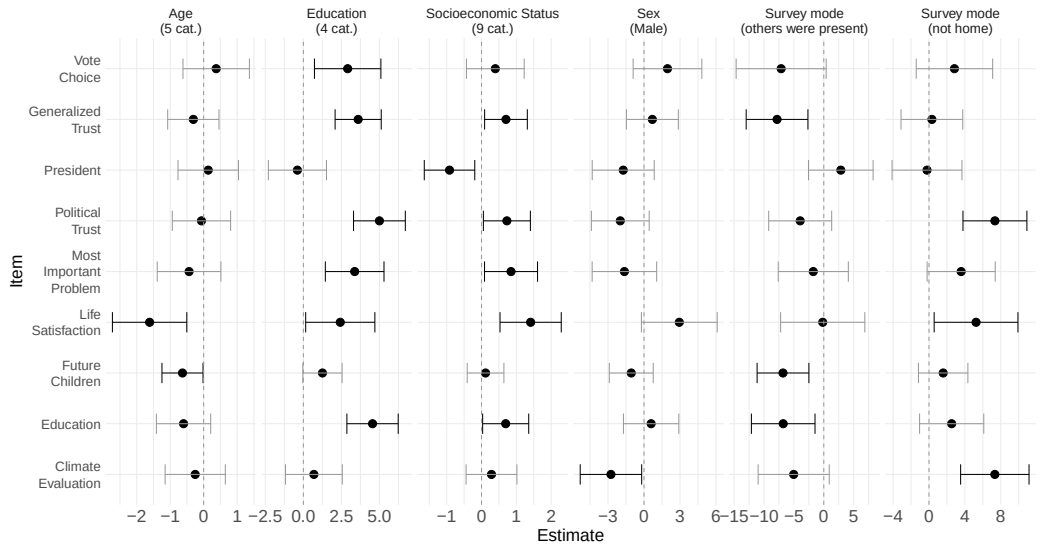

Figure 3.4: Answer length for respondent subgroups by item (audio vs. text). *Note*: Linear probability models of self-reported respondent characteristics on probability to answer open-ended survey questions comparing audio (audio=1) to text. Predictor Variables: Age (18-27, 28-37, 38-47, 48-57, 58-80+), Education (less than a high school diploma, high school diploma (+ GED), Bachelor's degree, Advanced Degree); Socioeconomic Status (1-9), Sex (female, male), Survey mode (alone, others were present), Survey mode (home, another location), Voice Input Daily Life (Never, 2, 3, 4, Daily), Survey Evaluation (Very difficult, Rather difficult, Rather easy, Very easy).

A respondent's age and gender on average do not seem to influence the answer length for spoken answers. Put differently, while audio formats positively influence the answer length (cf. Figure 3.1), the effect does not seem to be moderated by age and gender (across most OEQs).

The educational attainment and the socioeconomic status of a respondent on average have a positive influence on the answer length in audio formats compared to text formats.

Respondents who completed the questionnaire in solitude showed on average

[46] In addition to the effects on response length, we also analyzed the effect on entropy and the number of topics. The effects are in similar directions and can be found in Appendix A.4.

longer response lengths when using an audio-based response format. Additionally, respondents who answered the questions outside their homes, presumably while on the go, generally experienced the advantage of audio formats with respect to longer response lengths. This observation suggests that these respondents provided their answers in an environment that emulates contemporary voice-messaging patterns, potentially influencing their response behavior. Overall, our analysis suggests that certain independent variables have a significant impact on the effects of response format and information content, thus providing support for our fourth hypothesis (H4).

3.5 Discussion and Conclusion

Open-ended questions are a valuable source of information for survey researchers, providing insights into attitudes and opinions that cannot be captured by closed-ended questions alone. Naturally, the question arises as to how survey researchers could design open-ended questions so that their answers contain a high degree of informative answers. The present study examined the effect of asking respondents to answer by audio compared to text format on information content in responses to open-ended questions. Prior studies have emphasized the value of using voice-based formats in standardized web surveys to capture in-depth and nuanced information on respondents' attitudes, behaviors, and beliefs through open, more elaborated and detailed narrations (Lenzner & Höhne, 2022; Lütters et al., 2018; Revilla et al., 2020).

We conducted a smartphone survey among a U.S. quota-based sample (N=1,500) and asked some of the most common standardized self-report measures from popular social survey programs (e.g., most important problem) as well as follow-up ("probing") questions in an open-ended response format. We randomly assigned respondents to either give their answer in written or in spoken format.

We pursued two research questions: RQ1 asks whether there are differences in information content between responses given in audio and text responses. RQ2 asks whether sociodemographic and interview context-related characteristics are related to the effect of response format and information content (e.g., response length). To examine RQ1, we first analyzed the relationship between response

format and answer length. Previous research has shown that spoken answers tend to be longer than written ones (Berens & Hobert, 2022; Chen et al., 2022; Gavras et al., 2022; Höhne & Gavras, 2022; Revilla & Couper, 2021). Our results support these findings and indicate that requests for voice input for 5 out of 9 of our questions lead to significantly longer answers (cf. Figure 3.1). Next, we applied a topic model as well as response entropy analyses to gain further insights into the information content of the text answers. Both information content measures (i.e., number of topics and response entropy) significantly differed for the two response formats and indicate that audio formats can have a higher amount of information (cf. Figure 3.2 and Figure 3.3) (cf. Gavras et al., 2022 and Höhne and Gavras, 2022 for similar findings regarding topic models).

For our second research question, we conducted linear regression models which included interactions of response format (text/audio) and the covariates of interest (e.g., education, gender) (cf. Figure 3.4). We found an effect for a respondent's education and socioeconomic status on the response length in voice formats, indicating that higher educated respondents benefit from audio in terms of answer length. While we know from previous research (e.g., Zuell et al., 2015) that higher educated respondents provide longer answers to open-ended questions (i.e., lower burden for the formulation of an answer), we find that this particularly holds for voice formats. Moreover, we found that the presence of other individuals during survey participation has a negative effect on response length in audio formats. Conversely, when respondents are in a solitary environment, their tendency to provide longer responses in audio formats significantly increases. Additionally, we observed that survey settings that mimic contemporary voice-messaging practices, such as engaging in surveys while being in a non-home setting, can further augment response lengths in audio formats.

Research on comparisons of text and audio response formats for open-ended survey questions is in its early stages. Previous research has mainly focused on the relationship between response format and response length. In our study, we added two more measures: the number of topics and the response entropy. While the number of topics has already been used in research on text-audio comparisons (Gavras et al., 2022; Höhne & Gavras, 2022), entropy is a new approach that could contribute to the calls to include other measures of lexical diversity in such anal-

yses (Gavras et al., 2022). Our study presents the first account of entropy in the context of text-audio comparisons, reflecting the need for further research in this area. While entropy is commonly used as a measure of lexical richness in quantitative linguistics (Grabchak et al., 2013; Rajput et al., 2018; Shi & Lei, 2022), its linguistic significance in the context of open-ended survey answers requires further evaluation. In our computation of entropy, we employed a naive estimation known as Shannon Entropy (Shannon, 1948), which, despite its popularity and appropriateness, has some shortcomings (Shi & Lei, 2022). Alternatives (Grabchak et al., 2013; Shi & Lei, 2022) include the Zhang estimation (Zhang, 2012) or the Nemenman-Shafee-Bialek estimation (Nemenman et al., 2001). Moreover, besides entropy, there exist alternative measures for assessing lexical richness, such as Guiraud's Index (Daller & Xue, 2007; van Hout & Vermeer, 2007). More generally, lexical richness is just one of many potential measures of text complexity and information content, and could be used in combination with other methods to ensure a comprehensive analysis.

Importantly, our study highlights the importance of considering both the potential merits and downsides of voice-based answers. While we found voice-based answers to have higher information content, researchers should pay attention to the high item non-response rates which can limit their overall utility. In our sample, on average, only 47% of participants assigned to the voice condition provided open-ended voice answers, compared to 99% for the text format. Similar dropout rates have been observed in other studies (Gavras et al., 2022; Lütters et al., 2018; Revilla & Couper, 2021; Revilla et al., 2020). Further research is needed to understand the reasons behind these high dropout rates and whether certain respondent groups are more likely to drop out. In Appendix A.3 we provide a first, explorative analysis. In general, next to technical reasons, there could be several possible explanations for group differences in response rates. These explanations might be broadly categorized into two classes: differences in technology use and differences in survey response behavior. One potential solution is to allow respondents choose their preferred response format, as suggested by previous studies (Chen et al., 2022; Höhne, 2021). We included interview context-related characteristics and concluded that once the survey setting is appropriate for voice formats (e.g., respondent is alone) an audio format can have positive effects on data quality (e.g.,

answer length, cf. Figure 3.1). Lenzner & Höhne (2022) suggest applying ideas from the technology acceptance model (TAM) by Davis (1989). TAM suggests that there are two key factors that determine a person's acceptance and usage of new technology: their perceived usefulness of the technology and their perceived ease of use.

Ultimately, understanding and addressing these factors will be crucial for maximizing the potential of new survey response formats such as voice-based input. Recent advancements in computational methods, automated speech recognition (ASR) systems (Proksch et al., 2019), the widespread use of smartphones and voice input (VI) technology coupled with increased willingness to use such options within web surveys (Revilla et al., 2020), have opened up promising avenues for collecting audio data in survey research. These developments, in combination with insights concerning respondent preferences and behaviors, could help pave the way for innovative data collection methods in future survey research.

References

American Association for Public Opinion Research (AAPOR). (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys (9th ed*. Oakbrook Terrace, IL: AAPOR.

Barth, A., & Schmitz, A. (2021). Interviewers' and respondents' joint production of response quality in openended questions. a multilevel negativebinomial regression approach. *Methods, data, analyses: a*, *15*(1), 43–76.

Bauer, P. C., Barberá, P., Ackermann, K., & Venetz, A. (2017). Is the Left-Right scale a valid measure of ideology? *Political Behavior*, *39*(3), 553–583.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774.

Berens, F., & Hobert, S. (2022). Influence of the design of Open-Ended survey questions using conversational agents or Voice-Based input on the responses. *DAGStat Conference 2022*.

Budescu, D. V., & Budescu, M. (2012). How to measure diversity when you must. *Psychol. Methods*, *17*(2), 215–227.

Callegaro, M., & DiSogra, C. (2008). Computing response metrics for online panels. *Public Opin. Q.*, *72*(5), 1008–1032.

Chen, P., Sibia, N., Zavaleta Bernuy, A., Liut, M., & Williams, J. J. (2022). Investigating the impact of voice response options in surveys. *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 2*, 1124.

Daller, H., & Xue, H. (2007, August). Lexical richness and the oral proficiency of chinese EFL students. In *Modelling and assessing vocabulary knowledge* (pp. 150–164). Cambridge University Press.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Miss. Q.*, *13*(3), 319–340.

Gavras, K., & Höhne, J. K. (2022). Evaluating political parties: Criterion validity of open questions with requests for text and voice answers. *Int. J. Soc. Res. Methodol.*, *25*(1), 135–141.

Gavras, K., Höhne, J. K., Blom, A. G., & Schoen, H. (2022). Innovating the collection of open-ended answers: The linguistic and content characteristics of written and oral answers to political attitude questions. *J. R. Stat. Soc. Ser. A Stat. Soc.*, *tba*(tba), 1–19.

Grabchak, M., Zhang, Z., & Zhang, D. T. (2013). Authorship attribution using entropy. *J. Quant. Linguist.*, *20*(4), 301–313.

Groves, R. M. (2011). Three eras of survey research. *Public Opin. Q.*, *75*(5), 861–871.

Gummer, T., Quoß, F., & Roßmann, J. (2019). Does increasing mobile device coverage reduce heterogeneity in completing web surveys on smartphones? *Soc. Sci. Comput. Rev.*, *37*(3), 371–384.

Heffington, C., Park, B. B., & Williams, L. K. (2019). The "most important problem" dataset (MIPD): A new dataset on american issue importance. *Conflict Management and Peace Science*, *36*(3), 312–335.

Höhne, J. K., Gavras, K., & Qureshi, D. (2021). SurveyVoice (SVoice): A comprehensive guide for collecting voice answers in surveys. zenodo. available from: Https://doi.org/10.5281/zenodo.4644590.

Höhne, J. K. (2021). Are respondents ready for audio and voice communication channels in online surveys? *Int. J. Soc. Res. Methodol.*, 1–8.

Höhne, J. K., & Gavras, K. (2022). Typing or speaking? comparing text and voice answers to open questions on sensitive topics in smartphone surveys. *Available at SSRN: https://ssrn.com/abstract=4239015*.

Iyengar, S. (1996). Framing responsibility for political issues. *Ann. Am. Acad. Pol. Soc. Sci.*, *546*, 59–70.

Johnson, W. (1944). I. a program of research. *Psychol. Monogr.*, *56*(2), 1–15.

Lambert, A. D., & Miller, A. L. (2015). Living with smartphones: Does completion device affect survey responses? *Res. High. Educ.*, *56*(2), 166–177.

Landesvatter, C., Behnert, J., & Bauer, P. C. (2023). Comparing Speech-to-Text algorithms for transcribing voice data from surveys. *SocArXiv, osf.io/vk6wj*.

Lazarsfeld, P. F. (1935). The art of asking WHY in marketing research: Three principles underlying the formulation of questionnaires. *National Marketing Review*, *1*(1), 26–38.

Lenzner, T., & Höhne, J. K. (2022). Who is willing to use audio and voice inputs in smartphone surveys, and why? *International Journal of Market Research*, *64*(5), 594–610.

Lugtig, P. J., & Toepoel, V. (2016a). Mobile-Only web survey respondents. *Survey Practice*, *9*(3), 1–8.

Lugtig, P. J., & Toepoel, V. (2016b). The use of PCs, smartphones, and tablets in a Probability-Based panel survey: Effects on survey measurement error. *Soc. Sci. Comput. Rev.*, *34*(1), 78–94.

Lütters, H., Friedrich-Freksa, M., & Egger, M. (2018). Effects of speech assistance in online questionnaires. *General Online Research Conference, Vol. 18*.

Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Soc. Sci. Comput. Rev.*, *31*(6), 725–743.

Nemenman, I., Shafee, F., & Bialek, W. (2001). Entropy and inference, revisited. *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 471–478.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Peterson, G., Griffin, J., LaFrance, J., & Li, J. (2017, February). Smartphone participation in web surveys. In *Total survey error in practice* (pp. 203–233). John Wiley & Sons, Inc.

Peytchev, A., & Hill, C. A. (2010). Experiments in mobile web survey design. *Soc. Sci. Comput. Rev.*, *28*(3), 319–335.

Proksch, S.-O., Wratil, C., & Wäckerle, J. (2019). Testing the validity of automatic speech recognition for political text analysis. *Polit. Anal.*, *27*(3), 339–359.

Rajput, N. K., Ahuja, B., & Riyal, M. K. (2018). A novel approach towards deriving vocabulary quotient. *Digital Scholarship Humanities*, *33*(4), 894–901.

Revilla, M., & Couper, M. P. (2021). Improving the use of voice recording in a smartphone survey. *Soc. Sci. Comput. Rev.*, *39*(6), 1159–1178.

Revilla, M., Couper, M. P., Bosch, O. J., & Asensio, M. (2020). Testing the use of voice input in a smartphone web survey. *Soc. Sci. Comput. Rev.*, *38*(2), 207–224.

Revilla, M., Couper, M. P., & Ochoa, C. (2018). Giving respondents voice? the feasibility of voice input for mobile web surveys. *Surv. Pract.*, *11*(2), 1–11.

Revilla, M., Toninelli, D., Ochoa, C., & Loewe, G. (2016). Do online access panels need to adapt surveys for mobile devices? *Internet Research*, *26*(5), 1209–1227.

Richards, B. (1987). Type/Token ratios: What do they really tell us? *J. Child Lang.*, *14*(2), 201–209.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: R package for structural topic models. *Journal of Statistical Software*, *91*(2), 1–40.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *Am. J. Pol. Sci.*, *58*(4), 1064–1082.

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.

Scholz, E., & Zuell, C. (2012). Item non-response in open-ended questions: Who does not answer on the meaning of left and right? *Soc. Sci. Res.*, *41*(6), 1415–1428.

Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. *Am. Sociol. Rev.*, *31*(2), 218–222.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, *27*(3), 379–423.

Shi, Y., & Lei, L. (2022). Lexical richness and text length: An entropy-based perspective. *J. Quant. Linguist.*, *29*(1), 62–79.

Singer, E., & Couper, M. P. (2017). Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys. *methods, data, analyses*, *11*(2), 20.

Singer, M. (2011). Who says "it's the economy"? Cross-National and Cross-Individual variation in the salience of economic performance. *Comp. Polit. Stud.*, *44*(3), 284–312.

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.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychol. Bull.*, *103*(3), 299–314.

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

van Hout, R., & Vermeer, A. (2007, August). Comparing measures of lexical richness. In *Modelling and assessing vocabulary knowledge* (pp. 93–115). Cambridge University Press.

Willis, G. B. (2004, September). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE Publications.

Zhang, Z. (2012). Entropy estimation in turing's perspective. *Neural Comput.*, *24*(5), 1368–1389.

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. *Soc. Sci. Comput. Rev.*, *33*(1), 115–122.

A.1 Question Wordings

Below, we outline the wording of our different survey measures.

Measure	Question wording	Response scale & recoding
Climate Evaluation	How confident are you that actions taken by the international community will significantly reduce the effects of global climate change: very confident, somewhat confident, not too confident, or not at all confident?	Very confident; Somewhat confident; Not too confident; Not at all confident; Don't know
Climate Evaluation Probe	The previous question was: ... Your answer was: ... In your own words, please explain why you selected this answer.	open textbox
Education	What is the highest level of education you have completed?	Original scale: Less than a high school diploma; GED, High school diploma; Associates Degree (AA,AS,etc.), Bachelor's degree (BA,BS,etc.), Doctorate or professional degree (PhD, MD, JD, DDS, etc.); Don't know; Recoded scale: Less than a high school diploma; High school diploma (+ GED), Bachelor's degree, Advanced Degree
Education Probe	The previous question was: ... Your answer was: ... In your own words, please explain why you decided {not} to earn a college degree.	open textbox
Future Children	Are you planning to have {a} {another} child in the next two years?	Definitely yes; Probably yes; Probably not; Definitely not; Don't know; Prefer not to Answer
Future Children Probe	The previous question was: ... Your answer was ... In your own words, please explain why you are {not} planning to have {a} {another} child in the next two years	open textbox

Continued on next page

(continued)

Measure	Question wording	Response scale & recoding
Life Satisfaction	All things considered, how satisfied are you with your life as a whole these days?	Extremely satisfied; Very satisfied; Moderately satisfied; Slightly satisfied; Not satisfied at all, Don't Know
Life Satisfaction Probe	The previous question was: ... Your answer was: ... In your own words, please explain why you selected this answer	open textbox
Most Important Problem	What do you think is the most important problem facing this country? Please describe.	open textbox
Political Trust	How often can you trust the federal government in Washington to do what is right?	Always; Most of the time; About half of the time; Some of the time; Never; Don't Know
Political Trust Probe	The previous question was: ... Your answer was: ... In your own words, please explain why you selected this answer.	open textbox
President	What do you think about the US president Joe Biden? Please describe.	open textbox
Generalized Trust	Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people? Please tell us on a score of 0 to 6, where 0 means you can't be too careful and 6 means that most people can be trusted.	0 - You can't be too careful; 1; 2; 3; 4; 5; 6 - Most people can be trusted; Don't know
Generalized Trust Probe	The previous question was: ... Your answer was: ... In your own words, please explain why you selected this answer.	open textbox
Vote Choice	In 2020, Joe Biden ran for President on the Democratic ticket against Donald Trump for the Republicans. Did you vote for Joe Biden or Donald Trump?	Joe Biden; Donald Trump; Other Candidate; Don't Know, Prefer not to Answer

Continued on next page

(continued)

Measure	Question wording	Response scale & recoding
Vote Choice Probe	The previous question was: ... Your answer was: ... In your own words, please explain why you decided to vote the way you did.	open textbox
Age (numeric)	What is your current age in years?	Original scale: Simple numeric entry; Recoded scale: Recoded to factor with four levels 18-27, 28-37, 38-47, 48-57, 58-80+
Sex (factor)	What is your sex, as recorded on legal/official documents?	Original scale: Two answers options 'Male' and 'Female'; Recoded scale: Recoded to factor with two levels (1) Female and (2) Male.
Ethnicity (factor)	Please indicate your ethnicity (i.e. peoples' ethnicity describes their feeling of belonging and attachment to a distinct group of a larger population that shares their ancestry, colour, language or religion)?	Original scale: Five answer options 'Asian', 'Black', 'Mixed', 'Other' and 'White'; Recoded scale: Recoded to factor with corresponding levels. Reference category is 'Asian'.
Socioeconomic status (numeric)	Think of a ladder as representing where people stand in society. At the top of the ladder are the people who are best off—those who have the most money, most education and the best jobs. At the bottom are the people who are worst off—who have the least money, least education and the worst jobs or no job. The higher up you are on this ladder, the closer you are to people at the very top and the lower you are, the closer you are to the bottom. Where would you put yourself on the socioeconomic ladder? Choose the number whose position best represents where you would be on this ladder.	Original scale: Ten answer options; Recoded scale: Numeric with 10 values.

Continued on next page

(continued)

Measure	Question wording	Response scale & recoding
Household Income (numeric)	What is your total household income per year, including all earners in your household (after tax) in USD?	Original scale: Answer options 1 - Less than $10,000; 2 - 10,000-15,999; 3 - 16,000-19,999; 4 - 20,000-29,999; 5 - 30,000-39,999; 6 - 40,000-49,999; 7 - 50,000-59,999; 8 - 60,000-69,999; 9 - 70,000-79,999; 10 - 80,000-89,999; 11 - 90,000-99,999; 12 - 10,000-149,999;More than 150,000; Prefer not to Answer; Recoded scale: Numeric with 13 values, Don't know = NA.
Survey mode (alone/with others)	Were you alone when answering the questions or were there others present?	I was alone; Other persons were present, Don't Know
Survey mode (home/public space)	From where did you participate in this survey?	From home; from location; Prefer not to Answer

Table A3.1: Question wordings.

A.2 Summary Statistics

Covariate balance

Table A3.2 presents summary statistics for our two experimental conditions, audio (treatment) and text (condition) alongside the results of t-tests conducted to determine whether any statistically significant differences were present between the two groups. The analysis revealed no significant differences in any of the covariates between the two groups.

Covariate	Text			Audio			
	mean	sd	n	mean	sd	n	p
Age	40.97	14.04	800	41.95	14.75	661	0.19
Gender	0.53	0.50	800	0.55	0.50	661	0.07
Education	2.86	0.72	800	2.83	0.68	661	0.36
Socioeconomic Status	4.92	1.72	800	5.00	1.72	661	0.38
Survey mode (alone)	0.96	0.18	800	0.95	0.21	661	0.00
Survey mode (home)	0.88	0.32	800	0.88	0.32	661	0.71

Table A3.2: Summary statistics by experimental condition.

Summary Statistics: Numerical and Categorical

Table A3.3 and Table A3.4 provide summary statistics of our covariates, split by whether they are numerical or categorical variables. Whenever applicable, we also include population estimates from the US Census. Additionally, Table A3.5 shows frequency counts of the single values.

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
Socio-economic status (numeric)	9	0	4.95	1.72	1.00	5.00	9.00
Income (numeric)	13	0	7.48	3.73	1.00	7.00	13.00

Table A3.3: Summary statistics: Numeric covariates.

		Sample		Census	
Covariate	Categories	N	%	N	%
Age (factor)	18-27	290	20.4	43,355,638	17.9
	28-37	346	24.3	42,085,420	17.4
	38-47	285	20.0	39,974,287	16.5
	48-57	251	17.7	43,370,543	17.9
	58-80+	250	17.6	73,462,149	30.3
Sex (factor)	Female	771	54.2	125,196,929	51.7
	Male	651	45.8	117,051,108	48.3
Ethnicity (factor)	Asian	62	4.4	14,040,646	5.8
	Black	144	10.1	30,097,066	12.4
	Mixed	28	2.0	3,893,117	1.6

Continued on next page

(continued)

		Sample		Census	
Covariate	Categories	N	%	N	%
	Other	18	1.3	3,601,403	1.5
	White	1170	82.3	190,615,805	78.7
Education (factor)	Less than a high school diploma	11	0.8	NA	NA
	High school diploma (+ GED)	438	30.8	NA	NA
	Bachelor's degree (+ Associates Degree)	725	51.0	NA	NA
	Advanced Degree (Master/Doctorate/Professional)	248	17.4	NA	NA
Survey Participation (others were present)	alone	1366	96.1	NA	NA
	others were present	56	3.9	NA	NA
Survey Participation (another location)	home	1258	88.5	NA	NA
	another location	164	11.5	NA	NA

Table A3.4: Summary statistics: Categorical covariates.

Characteristic	N = 1,422
Age (factor)	
18-27	290 (20%)
28-37	346 (24%)
38-47	285 (20%)
48-57	251 (18%)
58-80+	250 (18%)
Sex (factor)	
Female	771 (54%)
Male	651 (46%)
Ethnicity (factor)	
Asian	62 (4.4%)
Black	144 (10%)
Mixed	28 (2.0%)
Other	18 (1.3%)
White	1,170 (82%)
Socio-economic status (numeric)	
1	33 (2.3%)
2	91 (6.4%)
3	189 (13%)
4	246 (17%)
5	279 (20%)
6	293 (21%)
7	218 (15%)
8	65 (4.6%)
9	8 (0.6%)
Income (numeric)	7 (4, 11)

Continued on next page

(continued)

Characteristic	**N = 1,422**
Education (factor)	
Less than a high school diploma	11 (0.8%)
High school diploma (+ GED)	438 (31%)
Bachelor's degree (+ Associates Degree)	725 (51%)
Advanced Degree (Master/Doctorate/Professional)	248 (17%)
Survey Participation (others were present)	
alone	1,366 (96%)
others were present	56 (3.9%)
Survey Participation (another location)	
home	1,258 (88%)
another location	164 (12%)

[1] n (%); Median (IQR)

Table A3.5: Frequency counts: Numeric covariates.

A.3. Nonresponse by Experimental Condition

Following the mean (item-)nonresponse rate values already described in Section "Data", Figure A3.1 visualizes nonresponse rates by item and experimental condition.

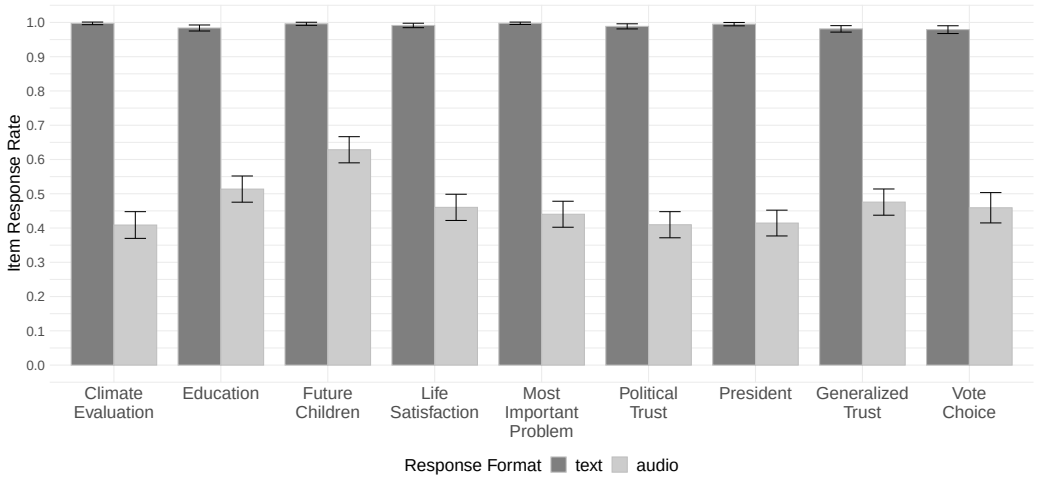

Figure A3.1: Response rates by experimental condition and item.

Figure A3.2 depicts item response rates by response format and respondent characteristics. Figure A3.2 shows how respondent's sociodemographics and interview context-related characteristics influence the likelihood to answer in the voice format. Results are from linear probability models and the following plots show interaction effects of the response format (0: text, 1: audio) and respective predictor variables (cf. Section "Measures and Operationalisation").

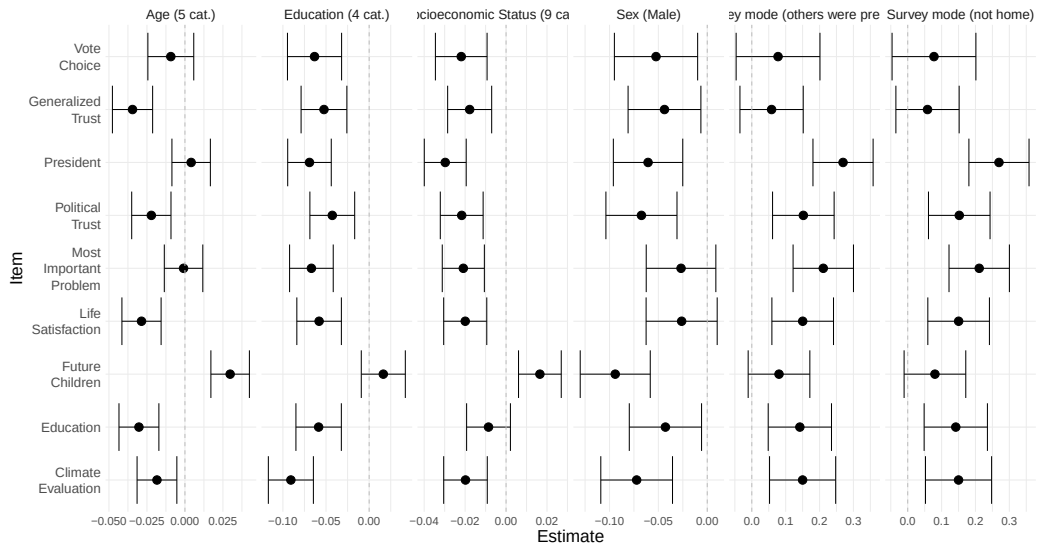

Figure A3.2: Probabilities to answer for respondent subgroups by item (audio vs. text). *Note*: Linear Probability models of self-reported respondent characteristics on probability to answer open-ended survey questions comparing audio (audio=1) to text. Predictor Variables: Age (18–27, 28–37, 38–47, 48–57, 58–80+), Education (less than a high school diploma, high school diploma (+ GED), Bachelor's degree, Advanced Degree); Socioeconomic Status (1–9), Sex (female, male), Survey mode (alone, others were present), Survey mode (home, another location), Voice Input Daily Life (Never, 2, 3, 4, Daily), Survey Evaluation (Very difficult, Rather difficult, Rather easy, Very easy).

For some items, the education and socioeconomic status (SES) of respondents seem to influence their willingness to participate in our OEQs using voice, with higher levels indicating lower probability to participate. Put differently, higher education and SES increases participation in the text format which we might attribute to the fact that these groups have sufficiently high literacy skills required for typewriting an answer (cf. Revilla et al., 2018). Female respondents are more likely to respond to the voice format than male respondents. There is no consistent effect of age on the likelihood to respond in voice which is in line with previous research (Revilla et al., 2020) showing no effect for age on answering at least one voice question]. Respondents who completed the survey in the presence of others or at a location other than their home are also more likely to respond to the voice requests. The latter finding might seem surprising (Revilla et al., 2018) but could

indicate a usage behavior of such voice surveys similar to mobile voice input messaging (especially when questions are unproblematic and not sensitive).
In general, there are several possible explanations for group differences in response rates. These explanations can be broadly categorized into two classes: differences in technology use and differences in survey response behavior. These factors can potentially influence the likelihood of responding in a certain mode, such as the voice input mode.
Overall, these findings are still exploratory in nature, and the effects are not consistently observed when the results are analyzed by individual items (cf. Figure A3.2). In previous research on the question of systematic drop-out in voice requests, it has been found that there are no significant differences in sociodemographic variables (e.g., Höhne & Gavras, 2022).

A.4 Regression tables

Response Length and Respondent Characteristics

Below, we show the results of a linear regression model of various covariates (age, education, socioeconomic status (SES), sex, survey mode (alone/others were present), survey mode (home/public space)) and the response length. The coefficients below can be assigned to Figure 3.4 in the main paper, where we plotted the respective interactions (i.e., audio*covariate).

	Climate Evaluation	Education	Future Children	Life Satisfaction	Most Important Problem	Political Trust	President	Generalized Trust	Vote Choice
Age									
(Intercept)	25.273***	21.107***	22.043***	25.426***	18.443***	23.133***	24.610***	24.528***	23.749***
	(1.478)	(1.463)	(1.189)	(1.896)	(1.592)	(1.384)	(1.486)	(1.334)	(1.743)
Audio	2.462	6.384*	9.103***	4.987	4.284	4.560+	1.319	5.463*	0.446
	(2.855)	(2.563)	(2.039)	(3.488)	(3.050)	(2.720)	(2.909)	(2.410)	(3.271)
Age (cat)	0.020	0.269	-1.918***	0.352	-0.129	0.382	-0.221	-0.320	0.260
	(0.470)	(0.464)	(0.374)	(0.602)	(0.507)	(0.441)	(0.473)	(0.422)	(0.543)
Audio*Age(cat)	-0.252	-0.601	-0.633	-1.615	-0.436	-0.066	0.137	-0.308	0.374
	(0.899)	(0.811)	(0.613)	(1.111)	(0.951)	(0.871)	(0.904)	(0.767)	(0.992)
Num.Obs.	984	1126	1137	1092	1089	1040	1070	1095	830
R2	0.002	0.015	0.091	0.002	0.005	0.014	0.002	0.018	0.003
R2 Adj.	-0.001	0.012	0.088	-0.001	0.002	0.011	-0.001	0.015	-0.001
AIC	8414.4	9679.4	9238.8	9957.8	9564.0	8810.9	9246.3	9187.9	7169.7
BIC	8438.8	9704.5	9264.0	9982.8	9589.0	8835.6	9271.2	9212.9	7193.3
Log.Lik.	-4202.187	-4834.688	-4614.404	-4973.908	-4777.008	-4400.436	-4618.165	-4588.970	-3579.852
RMSE	17.32	17.72	14.01	23.01	19.45	16.65	18.12	15.99	18.07
Education									
(Intercept)	21.876***	24.074***	18.466***	26.680***	17.435***	24.173***	18.924***	24.439***	26.529***
	(2.693)	(2.602)	(2.153)	(3.373)	(2.828)	(2.452)	(2.635)	(2.347)	(3.111)
Audio	0.027	-8.040+	3.161	-6.263	-6.272	-9.549+	2.972	-5.429	-6.683
	(5.330)	(4.873)	(3.794)	(6.535)	(5.532)	(4.917)	(5.453)	(4.385)	(6.363)
Education (cat)	1.196	-0.767	-0.669	-0.089	0.224	0.015	1.768*	-0.287	-0.697
	(0.906)	(0.880)	(0.727)	(1.143)	(0.958)	(0.829)	(0.893)	(0.794)	(1.041)
Audio*Education(cat)	0.692	4.546**	1.254	2.425	3.371+	4.999**	-0.392	3.602*	2.912
	(1.867)	(1.687)	(1.290)	(2.273)	(1.927)	(1.702)	(1.910)	(1.516)	(2.184)
Num.Obs.	984	1126	1137	1092	1089	1040	1070	1095	830

Continued on next page

(continued)

	Climate Evaluation	Education	Future Children	Life Satisfaction	Most Important Problem	Political Trust	President	Generalized Trust	Vote Choice
R2	0.005	0.021	0.049	0.001	0.009	0.024	0.006	0.023	0.004
R2 Adj.	0.002	0.019	0.046	-0.001	0.006	0.021	0.003	0.020	0.000
AIC	8411.4	9672.3	9290.4	9958.6	9559.9	8800.4	9242.0	9182.8	7168.7
BIC	8435.8	9697.4	9315.6	9983.5	9584.9	8825.2	9266.8	9207.7	7192.3
Log.Lik.	-4200.689	-4831.152	-4640.212	-4974.284	-4774.945	-4395.224	-4615.984	-4586.377	-3579.367
RMSE	17.29	17.67	14.33	23.02	19.41	16.56	18.08	15.95	18.06
Socioeconomic Status									
(Intercept)	25.073***	23.359***	18.430***	30.751***	20.386***	24.933***	21.339***	24.234***	24.408***
	(1.960)	(1.925)	(1.605)	(2.500)	(2.099)	(1.803)	(1.955)	(1.750)	(2.275)
Audio	0.298	1.192	6.197*	-6.548	-1.175	0.622	6.107	1.126	-0.691
	(3.801)	(3.471)	(2.808)	(4.576)	(3.946)	(3.483)	(3.718)	(3.193)	(4.376)
SES	0.061	-0.290	-0.380	-0.861+	-0.461	-0.141	0.551	-0.115	0.030
	(0.375)	(0.369)	(0.308)	(0.479)	(0.403)	(0.347)	(0.375)	(0.336)	(0.432)
Audio*SES	0.286	0.693	0.115	1.410	0.847	0.728	-0.919	0.702	0.396
	(0.730)	(0.662)	(0.527)	(0.879)	(0.761)	(0.675)	(0.725)	(0.613)	(0.829)
Num.Obs.	982	1120	1134	1087	1084	1037	1065	1091	826
R2	0.002	0.015	0.049	0.004	0.006	0.013	0.004	0.018	0.001
R2 Adj.	-0.001	0.012	0.046	0.001	0.003	0.010	0.001	0.015	-0.002
AIC	8397.1	9631.6	9267.2	9913.2	9522.7	8778.9	9203.4	9156.5	7128.9
BIC	8421.6	9656.7	9292.3	9938.2	9547.6	8803.6	9228.3	9181.4	7152.5
Log.Lik.	-4193.557	-4810.808	-4628.578	-4951.617	-4756.351	-4384.441	-4596.708	-4573.233	-3559.434
RMSE	17.31	17.75	14.33	23.02	19.47	16.59	18.12	16.00	18.00
Sex (Male)									
(Intercept)	24.797***	22.461***	16.851***	27.116***	17.733***	23.773***	23.943***	24.388***	25.441***
	(0.881)	(0.858)	(0.715)	(1.125)	(0.944)	(0.822)	(0.882)	(0.786)	(1.017)
Audio	3.051+	4.395**	7.075***	-0.897	3.852*	5.253***	2.349	4.234**	0.665
	(1.684)	(1.524)	(1.184)	(2.103)	(1.787)	(1.572)	(1.681)	(1.438)	(1.871)
Sex (Male)	0.993	-1.419	-0.645	-1.553	0.573	0.933	0.084	-1.656	-2.060
	(1.282)	(1.259)	(1.059)	(1.653)	(1.383)	(1.199)	(1.295)	(1.152)	(1.474)
Audio*Male	-2.756	0.591	-1.062	2.937	-1.634	-1.977	-1.733	0.694	1.954
	(2.559)	(2.305)	(1.829)	(3.155)	(2.686)	(2.407)	(2.582)	(2.162)	(2.860)
Num.Obs.	979	1120	1130	1085	1083	1034	1064	1089	825
R2	0.003	0.016	0.048	0.001	0.005	0.014	0.002	0.019	0.004
R2 Adj.	0.000	0.014	0.045	-0.002	0.003	0.011	-0.001	0.016	0.000
AIC	8364.3	9602.7	9236.9	9897.5	9505.8	8754.2	9195.0	9137.3	7120.5
BIC	8388.8	9627.8	9262.1	9922.5	9530.8	8778.9	9219.9	9162.3	7144.1
Log.Lik.	-4177.170	-4796.331	-4613.454	-4943.774	-4747.919	-4372.083	-4592.519	-4563.661	-3555.254
RMSE	17.25	17.52	14.35	23.05	19.40	16.60	18.13	15.99	18.00

Continued on next page

(continued)

	Climate Evaluation	Education	Future Children	Life Satisfaction	Most Important Problem	Political Trust	President	Generalized Trust	Vote Choice
Survey Mode (others were present)									
(Intercept)	25.342***	21.938***	16.481***	26.534***	18.187***	24.307***	24.049***	23.653***	24.363***
	(0.650)	(0.643)	(0.533)	(0.836)	(0.702)	(0.609)	(0.655)	(0.582)	(0.745)
Audio	2.057	5.128***	7.089***	0.554	3.234*	4.689***	1.555	5.034***	1.795
	(1.303)	(1.182)	(0.919)	(1.606)	(1.375)	(1.224)	(1.316)	(1.096)	(1.442)
Survey Mode (others were present)	-0.433	-1.831	1.978	-3.070	-3.116	-2.521	-1.835	-1.060	5.771
	(3.753)	(3.410)	(2.976)	(4.438)	(3.746)	(3.207)	(3.493)	(3.133)	(4.733)
Audio*NotAlone	-5.032	-6.785	-6.820	-0.176	-1.755	-3.945	2.849	-7.814	-7.129
	(5.950)	(5.321)	(4.332)	(7.042)	(5.864)	(5.268)	(5.404)	(5.164)	(7.530)
Num.Obs.	984	1126	1137	1092	1089	1040	1070	1095	830
R2	0.003	0.019	0.050	0.001	0.006	0.016	0.002	0.021	0.003
R2 Adj.	0.000	0.016	0.047	-0.002	0.004	0.013	-0.001	0.018	0.000
AIC	8413.0	9675.2	9288.8	9959.2	9562.7	8808.8	9246.2	9184.7	7169.0
BIC	8437.5	9700.4	9313.9	9984.1	9587.7	8833.5	9271.1	9209.7	7192.6
Log.Lik.	-4201.524	-4832.609	-4639.381	-4974.579	-4776.356	-4399.393	-4618.112	-4587.342	-3579.487
RMSE	17.30	17.69	14.32	23.02	19.43	16.63	18.12	15.97	18.06
Survey Mode (public)									
(Intercept)	25.557***	21.904***	16.400***	26.393***	17.994***	24.550***	24.001***	23.459***	24.449***
	(0.679)	(0.673)	(0.559)	(0.874)	(0.734)	(0.635)	(0.684)	(0.611)	(0.790)
Audio	0.815	4.455***	6.545***	-0.213	2.379	3.476**	1.777	4.620***	1.242
	(1.355)	(1.229)	(0.958)	(1.674)	(1.447)	(1.275)	(1.359)	(1.137)	(1.513)
Survey Mode (public)	-2.015	-0.262	1.200	0.274	0.715	-2.948	-0.144	1.326	0.437
	(2.019)	(1.969)	(1.613)	(2.546)	(2.149)	(1.885)	(2.023)	(1.771)	(2.185)
Audio*Public	7.340+	2.528	1.579	5.246	3.586	7.350*	-0.222	0.322	2.837
	(3.813)	(3.579)	(2.755)	(4.671)	(3.797)	(3.564)	(3.895)	(3.441)	(4.261)
Num.Obs.	984	1125	1136	1091	1088	1039	1069	1094	830
R2	0.006	0.015	0.049	0.002	0.006	0.017	0.002	0.018	0.003
R2 Adj.	0.003	0.013	0.047	-0.001	0.004	0.014	-0.001	0.015	-0.001
AIC	8410.8	9670.0	9282.1	9949.7	9554.6	8799.4	9238.3	9181.1	7169.7
BIC	8435.2	9695.1	9307.3	9974.7	9579.5	8824.2	9263.2	9206.1	7193.3
Log.Lik.	-4200.377	-4830.004	-4636.073	-4969.863	-4772.293	-4394.720	-4614.170	-4585.534	-3579.838
RMSE	17.28	17.71	14.33	23.02	19.44	16.62	18.13	16.00	18.07

Table A3.6: Regression of Response Length and Respondent Characteristics. *Note*: Stars indicate signifcance levels +=.1, *=.05, **=.01, ***=0.001.

Number of Topics and Respondent Characteristics

Below, we show the results of a linear regression model of various covariates (age, education, socioeconomic status, sex, survey mode (alone/others were present), survey mode (home/public space)) and the number of topics. We show the estimated regression coefficients and model fit statistics.

	Climate Evaluation	Education	Future Children	Life Satisfaction	Most Important Problem	Political Trust	President	Generalized Trust	Vote Choice
Age									
(Intercept)	3.396***	2.960***	2.911***	3.349***	2.196***	3.320***	3.116***	3.728***	3.118***
	(0.105)	(0.098)	(0.093)	(0.090)	(0.082)	(0.105)	(0.098)	(0.088)	(0.117)
Audio	0.591**	0.769***	0.092	0.435**	0.638***	0.650**	0.436*	0.054	0.631**
	(0.203)	(0.172)	(0.158)	(0.166)	(0.157)	(0.208)	(0.191)	(0.159)	(0.220)
Age (cat)	-0.084*	0.042	0.086**	-0.093**	-0.044+	-0.048	-0.056+	-0.020	0.003
	(0.033)	(0.031)	(0.029)	(0.029)	(0.026)	(0.034)	(0.031)	(0.028)	(0.036)
Audio*Age(cat)	-0.084	-0.040	0.062	0.002	0.095+	-0.087	0.034	0.070	0.044
	(0.064)	(0.055)	(0.048)	(0.053)	(0.049)	(0.067)	(0.060)	(0.051)	(0.067)
Num.Obs.	983	1123	1123	1089	1086	1038	1068	1093	828
R2	0.031	0.062	0.037	0.045	0.143	0.026	0.040	0.013	0.073
R2 Adj.	0.028	0.059	0.035	0.042	0.141	0.023	0.037	0.011	0.070
AIC	3201.4	3579.5	3378.8	3293.6	3088.5	3446.1	3414.0	3221.6	2677.7
BIC	3225.8	3604.6	3403.9	3318.6	3113.4	3470.9	3438.9	3246.6	2701.3
Log.Lik.	-1595.691	-1784.737	-1684.388	-1641.815	-1539.248	-1718.070	-1702.024	-1605.817	-1333.857
RMSE	1.23	1.19	1.08	1.09	1.00	1.27	1.19	1.05	1.21
Education									
(Intercept)	3.283***	2.361***	3.090***	3.111***	1.934***	3.429***	3.240***	3.632***	3.036***
	(0.192)	(0.173)	(0.165)	(0.161)	(0.146)	(0.188)	(0.174)	(0.155)	(0.209)
Audio	0.542	0.970**	0.516+	0.933**	1.279***	0.033	-0.155	-0.202	1.225**
	(0.382)	(0.326)	(0.290)	(0.313)	(0.286)	(0.379)	(0.360)	(0.290)	(0.429)
Education (cat)	-0.044	0.251***	0.023	-0.009	0.048	-0.086	-0.099+	0.014	0.031
	(0.065)	(0.059)	(0.056)	(0.055)	(0.049)	(0.064)	(0.059)	(0.052)	(0.070)
Audio*Education(cat)	-0.072	-0.105	-0.075	-0.177	-0.131	0.132	0.245+	0.163	-0.163
	(0.134)	(0.113)	(0.099)	(0.109)	(0.099)	(0.131)	(0.126)	(0.100)	(0.147)
Num.Obs.	983	1123	1123	1089	1086	1038	1068	1093	828
R2	0.016	0.077	0.017	0.035	0.141	0.021	0.040	0.015	0.074
R2 Adj.	0.013	0.075	0.015	0.032	0.139	0.018	0.038	0.013	0.070
AIC	3215.9	3560.9	3401.9	3304.4	3091.0	3451.8	3413.0	3219.2	2677.2
BIC	3240.3	3586.0	3427.0	3329.3	3116.0	3476.5	3437.8	3244.2	2700.8

Continued on next page

(continued)

	Climate Evaluation	Education	Future Children	Life Satisfaction	Most Important Problem	Political Trust	President	Generalized Trust	Vote Choice
Log.Lik.	-1602.936	-1775.439	-1695.956	-1647.192	-1540.524	-1720.892	-1701.491	-1604.610	-1333.586
RMSE	1.24	1.18	1.10	1.10	1.00	1.27	1.19	1.05	1.21
Socioeconomic Status									
(Intercept)	3.075***	2.891***	2.883***	3.089***	2.012***	3.036***	2.978***	3.341***	2.832***
	(0.140)	(0.129)	(0.124)	(0.120)	(0.108)	(0.138)	(0.129)	(0.115)	(0.153)
Audio	0.634*	1.172***	0.603**	0.492*	1.322***	0.562*	0.405+	0.443*	1.230***
	(0.272)	(0.232)	(0.215)	(0.219)	(0.203)	(0.267)	(0.245)	(0.210)	(0.294)
SES	0.017	0.038	0.055*	-0.002	0.012	0.030	-0.006	0.067**	0.060*
	(0.027)	(0.025)	(0.024)	(0.023)	(0.021)	(0.027)	(0.025)	(0.022)	(0.029)
Audio*SES	-0.058	-0.103*	-0.062	-0.009	-0.084*	-0.031	0.028	-0.039	-0.094+
	(0.052)	(0.044)	(0.040)	(0.042)	(0.039)	(0.052)	(0.048)	(0.040)	(0.056)
Num.Obs.	981	1117	1120	1084	1081	1035	1063	1089	824
R2	0.016	0.066	0.021	0.032	0.144	0.020	0.038	0.020	0.078
R2 Adj.	0.013	0.063	0.019	0.030	0.141	0.018	0.035	0.018	0.074
AIC	3210.7	3559.1	3388.4	3294.3	3075.8	3444.9	3401.0	3205.2	2661.1
BIC	3235.1	3584.2	3413.5	3319.3	3100.7	3469.6	3425.8	3230.1	2684.7
Log.Lik.	-1600.329	-1774.545	-1689.187	-1642.169	-1532.885	-1717.455	-1695.492	-1597.587	-1325.564
RMSE	1.24	1.18	1.09	1.10	1.00	1.27	1.19	1.05	1.21
Sex (Male)									
(Intercept)	3.216***	3.033***	3.128***	3.036***	2.127***	3.213***	3.045***	3.675***	3.206***
	(0.063)	(0.058)	(0.055)	(0.054)	(0.048)	(0.063)	(0.058)	(0.052)	(0.068)
Audio	0.271*	0.795***	0.401***	0.506***	0.916***	0.389**	0.574***	0.280**	0.732***
	(0.121)	(0.103)	(0.090)	(0.100)	(0.092)	(0.121)	(0.110)	(0.095)	(0.126)
Sex (Male)	-0.125	0.110	0.056	0.115	-0.133+	-0.072	-0.189*	-0.006	-0.175+
	(0.092)	(0.085)	(0.081)	(0.079)	(0.071)	(0.092)	(0.085)	(0.076)	(0.099)
Audio*Male	0.173	-0.337*	-0.255+	-0.157	-0.015	0.041	-0.117	-0.068	0.030
	(0.184)	(0.156)	(0.139)	(0.151)	(0.139)	(0.185)	(0.169)	(0.143)	(0.193)
Num.Obs.	978	1117	1116	1082	1080	1032	1062	1087	823
R2	0.017	0.063	0.019	0.032	0.144	0.020	0.045	0.012	0.075
R2 Adj.	0.014	0.060	0.017	0.030	0.142	0.017	0.043	0.009	0.072
AIC	3199.4	3557.1	3364.5	3283.6	3066.3	3436.2	3388.6	3206.8	2659.3
BIC	3223.9	3582.2	3389.5	3308.6	3091.2	3460.9	3413.5	3231.8	2682.9
Log.Lik.	-1594.722	-1773.549	-1677.228	-1636.821	-1528.161	-1713.089	-1689.316	-1598.412	-1324.644
RMSE	1.24	1.18	1.09	1.10	1.00	1.27	1.19	1.05	1.21
Survey Mode (others were present)									
(Intercept)	3.143***	3.084***	3.148***	3.079***	2.065***	3.178***	2.963***	3.678***	3.132***
	(0.046)	(0.043)	(0.041)	(0.040)	(0.036)	(0.047)	(0.043)	(0.038)	(0.050)
Audio	0.361***	0.648***	0.311***	0.455***	0.931***	0.417***	0.537***	0.264***	0.736***

Continued on next page

(continued)

	Climate Evaluation	Education	Future Children	Life Satisfaction	Most Important Problem	Political Trust	President	Generalized Trust	Vote Choice
	(0.093)	(0.079)	(0.071)	(0.077)	(0.071)	(0.094)	(0.087)	(0.072)	(0.097)
Survey Mode (others were present)	0.447+	-0.120	0.227	0.181	0.149	0.144	-0.213	-0.197	-0.266
	(0.268)	(0.229)	(0.228)	(0.216)	(0.193)	(0.245)	(0.230)	(0.206)	(0.317)
Audio*NotAlone	-0.380	0.178	-0.277	-0.241	-0.346	-0.209	0.047	-0.183	0.697
	(0.433)	(0.362)	(0.331)	(0.339)	(0.302)	(0.403)	(0.356)	(0.340)	(0.505)
Num.Obs.	983	1123	1123	1089	1086	1038	1068	1093	828
R2	0.018	0.060	0.018	0.032	0.141	0.019	0.037	0.014	0.074
R2 Adj.	0.015	0.058	0.015	0.030	0.138	0.016	0.035	0.011	0.071
AIC	3214.5	3581.0	3401.5	3307.6	3091.6	3453.4	3416.3	3220.7	2676.5
BIC	3238.9	3606.1	3426.6	3332.5	3116.5	3478.1	3441.2	3245.7	2700.1
Log.Lik.	-1602.238	-1785.508	-1695.726	-1648.784	-1540.799	-1721.691	-1703.147	-1605.340	-1333.249
RMSE	1.23	1.19	1.10	1.10	1.00	1.27	1.19	1.05	1.21
Survey Mode (public)									
(Intercept)	3.143***	3.084***	3.148***	3.079***	2.065***	3.178***	2.963***	3.678***	3.132***
	(0.046)	(0.043)	(0.041)	(0.040)	(0.036)	(0.047)	(0.043)	(0.038)	(0.050)
Audio	0.361***	0.648***	0.311***	0.455***	0.931***	0.417***	0.537***	0.264***	0.736***
	(0.093)	(0.079)	(0.071)	(0.077)	(0.071)	(0.094)	(0.087)	(0.072)	(0.097)
Survey Mode (others were present)	0.447+	-0.120	0.227	0.181	0.149	0.144	-0.213	-0.197	-0.266
	(0.268)	(0.229)	(0.228)	(0.216)	(0.193)	(0.245)	(0.230)	(0.206)	(0.317)
Audio*NotAlone	-0.380	0.178	-0.277	-0.241	-0.346	-0.209	0.047	-0.183	0.697
	(0.433)	(0.362)	(0.331)	(0.339)	(0.302)	(0.403)	(0.356)	(0.340)	(0.505)
Num.Obs.	983	1123	1123	1089	1086	1038	1068	1093	828
R2	0.018	0.060	0.018	0.032	0.141	0.019	0.037	0.014	0.074
R2 Adj.	0.015	0.058	0.015	0.030	0.138	0.016	0.035	0.011	0.071
AIC	3214.5	3581.0	3401.5	3307.6	3091.6	3453.4	3416.3	3220.7	2676.5
BIC	3238.9	3606.1	3426.6	3332.5	3116.5	3478.1	3441.2	3245.7	2700.1
Log.Lik.	-1602.238	-1785.508	-1695.726	-1648.784	-1540.799	-1721.691	-1703.147	-1605.340	-1333.249
RMSE	1.23	1.19	1.10	1.10	1.00	1.27	1.19	1.05	1.21

Table A3.7: Regression of Number of Topics and Respondent Characteristics. *Note*: Stars indicate signifcance levels +=.1, *=.05, **=.01, ***=0.001.

Response Entropy and Respondent Characteristics

Below, we show the results of a linear regression model of various covariates (age, education, socioeconomic status, sex, survey mode (alone/others were present), survey mode (home/public space)) and the response entropy. We show the estimated regression coeffcients and model fit statistics.

	Climate Evaluation	Education	Future Children	Life Satisfaction	Most Important Problem	Political Trust	President	Generalized Trust	Vote Choice
Age									
(Intercept)	3.396***	2.960***	2.911***	3.349***	2.196***	3.320***	3.116***	3.728***	3.118***
	(0.105)	(0.098)	(0.093)	(0.090)	(0.082)	(0.105)	(0.098)	(0.088)	(0.117)
Audio	0.591**	0.769***	0.092	0.435**	0.638***	0.650**	0.436*	0.054	0.631**
	(0.203)	(0.172)	(0.158)	(0.166)	(0.157)	(0.208)	(0.191)	(0.159)	(0.220)
Age (cat)	-0.084*	0.042	0.086**	-0.093**	-0.044+	-0.048	-0.056+	-0.020	0.003
	(0.033)	(0.031)	(0.029)	(0.029)	(0.026)	(0.034)	(0.031)	(0.028)	(0.036)
Audio*Age(cat)	-0.084	-0.040	0.062	0.002	0.095+	-0.087	0.034	0.070	0.044
	(0.064)	(0.055)	(0.048)	(0.053)	(0.049)	(0.067)	(0.060)	(0.051)	(0.067)
Num.Obs.	983	1123	1123	1089	1086	1038	1068	1093	828
R2	0.031	0.062	0.037	0.045	0.143	0.026	0.040	0.013	0.073
R2 Adj.	0.028	0.059	0.035	0.042	0.141	0.023	0.037	0.011	0.070
AIC	3201.4	3579.5	3378.8	3293.6	3088.5	3446.1	3414.0	3221.6	2677.7
BIC	3225.8	3604.6	3403.9	3318.6	3113.4	3470.9	3438.9	3246.6	2701.3
Log.Lik.	-1595.691	-1784.737	-1684.388	-1641.815	-1539.248	-1718.070	-1702.024	-1605.817	-1333.857
RMSE	1.23	1.19	1.08	1.09	1.00	1.27	1.19	1.05	1.21
Education									
(Intercept)	3.283***	2.361***	3.090***	3.111***	1.934***	3.429***	3.240***	3.632***	3.036***
	(0.192)	(0.173)	(0.165)	(0.161)	(0.146)	(0.188)	(0.174)	(0.155)	(0.209)
Audio	0.542	0.970**	0.516+	0.933**	1.279***	0.033	-0.155	-0.202	1.225**
	(0.382)	(0.326)	(0.290)	(0.313)	(0.286)	(0.379)	(0.360)	(0.290)	(0.429)
Education (cat)	-0.044	0.251***	0.023	-0.009	0.048	-0.086	-0.099+	0.014	0.031
	(0.065)	(0.059)	(0.056)	(0.055)	(0.049)	(0.064)	(0.059)	(0.052)	(0.070)
Audio*Education(cat)	-0.072	-0.105	-0.075	-0.177	-0.131	0.132	0.245+	0.163	-0.163
	(0.134)	(0.113)	(0.099)	(0.109)	(0.099)	(0.131)	(0.126)	(0.100)	(0.147)
Num.Obs.	983	1123	1123	1089	1086	1038	1068	1093	828
R2	0.016	0.077	0.017	0.035	0.141	0.021	0.040	0.015	0.074
R2 Adj.	0.013	0.075	0.015	0.032	0.139	0.018	0.038	0.013	0.070
AIC	3215.9	3560.9	3401.9	3304.4	3091.0	3451.8	3413.0	3219.2	2677.2
BIC	3240.3	3586.0	3427.0	3329.3	3116.0	3476.5	3437.8	3244.2	2700.8

Continued on next page

(continued)

	Climate Evaluation	Education	Future Children	Life Satisfaction	Most Important Problem	Political Trust	President	Generalized Trust	Vote Choice
Log.Lik.	-1602.936	-1775.439	-1695.956	-1647.192	-1540.524	-1720.892	-1701.491	-1604.610	-1333.586
RMSE	1.24	1.18	1.10	1.10	1.00	1.27	1.19	1.05	1.21
Socioeconomic Status									
(Intercept)	3.075***	2.891***	2.883***	3.089***	2.012***	3.036***	2.978***	3.341***	2.832***
	(0.140)	(0.129)	(0.124)	(0.120)	(0.108)	(0.138)	(0.129)	(0.115)	(0.153)
Audio	0.634*	1.172***	0.603**	0.492*	1.322***	0.562*	0.405+	0.443*	1.230***
	(0.272)	(0.232)	(0.215)	(0.219)	(0.203)	(0.267)	(0.245)	(0.210)	(0.294)
SES	0.017	0.038	0.055*	-0.002	0.012	0.030	-0.006	0.067**	0.060*
	(0.027)	(0.025)	(0.024)	(0.023)	(0.021)	(0.027)	(0.025)	(0.022)	(0.029)
Audio*SES	-0.058	-0.103*	-0.062	-0.009	-0.084*	-0.031	0.028	-0.039	-0.094+
	(0.052)	(0.044)	(0.040)	(0.042)	(0.039)	(0.052)	(0.048)	(0.040)	(0.056)
Num.Obs.	981	1117	1120	1084	1081	1035	1063	1089	824
R2	0.016	0.066	0.021	0.032	0.144	0.020	0.038	0.020	0.078
R2 Adj.	0.013	0.063	0.019	0.030	0.141	0.018	0.035	0.018	0.074
AIC	3210.7	3559.1	3388.4	3294.3	3075.8	3444.9	3401.0	3205.2	2661.1
BIC	3235.1	3584.2	3413.5	3319.3	3100.7	3469.6	3425.8	3230.1	2684.7
Log.Lik.	-1600.329	-1774.545	-1689.187	-1642.169	-1532.885	-1717.455	-1695.492	-1597.587	-1325.564
RMSE	1.24	1.18	1.09	1.10	1.00	1.27	1.19	1.05	1.21
Sex (Male)									
(Intercept)	3.216***	3.033***	3.128***	3.036***	2.127***	3.213***	3.045***	3.675***	3.206***
	(0.063)	(0.058)	(0.055)	(0.054)	(0.048)	(0.063)	(0.058)	(0.052)	(0.068)
Audio	0.271*	0.795***	0.401***	0.506***	0.916***	0.389**	0.574***	0.280**	0.732***
	(0.121)	(0.103)	(0.090)	(0.100)	(0.092)	(0.121)	(0.110)	(0.095)	(0.126)
Sex (Male)	-0.125	0.110	0.056	0.115	-0.133+	-0.072	-0.189*	-0.006	-0.175+
	(0.092)	(0.085)	(0.081)	(0.079)	(0.071)	(0.092)	(0.085)	(0.076)	(0.099)
Audio*Male	0.173	-0.337*	-0.255+	-0.157	-0.015	0.041	-0.117	-0.068	0.030
	(0.184)	(0.156)	(0.139)	(0.151)	(0.139)	(0.185)	(0.169)	(0.143)	(0.193)
Num.Obs.	978	1117	1116	1082	1080	1032	1062	1087	823
R2	0.017	0.063	0.019	0.032	0.144	0.020	0.045	0.012	0.075
R2 Adj.	0.014	0.060	0.017	0.030	0.142	0.017	0.043	0.009	0.072
AIC	3199.4	3557.1	3364.5	3283.6	3066.3	3436.2	3388.6	3206.8	2659.3
BIC	3223.9	3582.2	3389.5	3308.6	3091.2	3460.9	3413.5	3231.8	2682.9
Log.Lik.	-1594.722	-1773.549	-1677.228	-1636.821	-1528.161	-1713.089	-1689.316	-1598.412	-1324.644
RMSE	1.24	1.18	1.09	1.10	1.00	1.27	1.19	1.05	1.21
Survey Mode (others were present)									
(Intercept)	3.143***	3.084***	3.148***	3.079***	2.065***	3.178***	2.963***	3.678***	3.132***
	(0.046)	(0.043)	(0.041)	(0.040)	(0.036)	(0.047)	(0.043)	(0.038)	(0.050)
Audio	0.361***	0.648***	0.311***	0.455***	0.931***	0.417***	0.537***	0.264***	0.736***

Continued on next page

(continued)

	Climate Evaluation	Education	Future Children	Life Satisfaction	Most Important Problem	Political Trust	President	Generalized Trust	Vote Choice
	(0.093)	(0.079)	(0.071)	(0.077)	(0.071)	(0.094)	(0.087)	(0.072)	(0.097)
Survey Mode (others were present)	0.447+	-0.120	0.227	0.181	0.149	0.144	-0.213	-0.197	-0.266
	(0.268)	(0.229)	(0.228)	(0.216)	(0.193)	(0.245)	(0.230)	(0.206)	(0.317)
Audio*NotAlone	-0.380	0.178	-0.277	-0.241	-0.346	-0.209	0.047	-0.183	0.697
	(0.433)	(0.362)	(0.331)	(0.339)	(0.302)	(0.403)	(0.356)	(0.340)	(0.505)
Num.Obs.	983	1123	1123	1089	1086	1038	1068	1093	828
R2	0.018	0.060	0.018	0.032	0.141	0.019	0.037	0.014	0.074
R2 Adj.	0.015	0.058	0.015	0.030	0.138	0.016	0.035	0.011	0.071
AIC	3214.5	3581.0	3401.5	3307.6	3091.6	3453.4	3416.3	3220.7	2676.5
BIC	3238.9	3606.1	3426.6	3332.5	3116.5	3478.1	3441.2	3245.7	2700.1
Log.Lik.	-1602.238	-1785.508	-1695.726	-1648.784	-1540.799	-1721.691	-1703.147	-1605.340	-1333.249
RMSE	1.23	1.19	1.10	1.10	1.00	1.27	1.19	1.05	1.21
Survey Mode (public)									
(Intercept)	3.143***	3.084***	3.148***	3.079***	2.065***	3.178***	2.963***	3.678***	3.132***
	(0.046)	(0.043)	(0.041)	(0.040)	(0.036)	(0.047)	(0.043)	(0.038)	(0.050)
Audio	0.361***	0.648***	0.311***	0.455***	0.931***	0.417***	0.537***	0.264***	0.736***
	(0.093)	(0.079)	(0.071)	(0.077)	(0.071)	(0.094)	(0.087)	(0.072)	(0.097)
Survey Mode (others were present)	0.447+	-0.120	0.227	0.181	0.149	0.144	-0.213	-0.197	-0.266
	(0.268)	(0.229)	(0.228)	(0.216)	(0.193)	(0.245)	(0.230)	(0.206)	(0.317)
Audio*NotAlone	-0.380	0.178	-0.277	-0.241	-0.346	-0.209	0.047	-0.183	0.697
	(0.433)	(0.362)	(0.331)	(0.339)	(0.302)	(0.403)	(0.356)	(0.340)	(0.505)
Num.Obs.	983	1123	1123	1089	1086	1038	1068	1093	828
R2	0.018	0.060	0.018	0.032	0.141	0.019	0.037	0.014	0.074
R2 Adj.	0.015	0.058	0.015	0.030	0.138	0.016	0.035	0.011	0.071
AIC	3214.5	3581.0	3401.5	3307.6	3091.6	3453.4	3416.3	3220.7	2676.5
BIC	3238.9	3606.1	3426.6	3332.5	3116.5	3478.1	3441.2	3245.7	2700.1
Log.Lik.	-1602.238	-1785.508	-1695.726	-1648.784	-1540.799	-1721.691	-1703.147	-1605.340	-1333.249
RMSE	1.23	1.19	1.10	1.10	1.00	1.27	1.19	1.05	1.21

Table A3.8: Regression of Response Entropy and Respondent Characteristics. *Note*: Stars indicate signifcance levels +=.1, *=.05, **=.01, ***=0.001.

A.5 Exemplary Questionnaires Pages

Figure A3.3 shows response options for exemplary questionnaire pages.

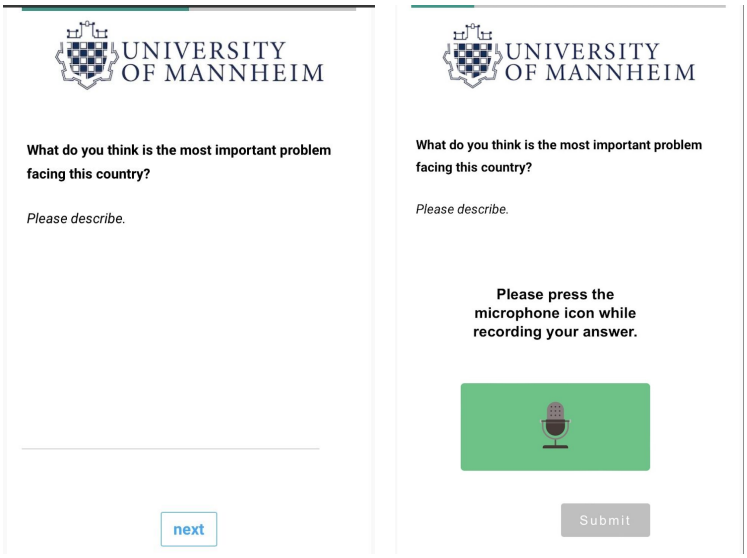

Figure A3.3: Exemplary questionnaire pages (audio and text request).

A.6 Results two-sample t-test

Answer Length

Below we show results from Welch two-sample t-test for the differences in Answer Length by Condition and Item. For 5 of our 9 OEQs we observe significantly longer answers for the audio than for the text format (p<.05) which in the main paper provides support for our first hypothesis (*H1*) (cf. Figure 3.1 in the main paper).

		Answer Length						
Item	Difference	Text	Audio	Statistic	p.value	Parameter	Conf.low	Conf.high
Climate Evaluation	-1.74	25	27	-1.56	0.12	559	-3.9	0.45
Education	-4.68	22	27	-4.43	0.00	790	-6.8	-2.61
Future Children	-6.75	17	23	-7.67	0.00	829	-8.5	-5.02
Life Satisfaction	-0.46	26	27	-0.37	0.71	940	-2.9	1.96
Most Important Problem	-3.01	18	21	-2.62	0.01	712	-5.3	-0.75
Political Trust	-4.36	24	29	-3.87	0.00	498	-6.6	-2.15
President	-1.70	24	26	-1.55	0.12	648	-3.8	0.45
Generalized Trust	-4.62	24	28	-4.45	0.00	617	-6.7	-2.58
Vote Choice	-1.59	25	26	-1.29	0.20	541	-4.0	0.83

Table A3.9: Average answer length and differences by condition and item. *Note*: Answer length: computed by counting the number of words. Method: Welch two-sample t-test.

Number of Topics

Below we show results from Welch two-sample t-test for the differences in Number of Topics by Condition and Item. For 9 of our 9 OEQs we observe significantly more topics in answers for the audio than for the text format (p<.05) which in the main paper provides support for our first hypothesis (*H1*) (cf. Figure 3.2 in the main paper).

		Number of Topics						
Item	Difference	Text	Audio	Statistic	p.value	Parameter	Conf.low	Conf.high
Climate Evaluation	-0.35	3.2	3.5	-4.2	0	491	-0.52	-0.18
Education	-0.66	3.1	3.7	-8.3	0	605	-0.81	-0.50
Future Children	-0.30	3.2	3.5	-4.3	0	765	-0.44	-0.16
Life Satisfaction	-0.45	3.1	3.5	-6.0	0	547	-0.59	-0.30
Most Important Problem	-0.91	2.1	3.0	-13.9	0	560	-1.04	-0.78
Political Trust	-0.41	3.2	3.6	-4.5	0	460	-0.58	-0.23
President	-0.53	3.0	3.5	-6.7	0	525	-0.69	-0.38
Generalized Trust	-0.25	3.7	3.9	-3.7	0	604	-0.39	-0.12
Vote Choice	-0.76	3.1	3.9	-8.5	0	444	-0.94	-0.59

Table A3.10: Average number of topics and differences by condition and item. *Note*: Number of Topics: computed using Structural Topic Models. Method: Welch two-sample t-test.

Entropy

Below we show results from Welch two-sample t-test for the differences in Entropy by Condition and Item. For 9 of our 9 OEQs we observe significantly more entropy in the answers for the audio than for the text format ($p<.05$) which in the main paper provides support for our first hypothesis (*H1*) (cf. Figure 3.3 in the main paper).

		Entropy						
Item	Difference	Text	Audio	Statistic	p.value	Parameter	Conf.low	Conf.high
Climate Evaluation	-0.12	4.2	4.3	-2.0	0.04	451	-0.24	0.00
Education	-0.33	3.8	4.2	-5.7	0.00	723	-0.44	-0.21
Future Children	-0.55	3.5	4.1	-9.8	0.00	990	-0.66	-0.44
Life Satisfaction	-0.11	4.1	4.2	-2.1	0.03	617	-0.22	-0.01
Most Important Problem	-0.65	3.2	3.8	-7.3	0.00	733	-0.83	-0.48
Political Trust	-0.21	4.1	4.3	-3.7	0.00	497	-0.33	-0.10
President	-0.19	4.0	4.2	-3.1	0.00	616	-0.32	-0.07
Generalized Trust	-0.21	4.1	4.3	-3.9	0.00	629	-0.31	-0.10
Vote Choice	-0.15	4.1	4.2	-2.3	0.02	456	-0.28	-0.02

Table A3.11: Average entropy and differences by condition and item. *Note*: Entropy: computed using Shannon Entropy. Method: Welch two-sample t-test.

4 Asking Why: Is there an Affective Component of Political Trust Ratings in Surveys?

Abstract

Our study explores the nature of political trust ratings in surveys by investigating the impact of affective rationales. This inquiry challenges the conventional notion that trustors rely on informed, rational, and consequential reasoning, suggesting instead that emotional states play a role. With a sample of approximately 1,500 respondents in the United States, we collect open-ended responses through smartphone microphones, prompting participants to articulate their thoughts during the response process. We classify these answers by leveraging methods from the fields of sentiment analysis as well as speech emotion recognition. Our findings, in terms of sentiment, reveal a high share of negative associations expressed by respondents when answering a question about political trust. The nature of these associations significantly influences trust scores, with positive associations positively impacting trust levels, and vice versa. For the spectrum of emotions conveyed in survey responses, a more detailed measure of the nature of associations, we observed limited variation with only one notable effect where respondents employing "happy" language and paralinguistics exhibited higher trust scores.

Keywords

sentiment analysis, affective computing, speech emotion recognition, BERT, pysentimiento, GPT, zero-shot prompting, political trust

4.1 Introduction

The study of determinants of political trust holds a longstanding position within the social sciences. This enduring interest is attributed for example to trust's influential role in various aspects of societies, including political participation such as voting, campaign involvement, and citizen compliance (Grönlund & Setälä, 2007; Levi & Stoker, 2000; Marien & Hooghe, 2011). Consequently, a substantial body of literature has emerged, further underscoring the importance of research on factors that shape political trust. In general, studying the process of trust judgments is not only inherently intriguing but in better understanding the factors influencing trust, we also shed light on how and why trust levels evolve over time.

Our study explores the nature of political trust ratings in surveys with regards to a specific element, namely the impact of emotional states on trust. Previous investigations on political trust are often characterized by the conventional notion of trust being rooted in informed, rational, and consequential judgments by trustors, who base their trust on their perceived knowledge about the trustee (e.g., politicians) accompanied by consequential reasoning. Our study challenges this prevailing "cognitive-based" approach of trust and follows scholarly debates that have emphasized that trust might include both, emotional and cognitive dimensions (e.g., Dunn & Schweitzer, 2005; Finucane et al., 2000; Lahno, 2020; Lee et al., 2023; Lewis & Weigert, 1985; McAllister, 1995; Midden & Huijts, 2009; Myers & Tingley, 2017; Theiss-Morse & Barton, 2017).

In particular, we are examining two research questions. First, we ask whether individual trust judgments in surveys include affective rationales. Second, we ask, whether the presence of affective responses is related to the strength of trust values.

In pursuing these questions, our study aims at making several contributions. First, we want to contribute to the ongoing debate about whether trust judgements are predominantly driven by rational reasoning or if they also include affective components. We will achieve this with the help of a survey where we first ask one of the most commonly used political trust questions (i.e., biennial ANES survey since 1964) and subsequently collect data on the response process using an

open-ended probing question.[47] With this probing question we ask respondents to describe (using their own words) how they came to their rating in the beforehand political trust question. These open-ended answers provide us with details on the answer process for the previous closed-ended decision, from which we can extract the extent and nature of affective components. In order to measure the affective component in these probing answers, we understand affect as "an umbrella term that is used to refer to both emotions and moods" (Lee et al., 2023, p.549) and measure sentiment (a simple negative versus positive feeling) as well as emotion (a more complex and multi-dimensional state of feeling).

Second, while previous research has already demonstrated the impact of emotional states on interpersonal trust judgments (e.g., Dunn & Schweitzer, 2005; Lee et al., 2023; Myers & Tingley, 2017), our study wants to extend this task to the domain of political trust. Assessing the role of emotions in political trust ratings might be crucial for our efforts to get a full understanding of political trust (Theiss-Morse & Barton, 2017, p.160) but this stream of research is still in its early stages and requires further empirical investigation (Theiss-Morse & Barton, 2017, p.167). Generally, this research objective seems promising, given that "politics" is considered as a profoundly emotional subject (Marcus, 2003).

Third, we want to make methodological contributions. The existing body of work on the determinants of political trust (for an overview see Schoon & Cheng, 2011) predominantly relied on closed-ended survey questions and regression-based analyses. In our study, we engage in direct inquiry, where we directly ask (i.e., probe) respondents to articulate reasons for their expressed level of political trust. For (political) trust items, probing has been carried out only a few times (Knudsen et al., 2021; Newman & Fletcher, 2017; Sturgis & Smith, 2010; Uslaner, 2002; Winsvold et al., 2023) and thus we want to contribute to this scarce literature.

Fourth, we instruct our participants to articulate their open-ended responses via speech input, requiring them to record their answer using their device's microphone within our web survey. Previous research has indicated that spoken re-

[47] Probing is a method that involves asking open-ended questions following closed-ended ones. The historical roots of probing in survey research lead back to the 1960s when Schuman introduced the concept of "random probes" in questionnaires (Given, 2008; Schuman, 1966). It has evolved over time, with modern developments such as verbal probing (Willis, 2004) now being a common practice in social science survey methodology (Behr et al., 2017; Neuert et al., 2021).

sponses tend to be more detailed, spontaneous, intuitive and that they show more extreme answers in terms of sentiment (Gavras et al., 2022). Our objective is to explore if audio responses offer insights into a respondent's emotional state. Additionally to analyzing the audio recordings from our respondents, we also transcribe these and analyze the transcriptions. By employing this multimodal approach to evaluate emotions in open-ended text responses, we hope to deliver a more complete picture into different aspects of emotions in such answers. The two approaches are achieved with state-of-the-art models for text and audio classification where we leverage deep-learning methods (i.e., SpeechBrain) (Ravanelli et al., 2021) and other transformer-based models (i.e., BERT and GPT-3.5).
In conclusion, this study's primary interest lies in uncovering affect-driven motivators for political trust that influence respondents' trust ratings. We hypothesize that a significant share of respondents, when providing trust ratings in surveys, employs judgment processes involving affective rationales. Answer processes that were guided by such affective rationales would stand in contrast to what is assumed by prevailing theoretical assumptions, namely that trust judgments are made on a cognitive and rational basis.

4.2 Theory, Empirical Evidence and Hypotheses

"Cognition-based" and "affect-based" trust

Generally, in trust research, a conventional notion prevails, in which trust originates from informed, rational, and consequential judgments. This "cognitive-based" approach to trust, in essence, suggests that individuals base their trust judgments on purposeful and thoughtful evaluations of objects that primarily take a cognitive form (Hooghe et al., 2012; Metzger & Flanagin, 2013). For example for political trust, trustors would increase or decrease their trust based on knowledge and information they think to have about a political entity, from which they then derive predictions about the trustee's future trustworthiness. Lahno (2020) traces this cognitive definition of trust back to the 1980s and describes how this notion has found its way into contemporary trust research by the predominance of using dilemma games to measure dyadic (i.e., interpersonal) trust. Trust games typify situations of individual rational decision making (Lahno, 2020) and the key expla-

nation of trusting behavior in these game-theoretical accounts, is often achieved by assumptions of rational choice theory where "the human agent will choose his actions rationally in the light of [...] aims" (Lahno, 2020, p.148). In sum, a cognitive-based understanding of trust aligns with the notion that trust judgments are made upon the basis of risk calculations and rational choice-making processes (Coleman, 1994; Hardin, 2002; Levi & Stoker, 2000).

However, in addition to this cognitive-based form of trust, others have suggested that there is another, a so-called "affect-based" type of trust and various authors have long emphasized that trust includes both emotional and cognitive dimensions (e.g., Dunn & Schweitzer, 2005; Finucane et al., 2000; Lahno, 2020; Lee et al., 2023; Lewis & Weigert, 1985; McAllister, 1995; Midden & Huijts, 2009; Myers & Tingley, 2017; Theiss-Morse & Barton, 2017). Grimmelikhuijsen for instance describes that "a decision to trust a government organization may [. . .] not always be conscious and/or rational" (Grimmelikhuijsen, 2012, p.57).

In affect-based trust the basis for trust (or distrust) lies in emotional ties (McAllister, 1995) and is further grounded in an individual's attributions concerning the motives for the trustee's behavior (McAllister, 1995, p.29). This conceptualization of trust seems to specify emotion and affect as a core part of the construct itself (Lee et al., 2023).

Our study employs this idea of affect-driven trust and is furthermore closely inspired by ideas, theory and empirical findings described in Lodge and Tabers "The Rationalizing Voter" (2013). Lodge and Taber argue that political judgment is driven by "affect-driven, dual-process modes of thinking and reasoning" (Lodge & Taber, 2013, p.2) that "account for when, how, and why thoughts, feelings and behavioral intentions come to mind automatically" (Lodge & Taber, 2013, p.17). The author's starting point is the limited capacity of an individual's working memory, which requires a highly selective retrieval process of information from long-term memory which also makes the "nature of the affective and semantic connections [. . .] critical" (Lodge & Taber, 2013, p.17). They then describe different propositions which can be summarized in a concept they call "Automatic Hot Cognition". Here, automaticity refers to the affect-driven, dual-process modes of thinking and reasoning that have developed over three decades in the cognitive sciences (Lodge & Taber, 2013, p.2). Central to such dual-process models is the

distinction between unconscious (system I) and conscious (system II) processing (Lodge & Taber, 2013, p.2). System 2 judgements are guided by automaticity in which the retrieval and processing of information is guided by affect. The authors also call such judgements "snap judgments" (Lodge & Taber, 2013, p.10). The term "hot cognition" refers to the idea that concepts (e.g. an idea, a group, a political entity) are instantly and without intentional control classified as either good or bad (Lodge & Taber, 2013, p.44), based on the integration of thoughts and feelings associated with one's conscious and unconscious assessments. The valence of concepts is thus retrieved from an associated affect, allowing the brain "to use affect as real-time information to promote quick, efficient, spontaneous responses" (Lodge & Taber, 2013, p.48).

Previous empirical evidence and hypotheses

Various studies have examined the influence of emotions on trust (for an overview of studies see Lee et al., 2023). Unfortunately, to date, the investigation of the role of emotions in trust decisions, mainly encompasses investigations about interpersonal trust and not political trust (cf. Grimmelikhuijsen, 2012; Theiss-Morse & Barton, 2017). Theiss-Morse and Barton (2017) for example lament that the vast majority of previous research on political trust takes a cognitive approach (measured in terms of the number of studies devoted to understanding cognitive processes versus those attending to affective or emotional reactions). The authors stress that "ignoring emotions is detrimental to our efforts to get a full understanding of political trust" (2017, p.160).

Still, in this section, we want to review empirical studies from research on the effect of emotions on social trust as they can provide useful approaches for generating hypotheses about emotions and political trust.

Dunn and Schweitzer (2005) explore the influence of emotions on interpersonal trust using survey experiments in which they experimentally induce emotions with the help of a writing task and subsequently measure dyadic trust. They find that moods with positive valence increase trust, while moods with negative valence decrease it. Here, the authors outline a theoretical framework based on so-called "mood models" (Dunn & Schweitzer, 2005, p.737). For example, the "affect-as-

information" (sometimes "feelings-as-information") model suggests that, when making trust judgments, people frequently attribute their current mood (i.e., positive or negative) to the judgement they are evaluating. Moreover, the authors looked beyond valence, by exploring the impact of specific emotions, and found that happy participants were more trusting than sad participants and that sad participants were more trusting than angry participants. They explain these differences based on the "cognitive nature" of the respective emotion. Anger, for example, they argue, is an emotion driven by "other-person control" (Dunn & Schweitzer, 2005, p.738) meaning that it is an emotion which is typically accompanied by attributing responsibility to others. For negative emotions with other-person control (e.g., anger), they argue to find a larger negative influence on trust compared to negative emotions with lesser other-person control (i.e., sadness). On the other hand, emotions of weak control appraisals, such as happiness, they argue, have a positive effect on trust.[48]

Other empirical accounts on the influence of emotions on trust apply similar cognitive appraisals, but they argue for "certainty appraisals" instead of "control appraisals" (e.g., Albertson & Gadarian, 2015; Myers & Tingley, 2017). Myers & Tingley (2017) for example investigate the effect of emotions on dyadic trust and find that negative emotions can decrease trust, but only if negative emotions make people feel less certain about their current situation (e.g., anxiety). In contrast to results from Dunn & Schweitzer, they find that emotions with strong other-control appraisal, but with high certainty, such as anger, have no significant effect on trusting behavior. Guilt, an emotion of equally high certainty but weak other-control appraisal also has no large effect on trust in their sample. Put differently, they argue that irrespective of the extent of other-person control, emotions that exhibit high certainty (e.g., anger) should have no large impact on trusting behavior, compared to emotions of low certainty (e.g., anxiety). Apart from this substantially different argumentation, the authors provide alternative explanations about why their results with regards to other-control appraisals differ from the ones in previous research (e.g., Dunn & Schweitzer, 2005). For example, they highlight potential differences due to different measurement approaches, such as trust sur-

[48]The positive effect of happiness on trust can be found in other studies as well, for example in (Mislin et al., 2015).

vey items in previous studies compared to their approach of utilizing trust games. In sum, different theoretical arguments about the relationship of emotions on trust provide competing claims concerning which elements of a person's emotional state impact trust (Myers & Tingley, 2017). However, what unites these studies is their distinction of an emotion's valence (a "mood") and the actual emotion. Lee et al. (2023) describe affect as an "umbrella term that is used to refer to both emotions and moods" (Lee et al., 2023, p.549) however both, moods and emotions differ in various aspects.

A mood only encompasses a valence that represents whether there is a negative, positive or neutral affective state (i.e., sentiment) and neglects any specific details or the nuances of emotions experienced. In contrast, emotional states (or short emotions), are shorter in duration and more intense (Dunn & Schweitzer, 2005, p.737). Moreover, emotions are driven by a number of different cognitive evaluations (e.g., other-person control), which makes them more complex than moods (Smith & Ellsworth, 1985). These cognitive appraisals allow predictions about how different emotions might have different effects on trust: in the control appraisals framework, the prediction is that other-control emotions such as anger will diminish trust, whereas in the certainty appraisals framework, the prediction is that low-certainty emotions such as anxiety will decrease trust (Myers & Tingley, 2017).

In line with this previous research, in the following we will analyze both, moods (i.e., sentiment) and particular emotions. In doing so, we hope to provide a more complete picture, while offering a multi-dimensional understanding of the survey answers and their affective components. Our first hypothesis aims at investigating whether trust judgements in surveys contain affect and we argue in line with the previously outlined research that argues in favor of affect-based trust.

| **H1**: A significant share of respondents, when requested to give reasons for their level of trust, include non-neutral sentiment (**H1a**) as well as emotional language (**H1b**) in their answers.

In our second hypothesis we aim to explore the link between these affective rationales and actual trust ratings. In line with previous research that investigated the impact of emotions on trust using survey items, we adopt an approach that acknowledges the valence (i.e., positive or negative) of an emotion as well as its

"cognitive nature". In particular, following Dunn & Schweitzer (2005), we assume that both sadness and anger due to their negative valence have a negative effect on political trust. However, due to the idea that anger is an emotion characterized by other-person control, again following Dunn & Schweitzer (2005), we assume that the effect of anger on trust is larger than that of sadness.[49] Happiness due to its positive valence should have a positive effect on trust.

| **H2**: The sentiment or emotion expressed in an open-ended survey answer is correlated with the closed-ended trust score. Specifically, we expect sentiment of positive valence to increase trust and sentiment of negative valence to decrease trust (**H2a**). Further on, we expect the emotion of happiness to increase trust, anger to decrease trust and sadness, according to its negative valence to also decrease trust but to a much lesser extent than anger (**H2b**).

4.3 Methods

4.3.1 Data and Questionnaire

The survey was conducted from September 6 to October 27, 2023. The sample for our study was drawn with quotas aligned with 2015 U.S. Census Bureau estimates for gender, age, and ethnicity. Out of 35,153 eligible participants, we collected data from 1,431 individuals. Among them, 155 were screened-out due to non-completion. Consequently, the final sample size of this study comprises 1,276 respondents who successfully completed the survey, resulting in a break-off rate of 11% (American Association for Public Opinion Research (AAPOR), 2016; Callegaro & DiSogra, 2008).

Since this study faced challenges in obtaining sufficient participants in the oldest age category (58+), we continued collecting additional data only allowing for participants aged 58 and above to participate in our study, without imposing additional quotas (i.e., gender, ethnicity).[50] We collected another n=202 (n=216

[49] Despite the mixed findings described above (certainty versus control appraisals), we follow Dunn & Schweitzer (2005), because similarly to us they used survey measures, as opposed to other measures, for example trust games. Myers & Tingley (2017, p.7) explain that these two techniques possibly measure different constructs, surveys are better able to measure trustworthy behavior and not necessarily trusting behavior.

[50] This disparity might be attributed to several factors. Primarily, the usage of a smartphone, a requirement for participation in our study, is widely acknowledged to be more prevalent among

participants, n=14 break off) observations, excluded four responses from participants who reported technical issues with our survey, resulting in a final sample size of n=1,474.

Participants were recruited through the recruitment platform Prolific (Palan & Schitter, 2018). The average time to complete the questionnaire was 7.5 minutes (Mdn = 5.7) and was compensated with an average wage of 12$/hr.

The topic of the survey was "Politics and Trust". After a brief trial question where respondents could practice using the voice recording tool, we asked one of the most commonly used closed-ended survey question about political trust (i.e., biennial ANES survey since 1964, Citrin & Stoker, 2018).

| *"How often can you trust the federal government in Washington to do what is right?"*

| *Always – Most of the time – About half of the time – Never – Don't Know*

On the next questionnaire page, respondents received an open-ended follow-up question where we also repeated the question wording and the previously given answer, for example:

| *"The previous question was: 'How often can you trust the federal government in Washington to do what is right?'. Your answer was: 'About half of the time'. In your own words, please explain why you selected this answer."*

This strategy of using open-ended category-selection probing questions and the particular wording of the probe aims at offering respondents an opportunity to openly reflect upon their answer process and ideally uncover reasons for their previous closed-ended decision. The aim of this question was to uncover thought processes and associations that the respondents had while answering the closed-ended questions, from which we could then derive the sentiment and emotionality. Respondents could skip the open-ended question but they couldn't go back to previous questions. In total, we collected 491 open-ended voice responses that we can analyze.[51] This corresponds to a response rate of 33% to the audio open-

younger demographics. Additionally, the behavior of responding through voice recordings, as required in our study, is a behavior commonly observed among individuals, for instance, in the form of voice messaging.

[51]This only includes files of recorded answers exceeding a file size of 110KB which corresponds to approximately 2 seconds of content (this is a requirement by the speech-to-text algorithm). Additionally, we excluded voice answers that were longer than 2 seconds, but had no content or only contained random sounds (n=3). Additionally, we identified another answer (n=1) that

ended question. The tool used to record voice answers is SurveyVoice (Svoice) (Höhne et al., 2021). The recorded answers were automatically transcribed using Whisper from OpenAI (Radford et al., 2022).

4.3.2 Analytical Strategy

We analyze the open-ended answers in terms of whether and to what extent they exhibit affective components, including sentiment as well as emotions. To detect sentiment, we are analyzing information from the transcribed answers whereas for detecting emotions we will consider the raw, originally spoken answers (i.e., the audio files). Put differently, we not only use traditional text-based methods for Sentiment Recognition but also Speech Emotion Recognition (SER). Both tasks have the objective to take as input a spoken or written sentence, and output a classification of the expressed sentiment or emotion, such as neutral, negative, positive, happy, sad, and more.
Figure 4.1 illustrates the various steps in our workflow and the different approaches chosen to detect sentiment and emotions.

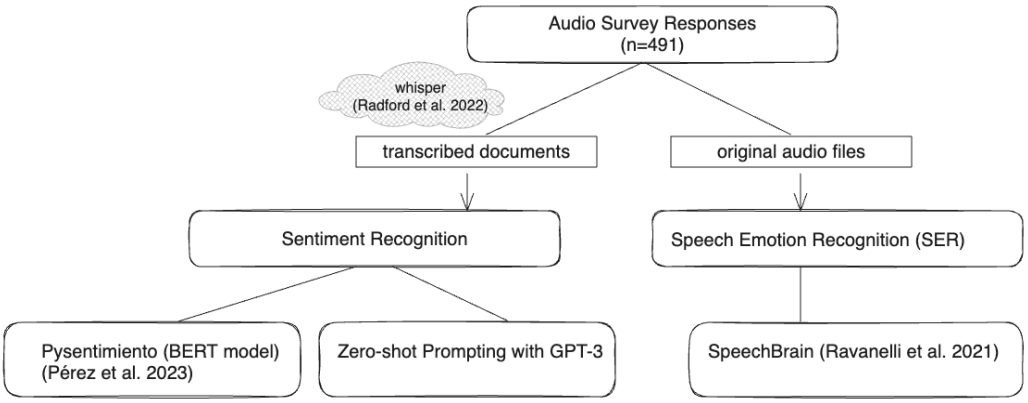

Figure 4.1: Methods for Sentiment and Emotion Analysis in the context of our Analytical Strategy. *Note*: Our Appendix additionally shows results where emotions were classified with transcribed documents using the NRC word-emotion lexicon (Appendix A.3). Appendix A.2 shows findings for a sentiment classification based on the transcribed documents using different dictionary approaches (AFINN and VADER).

contained non-sensical content and removed it.

Below, we describe in detail how these different steps depicted in Figure 4.1 are achieved.

Sentiment Analysis

We start by analyzing the sentiment of the survey responses, classifying them into positive, negative, or neutral tones. In this study we are exploring two approaches for this task, the first one is pysentimiento, a fine-tuned BERT model and the second approach is zero-shot prompting with GPT-3.5. While both approaches constitute a deep-learning architecture, they for example differ with regards to whether they are fine-tuned, that means specifically trained for the sentiment task or not.

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) is a language model with weights that contain contextual word representations (i.e., word embeddings), which can be fine-tuned for diverse natural language processing tasks. Fine-tuning is the process of "refining" a pre-trained model on a smaller task-specific dataset to learn a specific downstream task (e.g., sentiment classification). Nowadays, several pre-trained BERT models are available (Chiorrini et al., 2021) and in this study we are using a fine-tuned BERT model called pysentimiento (Pérez et al., 2023). Pysentimiento is built on top of Hugging Face's Transformers library (Wolf et al., 2019) and represents a Python multilingual toolkit for Social NLP tasks. Pysentimiento's models for Sentiment and Emotion Analysis, developed by Pérez et al. (2023), are deployed on the Hugging Face hub and can be easily accessed through the library. In particular, in our study, we are using the available sentiment model which uses BERTweet (a RoBERTa model trained on English tweets) as it's base model and was further trained with a dataset, which contains 61,900 tweets annotated for polarity detection (SemEval 2017 Task 4 Subtask 1, see Pérez et al., 2023 for details).

Our second approach to detect sentiment in the survey responses does not depend on fine-tuning but instead leverages the power of language models via zero-shot prompting. In general, prompting is a modern classification approach that came into being through the emergence of large language models such as GPT-3. Prompting can be further dissected into zero-shot or few-shot prompting. Both ap-

proaches were shown to match or even exceed the performance of typical human coders for a variety of classification tasks (Burnham, 2023; Rytting et al., 2023). Zero-shot prompting is an approach that makes use of the “instructions contained in prompts without any training data" (Chae & Davidson, 2023, p.3). Similarly to sentiment dictionaries, zero-shot prompting “requires no training data and minimal programming to implement” (Burnham, 2023, p.2), however, “[u]nlike sentiment dictionaries, it produces results comparable to, and sometimes better than, supervised classification” (Burnham, 2023, p.2). In our study, we are going to pursue zero-shot prompting with GPT-3.5-turbo, the successor to GPT-3, which, to date, is one of the largest existing language models (Brown et al., 2020). To achieve the classification of survey responses with GPT without providing human-annotated examples, we needed to provide the model with a specific prompt. Here, our goal was to provide a very clear and straightforward prompt which returns a single token (i.e, positive, negative, neutral) to detect the nature of associations in its most standard form: positive, negative or neutral. Our prompt reads as follows (see Appendix A.4 for more details on our prompt engineering):

“Classify the sentiment of the following open-ended survey answer into neutral, negative or positive. Text: *i* Sentiment:”, where *i* is the survey response.

With both the BERT and the GPT approach described above, we chose to pursue deep learning based models due to the finding that other, more traditional and more simple approaches (i.e., dictionary approaches) exhibited very low accuracy in our sample. Historically, the analysis of sentiment in survey responses was achieved with theory-based dictionaries of affectively scored words (Mosshholder et al., 1995). Results where we applied such dictionary methods to our data alongside a discussion of issues and challenges can be found in Appendix A.2.

Emotion Analysis

The detection of emotions can be understood as a more difficult endeavor compared to sentiment analysis, “given the greater variety of classes and the more subtle differences between them” (Chiorrini et al., 2021, p.1). Automatic emotion recognition is a young research area, but we already know that recognizing emotions from text is challenging (for an overview of transformer models for emotion

recognition from text see Nandwani and Verma, 2021, Kratzwald et al., 2018 and Adoma et al., 2020 as well as Appendix A.3 for an analysis with the EmoLex Dictionary) because it is stripped of all paralinguistic and acoustic features. We can assume that the inclusion of paralinguistic features such as intonation, pitch, volume, pauses, but also laughter or breathing noises (Lu et al., 2019) can be very helpful in recognizing emotional states in speech. This is why, in our study, for the task of emotion recognition we are going to use methods from the field of Speech Emotion Recognition.

Speech Emotion Recognition (SER) can be subordinated to the field of Automated Emotion Recognition and is a part of other disciplines such as Neurocomputing (de Lope & Graña, 2023) and Affective Computing (Atmaja et al., 2022; Madanian et al., 2023). Speech Emotion Recognition can include traditional machine learning (for example, logistic regression) (Madanian et al., 2023) but nowdays a large corpus of SER applications make use of deep learning architectures (de Lope & Graña, 2023; Singh & Goel, 2022).

To achieve emotion detection in our survey answers, our study follows this trend of utilizing deep learning and in particular we utilize SpeechBrain (Ravanelli et al., 2021). SpeechBrain is an open-source deep learning toolkit built upon PyTorch specifically designed for speech and audio processing tasks. In general, SpeechBrain provides a comprehensive set of pre-built tools ("recipes") and models tailored for various speech-related tasks, including Speech Emotion Recognition.

For emotion detection, Wang, Boumadane and Heba (2021) pre-trained a model which is based on the wav2vec 2.0 architecture. Wav2vec 2.0 (Baevski et al., 2020) originally has been proposed for Automatic Speech Recognition, but is nowadays also frequently used for speech emotion recognition (Chen & Rudnicky, 2021; Pepino et al., 2021; Wang et al., 2021). By training a wav2vec 2.0 model with data from the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, the authors achieve state-of-the-art results for different tasks.[52] For

[52]The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset comprises approximately 12 hours of annotated recordings, encompassing dialogues from 10 speakers. To capture the emotional content of databases six human evaluators assessed the emotional categories of the database (three per utterance) (Busso et al., 2008). To date, IEMOCAP is one of the three most widely used and most representative datasets for SER (Wang et al., 2021, p.3)

example, for the task of SER using the IEMOCAP public emotion recognition dataset they achieved a weighted (averaged on 5 different seeds) accuracy rate of 79.58% across multiple emotions (model "PF-hbt-large").[53] Similar performances for fine-tuned wav2vec 2.0 models for SER can be found in Chen and Rudnicky (2021) (e.g., 74.3% unweighted accuracy for a wav2vec 2.0 model fine-tuned using a P-TAPT method). Wang et al. (2021) open-sourced their code within the SpeechBrain framework which allows for easy implementation for researchers without extensive training resources.

SpeechBrain's recipe for Speech Emotion Recognition classifies utterances into four emotions: anger, happiness, sadness and neutrality (Wang et al., 2021, p.4). This decision of including four emotions is frequently found in SER research utilizing the IEMOCAP database and aims at ensuring consistency and comparability between different other studies on the same database (Fayek et al., 2017). In general, research that includes data of linguistic, acoustic as well as paralinguistic features is still in its infancy. We are not aware of any studies that use the SpeechBrain emotion classifier[54], and social science research that takes advantage of such detailed language characteristics of language is not yet widespread. Noteworthy exceptions and applications lie within the field of political science (e.g., Dietrich et al., 2019; Rittmann, 2023).

Our hypotheses will be investigated as follows. For our first hypothesis (**H1**) ("Respondents, when requested to give reasons for their level of trust, include non-neutral sentiment (**a**) as well as emotional language (**b**) in their answers."), we will examine the open-ended responses to determine which types of sentiment and emotion they include. As described above, sentiment will be measured with three categories (positive, negative, neutral) and emotions will be measured with four categories (anger, happiness, sadness and neutrality).

Our second hypothesis (**H2**), that the type of sentiment (**a**) or emotion (**b**) expressed in an open-ended survey answer impacts the closed-ended trust score, is analyzed by predicting the quantitative closed-ended measure of political trust with our sentiment and emotion variables. For this we will use regression anal-

[53] PF-hbt-large = Partially Fine-Tuned HuBERT based model Large Pre-Trained (Speaker-dependent setting) model. For a competing model named PF-w2v-large (Speaker-dependent) they achieve a weighted accuracy of 77.47 (Wang et al., 2021, p.4).

[54] A benchmark study that evaluates the SpeechBrain Classifier can be found in Vu et al. (2022).

ysis and show findings for a model that includes common covariates to explain political trust (age, education, and socioeconomic status).

4.4 Results

We start by presenting the results regarding the sentiment in the open-ended answers. Figure 4.2 illustrates the distributions over the three sentiment categories by classification approach. In sum, both classifiers result in a notable portion of open-ended answers with negative sentiment. From our total of 491 observations, the two classifiers classify between 59% (BERT) and 62% (GPT) as negative and only between 30% (GPT) and 33% (BERT) as neutral. The share of positive associations in the survey responses is small for both classifiers (8% for GPT and BERT).

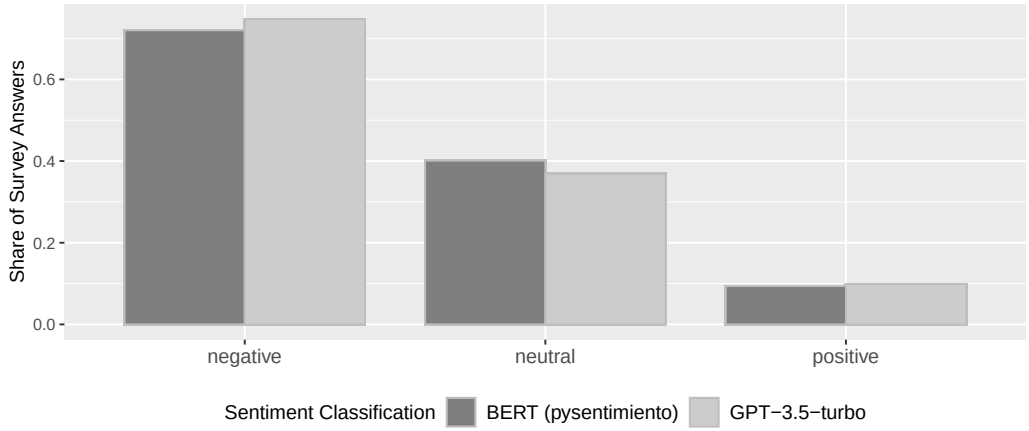

Figure 4.2: Sentiment Classification with three categories by classifier (BERT vs. GPT). *Note*: n=491 open-ended answers. BERT represents findings from sentiment analysis achieved with pysentimentio (Pérez et al. 2023) and GPT represents findings from zero-shot prompting with GPT-3.5-turbo.

Importantly, both classifiers result in similar results and exhibit a degree of agreement of 79%. Appendix A.1 depicts the confusion matrix of both classifiers and further demonstrates how they differ in terms of accuracy when we compare their results to a human-labeled subset of the data. Here, the GPT classifiers exhibited a slightly better accuracy compared to the BERT classifier.

Next, we delve into this sentiment variability and its impact on actual trust scores. For this, we conducted regression analysis, including both bivariate and multivariate models. Bivariate analysis examines the direct effect of sentiment on political trust, while multivariate analysis adjusts for potential confounding factors (age, education, and socioeconomic status) enhancing the robustness of our findings. Figure 4.3 illustrates the results from these regression models.

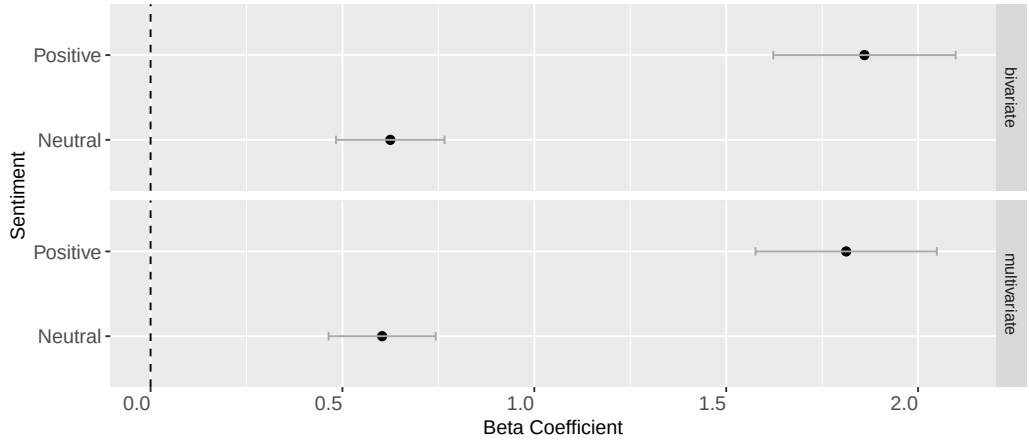

Figure 4.3: Linear model of sentiment and a five-category trust score (bi- and multivariate). *Note*: The data for the sentiment classification displayed in this figure stems from the GPT-based classification. Respective findings for the BERT classification can be found in Appendix A.1. Results from an ordered logit model can be found in Appendix A.1.

For computing the above displayed coefficients, we designated the "negative" sentiment as the reference category and consequently Figure 4.3 illustrates that irrespective of model specification, an increase in positive sentiment is associated with an increase in political trust (beta_positive = 1.8, $p<0.001$ and beta_neutral = 0.6, $p<0.001$ in the multivariate setting). This finding gives support for H2a suggesting that the nature of associations measured by sentiment in a survey response is correlated with trust scores.

We continue with our findings for the presence of emotions in the survey answers. Figure 4.4 displays emotion classification results obtained through audio-based analysis using SpeechBrain. In sum, Figure 4.4 demonstrates that the majority of open-ended responses do not contain explicit affective speech characteristics and

are characterized by a neutral emotional tone which does not support H1b, and thus indicates that there is no significant share of emotions in our survey answers.

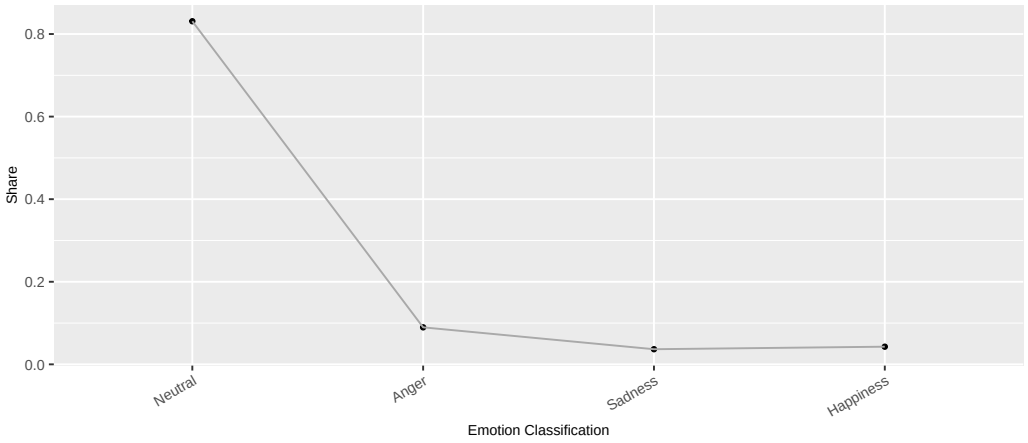

Figure 4.4: Emotion Classification obtained from SpeechBrain. *Note*: Analysis of n=491 open-ended answers. Number of observations for each sentiment category: 408 (neutral), 44 (angry), 18 (sad), 21 (happy).

We continue with a regression analysis to examine the impact of emotion categories on trust scores. Figure 4.5 shows findings from this linear regression (bi- and multivariate) where the "neutral" emotion category was used as the reference category.

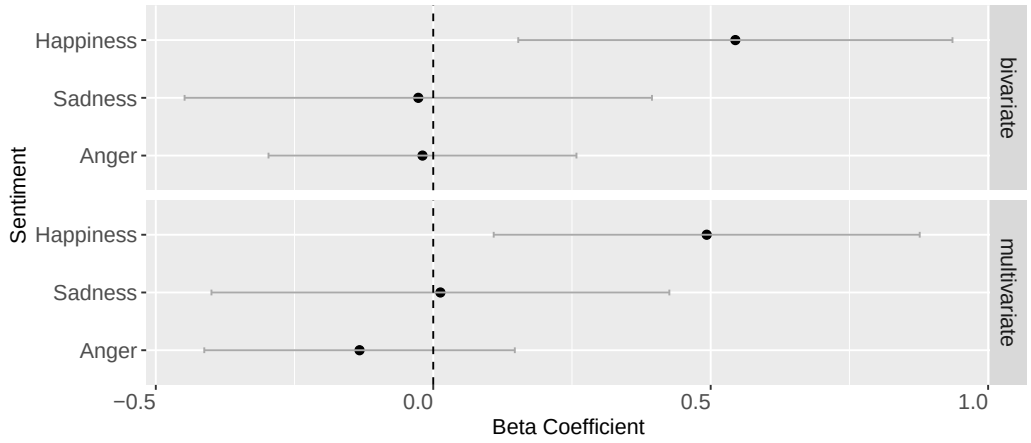

Figure 4.5: Linear model of emotion and a five-category trust score (bi- and multivariate). *Note*: The data for the emotion classification displayed in this figure stems from the SpeechBrain classification.

Figure 4.5 shows that irrespective of covariate specification, a positive emotion of happiness positively influences trust (for example, beta_happiness = 0.49, p<0.1 in the multivariate setting). The effect of negative emotions such as sadness and anger is negative, however not significant (beta_sadness = 0.013, p>0.1, beta_anger = -0.13, p>0.1 in the multivariate setting, beta_anger = -0.019, p>0.1 in the bivariate setting).
These findings are partly in line with our hypothesis H2b stating that the nature of associations measured by emotions in a survey response has an impact on trust scores, however we were only able to find a significant and meaningful effect for the positive emotion in our sample.

4.5 Discussion and Conclusion

Our study examined whether and to what extent trust judgments involve affective reasoning. The theoretical model we presented proposes different bases about how individuals create trust judgments. Recent research has suggested that additionally to a "cognitive route", there might be an "affective route" to explaining trust (Grimmelikhuijsen, 2012, p.56). In line with previous scholars (e.g., Dunn & Schweitzer, 2005; Finucane et al., 2000; Lahno, 2020; Lee et al., 2023; Lewis & Weigert, 1985; McAllister, 1995; Midden & Huijts, 2009; Myers & Tingley, 2017;

Theiss-Morse & Barton, 2017), we argued that individuals employ an affective decision making process and thereby challenged the idea of a fully informed, rational individual that makes trust judgments based on risk calculations and rational choice-making processes (Coleman, 1994; Hardin, 2002; Levi & Stoker, 2000). Our central ideas were grounded in existing findings from the cognitive sciences, thereby touching literature about affective components in political behavior and judgment such as "hot cognition" originally outlined in Lodge and Taber (Lodge & Taber, 2013).

To shed light on the nature of trust judgments, we employed a data generation approach that involves response probing, i.e. the use of open-ended questions. In contrast to previous covariate-adjusted regression approaches, our probing design "start[s] inductively by directly asking people what comes to their mind when they think about trust [...] without stipulating anything beforehand" (Knudsen et al., 2021, p.4). The goal of our probing question was to elicit the thought processes and the associations respondents had when making their trust judgment. We collected data from a sample of approximately n=1,500 respondents from the United States and requested respondents to give open-ended answers via speaking into the microphone of their smartphone. Eventually, we employed a multimodal analysis, where we analyze the transcripts of the audio files as well as the original audio files. Findings were obtained from methods from sentiment and emotion recognition in order to predict the sentiment and emotional content in a survey answer (e.g., happy, sad, neutral, and angry) using deep-learning based methods.

In our analysis, we first found a significant share of non-neutral sentiment for a question about political trust which is predominantly negative (59% of negative sentiment for our BERT classifier and 62% for our GPT classifier, see Figure 4.2). In contrast, we detected only very few instances of positive sentiment (8% for both BERT and GPT). Furthermore, we found that the valence of these open-ended survey responses (e.g., positive, negative) have a strong influence on a 5-point trust scale (e.g., beta_positive = 1.8, $p<0.001$ in a multivariate setting, see Figure 4.3). Additionally, we analyzed the emotions given in the survey responses. Emotions can be considered a more detailed and insightful way of capturing affect and we achieved this task by analyzing the original audio answers in order to include paralinguistic features of speech (i.e., pitch, intonation, etc). In sum, we found very

small shares of emotional language in the audio survey answers (see Figure 4.4). The overall presence of emotions (i.e., happy, sad, angry) was at 17% compared to a share of 83% of answers with no explicit emotion (neutral answers). We found that the positive emotion of happiness in our sample has a positive effect on political trust (beta_happiness = 0.49, $p<0.1$ in a multivariate setting, see Figure 4.5). Negative emotions such as sadness and anger do not seem to have an effect on trust scores in our sample. This latter finding (i.e., no effect of sadness and anger on trust) contradicts our hypothesis. We can imagine that, while politics and trust are often regarded as emotionally charged or rich in affect (Dunn & Schweitzer, 2005; Marcus, 2003), a formal survey setting may not be the most effective way to elicit emotionally charged paralinguistics. The requirement for respondents to input their answers through speech, especially in potentially uncomfortable settings with the presence of others, may contribute to this limitation. Conversely, in the sentiment task, we observed a substantial range of sentiment. This might be attributed to the fact that respondents possibly do not express strong paralinguistics in their responses, yet convey their emotions through the careful selection of words. For instance, words like "lovely" and "awesome" inherently carry stronger emotions compared to more generic, non-emotional terms like "person" and "day" (Yoon et al., 2018, p.1).

Research about the question whether affective information serves as a component of political trust ratings is still in its early stages and requires further empirical investigation (Theiss-Morse & Barton, 2017, p.167). Our study yields various connection points to obtain a more detailed and robust understanding of the affective components in trust survey measures.

First, exploring other probing wordings that aim at more general descriptions of experiences and associations related to politics (for example, how have you felt about politics lately?), could have potentially yielded stronger presence of emotions in answers.

Second, our study faced challenges in terms of relatively small response rates for the open-ended audio question (i.e., 33%). A higher response rate would have significantly enriched the depth and generalizability of our findings. Future studies could explore strategies to increase participation with this format (e.g., incentives, user experience).

Third, our paper's utilization of Natural Language Processing (NLP) techniques to analyze open-ended responses demonstrated the methodological possibilities (e.g., the use of deep learning) for such research questions. The rapid evolution in this field continually introduces new opportunities and promising prospects for enhancing (Speech) Emotion Recognition. Regarding that "[d]escribing emotion is an inherent complex problem" (Busso et al., 2008, p.347), one avenue is to argue that its adequate depiction should probably include both sound and spoken (i.e., the words used) content. Future studies hopefully yield multimodal models, as "different modalities contain different information, and all are slightly flawed, so multi-modal based information can better identify the speaker's emotion than a single modality in ERC" (Li et al., 2022, p.1). Recent advancements have shown promising ways to effectively and simultaneously use both audio and text data for emotion recognition in speech (Li et al., 2022; Yoon et al., 2018). Broadly speaking, the application of these techniques demonstrates the potential of NLP methods for understanding and extracting valuable information from unstructured data. Affective computing, a multidisciplinary research field involving engineering, psychology, education, cognitive science, sociology, and more, could significantly benefit from the developments in these fields (Daily et al., 2017). We strongly encourage future studies to leverage the rich information embedded in text and audio responses while exploring the latest methodological toolkits available.

References

Adoma, A. F., Henry, N.-M., & Chen, W. (2020). Comparative analyses of bert, roberta, distilbert, and xlnet for Text-Based emotion recognition. *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 117–121.

Albertson, B., & Gadarian, S. K. (2015, August). *Anxious politics: Democratic citizenship in a threatening world*. Cambridge University Press.

American Association for Public Opinion Research (AAPOR). (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys (9th ed*. Oakbrook Terrace, IL: AAPOR.

Atmaja, B. T., Sasou, A., & Akagi, M. (2022). Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Commun.*, *140*, 11–28.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for Self-Supervised learning of speech representations. *arXiv preprint:2006.11477*.

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). Web probing – implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions. *Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines)*.

Bello, A., Ng, S.-C., & Leung, M.-F. (2023). A BERT framework to sentiment analysis of tweets. *Sensors*, *23*(1).

Bharti, S. K., Varadhaganapathy, S., Gupta, R. K., Shukla, P. K., Bouye, M., Hingaa, S. K., & Mahmoud, A. (2022). Text-Based emotion recognition using deep learning approach. *Comput. Intell. Neurosci.*, *volume 2022 (special issue)*, 8 pages.

Blaikie, N. (2003). *Analyzing quantitative data*. SAGE Publications Ltd.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are Few-Shot learners.

Burnham, M. (2023). Stance detection with supervised, Zero-Shot, and Few-Shot applications. *arXiv preprint:2305.01723*.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.*, *42*(4), 335–359.

Callegaro, M., & DiSogra, C. (2008). Computing response metrics for online panels. *Public Opin. Q.*, *72*(5), 1008–1032.

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. *J. Soc. Sci.*, *3*(3), 106–116.

Chae, Y., & Davidson, T. (2023). Large language models for text classification: From Zero-Shot learning to Fine-Tuning. *OSF, osf.io/5t6xz*.

Chen, L.-W., & Rudnicky, A. (2021). Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Chiorrini, A., Diamantini, C., Mircoli, A., & Potena, D. (2021). Emotion and sentiment analysis of tweets using BERT. *EDBT/ICDT Workshops*.

Citrin, J., & Stoker, L. (2018). Political trust in a cynical age. *Annu. Rev. Polit. Sci.*, *21*(1), 49–70.

Coleman, J. S. (1994). *Foundations of social theory*. Harvard University Press.

Daily, S. B., James, M. T., Cherry, D., J. Porter, J., Darnell, S. S., Isaac, J., & Roy, T. (2017, January). Chapter 9 - affective computing: Historical foundations, current applications, and future trends. In M. Jeon (Ed.), *Emotions and affect in human factors and Human-Computer interaction* (pp. 213–231). Academic Press.

de Lope, J., & Graña, M. (2023). An ongoing review of speech emotion recognition. *Neurocomputing*, *528*, 1–11.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint: 1810.04805*.

Dietrich, B. J., Hayes, M., & O'brien, D. Z. (2019). Pitch perfect: Vocal pitch and the emotional intensity of congressional speech. *Am. Polit. Sci. Rev.*, *113*(4), 941–962.

Dunn, J. R., & Schweitzer, M. E. (2005). Feeling and believing: The influence of emotion on trust. *J. Pers. Soc. Psychol.*, *88*(5), 736–748.

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Netw.*, *92*, 60–68.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, *13*(1), 1–17.

Gavras, K., Höhne, J. K., Blom, A. G., & Schoen, H. (2022). Innovating the collection of open-ended answers: The linguistic and content characteristics of written and oral answers to political attitude questions. *J. R. Stat. Soc. Ser. A Stat. Soc.*, *tba*(tba), 1–19.

Given, L. M. (2008). *The sage encyclopedia of qualitative research methods*. SAGE.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Rev. Educ. Res.*, *42*(3), 237–288.

Grimmelikhuijsen, S. (2012). Linking transparency, knowledge and citizen trust in government: An experiment. *International Review of Administrative Sciences*, *78*(1), 50–73.

Grönlund, K., & Setälä, M. (2007). Political trust, satisfaction and voter turnout. *Comparative European Politics*, *5*(4), 400–422.

Hardin, R. (2002, March). *Trust and trustworthiness*. Russell Sage Foundation.

Höhne, J. K., Gavras, K., & Qureshi, D. (2021). SurveyVoice (SVoice): A comprehensive guide for collecting voice answers in surveys. zenodo. available from: Https://doi.org/10.5281/zenodo.4644590.

Hooghe, M., Marien, S., & de Vroome, T. (2012). The cognitive basis of trust. the relation between education, cognitive ability, and generalized and political trust. *Intelligence*, *40*(6), 604–613.

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious Rule-Based model for sentiment analysis of social media text. *ICWSM*, *8*(1), 216–225.

Knudsen, E., Dahlberg, S., Iversen, M. H., Johannesson, M. P., & Nygaard, S. (2021). How the public understands news media trust: An open-ended approach. *Journalism*.

Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., & Prendinger, H. (2018). Deep learning for affective computing: Text-based emotion recognition in decision support. *Decis. Support Syst.*, *115*, 24–35.

Lahno, B. (2020). Trust and emotion. In *The routledge handbook of trust and philosophy* (1st Edition, pp. 147–159). Routledge.

Lee, J. I., Dirks, K. T., & Campagna, R. L. (2023). At the heart of trust: Understanding the integral relationship between emotion and trust. *Group & Organization Management*, *48*(2), 546–580.

Levi, M., & Stoker, L. (2000). Political trust and trustworthiness. *Annu. Rev. Polit. Sci.*, *3*(1), 475–507.

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Soc. Forces*, *63*(4), 967–985.

Li, Z., Tang, F., Zhao, M., & Zhu, Y. (2022). EmoCaps: Emotion capsule based model for conversational emotion recognition. *arXiv preprint:arXiv:2203.13504*.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, *2*(2010), 627–666.

Lodge, M., & Taber, C. S. (2013). *The rationalizing voter*. Cambridge University Press.

Lu, Z., Cao, L., Zhang, Y., Chiu, C.-C., & Fan, J. (2019). Speech sentiment analysis via pre-trained features from end-to-end ASR models. *arXiv preprint:1911.09762*.

Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning — a systematic review. *Intelligent Systems with Applications*, *20*, 200266.

Marcus, G. E. (2003). Emotions in politics. *Annual Review of Political Science*, *3*(1), 221–2250.

Marien, S., & Hooghe, M. (2011). Does political trust matter? an empirical investigation into the relation between political trust and support for law compliance. *Eur. J. Polit. Res.*, *50*(2), 267–291.

McAllister, D. J. (1995). Affect- and Cognition-Based trust as foundations for interpersonal cooperation in organizations. *AMJ*, *38*(1), 24–59.

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *J. Pragmat.*, *59*, 210–220.

Midden, C. J. H., & Huijts, N. M. A. (2009). The role of trust in the affective evaluation of novel risks: The case of CO2 storage. *Risk Anal.*, *29*(5), 743–751.

Mislin, A., Williams, L. V., & Shaughnessy, B. A. (2015). Motivating trust: Can mood and incentives increase interpersonal trust? *Journal of Behavioral and Experimental Economics*, *58*, 11–19.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Comput. Intell.*, *29*(3), 436–465.

Mossholder, K. W., Settoon, R. P., Harris, S. G., & Armenakis, A. A. (1995). Measuring emotion in open-ended survey responses: An application of textual data analysis. *J. Manage.*, *21*(2), 335–355.

Myers, C. D., & Tingley, D. (2017). The influence of emotion on trust. *Polit. Anal.*, *24*(4), 492–500.

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Soc Netw Anal Min*, *11*(1), 81.

Neuert, C., Meitinger, K., & Behr, D. (2021). Open-ended versus closed probes: Assessing different formats of web probing. *Sociol. Methods Res.*, *52*(4), 1981–2015.

Newman, N., & Fletcher, R. (2017). Bias, bullshit and lies: Audience perspectives on low trust in the media. *Available at SSRN 3173579*.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint:1103.2903*.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint:2104.03502*.

Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alemany, L. A., & Martínez, M. V. (2023). Pysentimiento: A python toolkit for opinion mining and social NLP tasks.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via Large-Scale weak supervision. *arXiv preprint:2212.04356*.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., … Bengio, Y. (2021). SpeechBrain: A General-Purpose speech toolkit.

Rittmann, O. (2023). Legislators' emotional engagement with women's issues: Gendered patterns of vocal pitch in the german bundestag. *Br. J. Polit. Sci.*, 1–9.

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., & Wingate, D. (2023). Towards coding social science datasets with language models. *arXiv preprint: 2306.02177*.

Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from twitter text. *J. Comput. Sci.*, *36*, 101003.

Schoon, I., & Cheng, H. (2011). Determinants of political trust: A lifetime learning model. *Dev. Psychol.*, *47*(3), 619–631.

Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. *Am. Sociol. Rev.*, *31*(2), 218–222.

Singh, Y. B., & Goel, S. (2022). A systematic literature review of speech emotion recognition approaches. *Neurocomputing*, *492*, 245–263.

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *J. Pers. Soc. Psychol.*, *48*(4), 813–838.

Sturgis, P., & Smith, P. (2010). Assessing the validity of generalized trust questions: What kind of trust are we measuring? *Int J Public Opin Res*, *22*(1), 74–92.

Theiss-Morse, E., & Barton, D.-G. (2017, January). Emotion, cognition, and political trust. In *Handbook on political trust* (pp. 160–175). Edward Elgar Publishing.

Uslaner, E. M. (2002). *The moral foundations of trust*. Cambridge University Press.

Vu, L., Phan, R. C.-W., Han, L. W., & Phung, D. (2022). Improved speech emotion recognition based on music-related audio features. *2022 30th European Signal Processing Conference (EUSIPCO)*, 120–124.

Wang, Y., Boumadane, A., & Heba, A. (2021). A fine-tuned wav2vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint:2111.02735*.

Willis, G. B. (2004, September). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE Publications.

Winsvold, M., Haugsgjerd, A., Saglie, J., & Segaard, S. B. (2023). What makes people trust or distrust politicians? insights from open-ended survey questions. *West Eur. Polit.*, 1–25.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. M. (2019). HuggingFace's transformers: State-of-the-art natural language processing.

Yoon, S., Byun, S., & Jung, K. (2018). Multimodal speech emotion recognition using audio and text. *arXiv preprint:1810.04635*.

A.1 Sentiment Classification: Evaluation of automated approaches, and alternative Dictionary Approaches

Evaluation of BERT and GPT

In this subsection of the Appendix, we assess the performance of the two automated sentiment classification approaches introduced in the main paper: BERT and GPT. Table A4.1 presents the confusion matrix for GPT and BERT sentiment classification. Notably, when these classifiers diverge, their discrepancies are limited to the neutral category, except for a single instance where BERT identifies positive sentiment and GPT identifies negative sentiment. Both classifiers exhibit an overall agreement of 79%.

	negative	neutral	positive
negative	255	36	0
neutral	46	104	12
positive	1	9	28

Table A4.1: Confusion Matrix BERT and GPT classification.

To systematically evaluate the performance of the two sentiment classifications presented in our paper, we manually annotated a randomly drawn subset of the data (n=197). This dataset was generated by an independent human coder, given no specific instructions beyond categorizing survey responses into positive, negative, or neutral.

By using the human-annotated labels as a reference, we calculated the accuracy (number of correct predictions / total number of predictions) for the two automated approaches, GPT and BERT. Our BERT classification result in an agreement of 78%. In contrast, the GPT classification exhibited slightly better higher accuracy with an agreement of 81% in relation to our manually annotated dataset.

Table A4.2 illustrates the distributions over categories for the three classifiers. Notably, instances of disagreement between the human and the automated approach predominantly involve negative versus neutral assignments.

Category	BERT	GPT	manual
negative	117	112	99
neutral	66	69	87
positive	14	16	11

Table A4.2: Sentiment by BERT, GPT and manual.

A.2 Alternative Dictionary Approaches

Sentiment Classification can be achieved through a variety of modeling architectures. In the main paper we pursued appraoches with Deep Learning Architectures (BERT and GPT), however in this appendix we wanted to provide results from a more simple approach, namely dictionary approaches to this task. Historically, the analysis of sentiment in survey responses was achieved with theory-based dictionaries of affectively scored words (Mossholder et al., 1995) and still nowadays there are various, popular dictionaries available. For this appendix, we utilize two of the most popular ones: AFINN (Nielsen, 2011) and VADER (Hutto & Gilbert, 2014). Both AFINN and VADER are open-source lexicons, however VADER includes more capabilities as it is more comprehensive, context-aware, and granular than AFINN. For example, it is able to correctly handle negations as it incorporates knowledge about grammatical rules. Additionally, VADER is better suited for analyzing short texts (e.g., survey answers), since it is specifically attuned to sentiment expressed in social media.

Since both dictionaries operate on different outcome scales (i.e., -5,5 for AFINN and -1,1 for VADER) we create a categorical outcome variable utilizing rule-based categorization: Assigning '0' as neutral, categorizing the lower 50% of negative scores as highly negative and the upper 50% as slightly negative. Conversely, the upper 50% of positive scores were designated as highly positive, while the lower 50% were marked as slightly positive.

Figure A4.1 depicts the shares of sentiment classifications detected by modality (AFINN vs. VADER). The two classifiers show slight variances, but both convey the impression that the data consists of mainly positive sentiment (70% for AFINN, and 60% for VADER).

Additionally, the correlation between the sentiment categories and political trust scores, as indicated by the red line in the plot, demonstrates a slight increase in political trust as response sentiment is more positive.

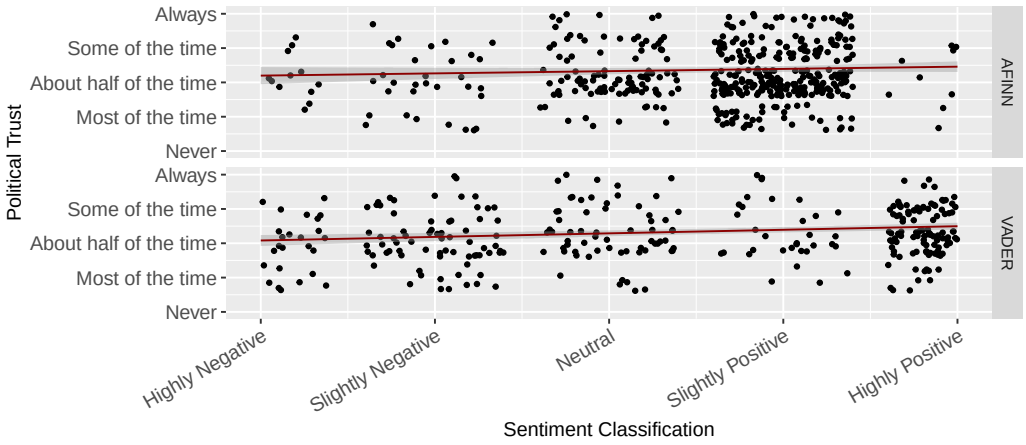

Figure A4.1: Sentiment Classification with five categories by classifier (AFINN vs. VADER). *Note*: Analysis of n=496 open-ended answers, "afinn" achieved using the valence-based AFINN dictionary, "Vader" achieved via Vader compound score, 5-category breaks were achieved with '0' as neutral, categorizing the lower 50% of negative scores as highly negative and the upper 50% as slightly negative, the upper 50% of positive scores as highly positive, while the lower 50% as slightly positive.

In summary, these findings deviate from the results outlined in the main paper, primarily attributed to the dictionary approaches displaying a higher prevalence of neutral and positive sentiment classifications. This discrepancy may stem from various factors. Notably, dictionary approaches struggle to identify sentiment when the response contains terms absent from the dictionary. For example, in survey responses like "Joe Biden is a pedophile. Hillary Clinton is a satanic", both dictionary methods failed to detect any sentiment, leading to their classification as neutral in our coding procedure. In contrast, BERT and GPT successfully identified the unequivocally negative sentiment in such responses.
Dictionary-based classifiers "work in cases where clearly defined sets of words indicate the presence of particular content but struggle with nuance and generalization" (Rytting et al., 2023, p.2). One set of solutions to these challenges stem from the field of machine learning and in particular supervised machine learning

such as naive bayes, random forests, and SVMs have been shown to work well for sentiment classification. The downside however is that they require large datasets of human-annotated training data. Most recent developments and the emergence of large language models that are based on deep learning circumvent even this necessity and the use of fine-tuned deep learning models (fine-tuning only requires a small set of fine-tinung data) or even zero-shot prompting "requires no training data and minimal programming to implement" (Burnham, 2023, p.2), however, "[u]nlike sentiment dictionaries, it produces results comparable to, and sometimes better than, supervised classification" (Burnham, 2023, p.2). Learning-based approaches (Bello et al., 2023 and Chiorrini et al., 2021 use BERT models for sentiment classification of Twitter Data; Sailunaz and Alhajj, 2019 use Naive Bayes), such as deep learning models have shown better performance in terms of sentiment classification compared to lexicon-based approaches (Bharti et al., 2022; Chiorrini et al., 2021). Their largest contribution might be that they move away from a simple bag-of-words approach but instead the context of the respective words. Eventually, similarly to the preceding subsection, we computed the degree of agreement between these dictionary-generated classifications and the manually annotated ones. This analysis serves as a demonstration that our more complex, deep-learning-based classifiers are significantly more effective in capturing nuanced sentiment.
Results from using AFINN compared to the reference of or manually annotated data, achieves an agreement of 21%. For VADER, the same comparison results in an agreement of 26%.

A.3 Alternative Regression Models

Ordinal Logistic Regression: Political Trust on Sentiment (GPT)

Figure 4.3 in the main paper showed findings from regression analysis of sentiment on trust scores. This appendix provides robustness measures for an alternative model specification, namely findings for an ordinal (or ordered) logistic regression. Ordinal logistic regression is a suitable model when outcome variables are ordinal in nature, meaning they possess a meaningful and ordered categorical structure, but the intervals between categories are not assumed to be equal – which

might be the case for our 5-category political trust item.
Figure A4.2 shows predicted probabilities for each of the outcome categories (e.g., trust-never (1), trust-some of the time (2), etc) by predicting different intercepts and slopes for each category.

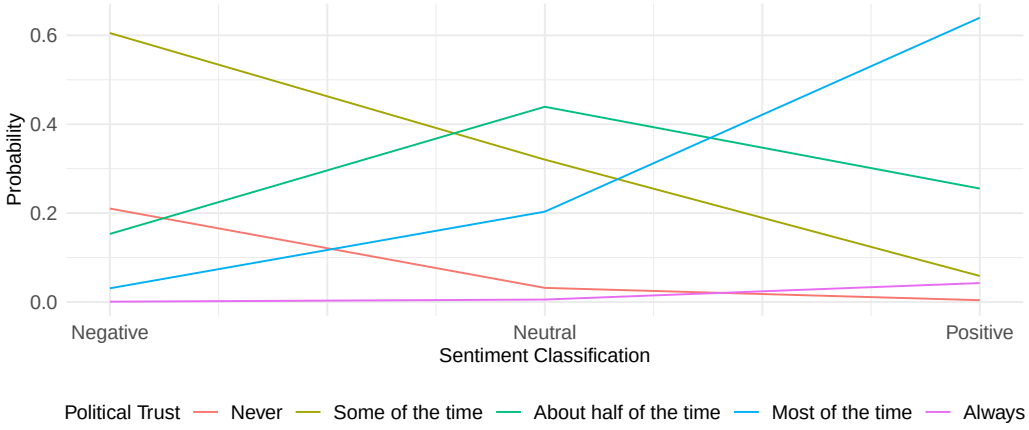

Figure A4.2: Ordered logit model of sentiment and a five-category trust score (bivariate).

In the main paper we intentionally chose to treat our outcome variable as continuous due to various reasons. Firstly, it enhances the overall readability and accessibility of our research findings as we choose the simplest model that is appropriate. Secondly, it. yields similar results as an ordered logistic model and thirdly, our decision aligns with previous research that describes the understanding that it "[...] has become common practice to assume that Likert-type categories constitute interval-level rather than ordinal-level measurement" (Blaikie, 2003, p.231). Monte Carlo Simulations have also shown that parametric tests, such as a F-Test in a linear regression, was strongly robust to the interval data assumption (as well as moderate skewing) when data was collected using a 5 to 7 point Likert response format with no resulting bias (Carifio & Perla, 2007; Glass et al., 1972).

Linear Regression: Political Trust on Sentiment (BERT)

Figure A4.3 shows findings for a linear regression of political trust (5 categories) on our sentiment variable achieved with BERT. Figure A4.3 is the counterpart to

Figure 4.2 in the main paper.

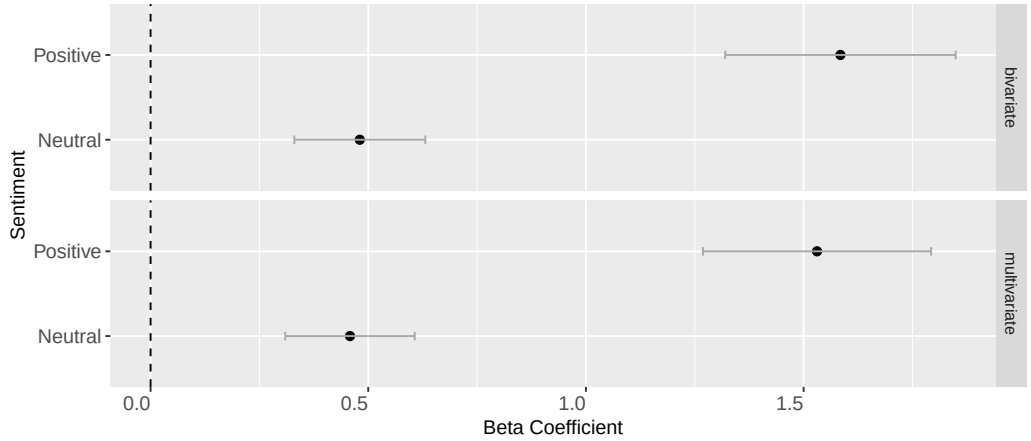

Figure A4.3: Linear model of sentiment (BERT) and a five-category trust score (bi- and multivariate). *Note*: The data for the sentiment classification displayed in this figure stems from the BERT-based classification. Respective findings for the GPT classification can be found in the main paper.

Alternative Emotion Classification: NRC word-emotion lexicon

For the task of emotion classification, in the main paper, we chose to analyze the original, "raw" audio files, as they might contain paralinguistic features, such as intonation, pitch, volume, pauses, but also laughter or breathing noises. However, one can also research text (i.e., our transcripts from the audio files) for emotions. Despite the lack of paralinguistic characteristics in text, they can still convey important information since the emotional essence of an utterance can also be significantly conveyed through the choice of words. For example, words such as "lovely" and "awesome" carry strong emotions compared to more generic, non-emotional words, such as "person" and "day" (Yoon et al., 2018, p.1) and a sentence like "This phone is a piece of junk" has a stronger valence than "I think this phone is fine" (Liu, 2010, p.632). Hence, we additionally classified emotions using a text-based emotion recognition classifier, namely the NRC word-emotion lexicon (Mohammad & Turney, 2013). While there are multiple off-the-shelf methods available to achieve this task, for example pysentimiento's emotion classifier (Pérez et al., 2023), the NRC lexicon can be considered to be one of the

most popular resources for the analysis of emotions in texts (Nandwani & Verma, 2021). It contains associations of words with eight basic emotions: joy, surprise, sadness, anger, disgust, fear, contempt, and anticipation. By analyzing a given text, words are assigned to the corresponding word-emotion pairs in the dictionary. Since each word is assigned to a specific emotion (in some cases more than one), we decided to use a multi-class classification. In this multi-class classification each sentence can be assigned multiple emotions at the same time (only the neutral category is of course exclusive). Figure A4.4 displays the findings from our emotion classification results obtained through applying the NRC lexicon. For comparison, Figure 4.4 displays the audio-based analysis using SpeechBrain from the main paper.

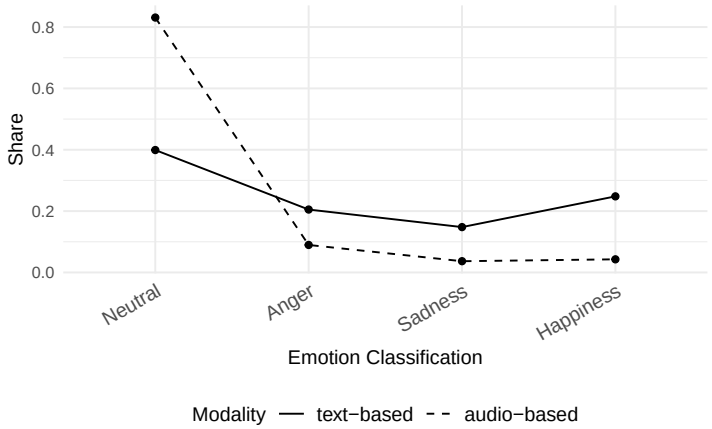

Figure A4.4: Emotion Classification obtained from EmoLex. *Note*: Text-based analysis achieved with the NRC Emotion Lexicon, Audio-based analysis achieved with SpeechBrain. Classification of n=404 open-ended answers, note that text-based NRC classifier allows for multi-class assignment.

Figure A4.4 demonstrates that irrespective of modeling approach the majority of open-ended responses do not contain explicit emotional language and are characterized by a neutral emotional tone. However and notably, the share of neutral statements is significantly higher in the audio-based condition compared to the text-based one (85% vs. 39%). In other words, the text-based analysis identifies a greater share of emotions within the survey answers.

A.4 Prompting Strategy

This appendix aims at describing some of the details we considered when drafting the prompt for our zero-shot classification of sentiment using GPT-3.5-turbo. As was described in the main paper, our prompt aimed at provide a very clear and straightforward instruction which returns a single token (i.e, positive, negative, neutral) to detect the nature of associations in its most standard form: positive, negative or neutral. Our prompt read as follows: | "Classify the sentiment of the following open-ended survey answer into neutral, negative or positive. Text: i Sentiment:", where i is the survey response.

Since sentiment is a comparatively easy task, we refrained from providing specific examples, which allowed us to use zero-shot (compared to few-shot) prompting. Prompt engineering plays a crucial role in influencing the final results, as slight modifications in the prompt, such as formatting changes, can alter the probability distribution over tokens and, consequently, impact the outcome. Nevertheless, in a series of experiments, Rytting et al. (2023) demonstrated that, on the whole, prompt engineering has limited influence on GPT-3's performance when coding social science datasets. Their findings emphasize the importance of selecting unique first tokens for each category, avoiding overly extreme descriptors like "very positive" and "very negative" in favor of more neutral terms like "positive" and "negative." Moreover, it is recommended to employ substantive alternatives in the prompt, replacing binary choices like "yes" and "no" with more meaningful options such as "positive" and "negative." Interestingly, the format of the prompt itself does not appear to significantly impact results. This includes practices like enclosing categories in quotes or using various delimiters such as slashes or pipes to separate the prompt task and input. Such findings underscore the relative stability of the information retrieval process in the face of diverse prompt formats.

5 Conclusion and Discussion

> *"[...] the current way of using and analyzing open-ended questions is not satisfactory; at least, their potential value has not been fully explored."* (Zhang et al., 2022, p.2)

The social sciences have a long history of incorporating open-ended survey questions into their research methodologies. Thus, it does not come as a surprise that various methodological research was conducted to investigate the analysis of this special type of textual data. Despite their inconsistent use in the past, open-ended questions are currently witnessing increased popularity, partly due to the increasing number of methods and models available to analyze natural language. In the past, such analyses were frequently conducted manually, using methods like content analysis. However, automated text analysis is now becoming more common. Methods for such automated workflows are still under development and thus this endeavor remains an ever-evolving one since it is accompanied by the frequent and regular development of new technologies.

This thesis aimed at contributing to these developments by introducing the various methods available and by demonstrating their application in three empirical studies. Chapter 1.1 introduced the survey question type of open-ended questions alongside a depiction of their characteristics, benefits and fields of application. Chapter 1.2 provided an overview of methods available for the analysis of data from open-ended questions. It distinguished between three types of workflows – manual, semi-automated, and fully automated – and described how the social sciences have developed an increasing number of computational methods that stem from NLP subfields such as text mining. This overarching theoretical background was followed by three empirical studies, each of which a) collected open-ended survey answers and b) automatically analyzed them with regards to different outcomes. Study 1 was interested in the content of associations respondents have with standard trust items in surveys and fine-tuned two transformer-based BERT models to classify such associations. Study 2 aimed at classifying the information content in open-ended responses (according to their response mode) and for this applied various different computational methods for textual data (e.g., topic models). Study 3 was concerned with emotions in open-ended answers and analyzed

data from open-ended voice answers to detect whether there are emotional associations using language models (i.e., BERT, GPT).

The empirical findings in Chapters 2, 3, and 4, together with the theoretical foundation from Chapter 1, allow a few conclusions to be drawn. First, this thesis illustrates how state-of-the-art models, in particular large language models (LLMs) can achieve relatively high accuracy scores. For example, in Study 1, fine-tuned BERT models outperformed a respective random first classifier trained with the same data, reaching accuracy rates of 87% and 95% compared to 83% and 92%, respectively.

Second, such accuracies can be achieved with a relatively small number of human-labeled examples in the fine-tuning dataset (e.g., only 13-20% in Study 1), resulting in significant gains in efficiency compared to fully automated approaches.

Third, for researchers that want to refrain from crafting individual fine-tuning / training data (as presumed in points 1 and 2), drawing on "off-the-shelf", thus readily available, fine-tuned models, might be an option, as exemplified in Study 3 (i.e., pysentimiento). These models can represent useful and efficient workflows, but at the same time might be accompanied by the cost of reduced accuracies (e.g., 78% in Study 3) as well as a trade-off in control and explainability due to the absence of independently crafted and tailored fine-tuning data.

Fourth, this thesis introduced zero-shot prompting as a successful strategy for relatively straightforward and common tasks (e.g., sentiment classification), requiring only minimal human input (only a prompt with no labeled examples). Study 3 demonstrated a high accuracy of 81% for this approach.

In sum, these findings and their interpretations suggest that advanced methods, such as language models exhibit remarkable performance, which might be no surprise when we consider their substantial computational capabilities. However, this advantage is counterbalanced by an accuracy-explainability tradeoff which was faced in all of the three studies in this thesis. This tradeoff for example means that employing deep-learning methods may enhance accuracy but simultaneously diminish the ability to explain the underlying processes, thus compromising the transparency of the research. For example, in Study 2 the usage of unsupervised clustering in the form of topic models, represents a powerful method for clustering, however comes with small transparency as to how these clusters are created

in the first place.

Another limitation of this thesis is that despite this inclusion of unsupervised learning, it did not assess fully automated, unsupervised approaches to the same extent as semi-automated approaches. One rationale for the focus on semi-automated approaches is that methods that allow for classification based on independently predefined categories (i.e., fully manual and semi-automated) are very popular approaches in the social sciences, as most research includes theories, hypotheses and operationalizations developed beforehand.

Ultimately, decisive factors for the final decision on whether to pursue a fully manual, semi-automated or fully automated approach can include the difficulty of the given task, the size of the available dataset, the structure of the open-ended text answers (e.g., length, amount of context), as well as the available resources. For example, in the latter case, working with large language models may require high computation power such as GPU, while simpler approaches can be executed locally.

Future research in the field of classification of open-ended survey responses is likely to experience many developments and changes in the near future and the interplay of manual and automated classifications is only one possible debate. Above all, it is important to further develop new methods and evaluate existing ones. Mosca et al., in the year 2022, painted a rather negative picture in stating that "[c]urrent analysis practices employ shallow machine learning methods or rely on (biased) human judgment" (Mosca et al., 2022, p.49). Even since 2022, a large number of new methods have emerged. Large language models such as BERT models have been integrated into the task of analyzing text data from surveys (Grootendorst, 2022; Gweon & Schonlau, 2023; Mosca et al., 2022; Schonlau et al., 2023), and zero- and few-shot prompting methods currently being explored in the social sciences (Gilardi et al., 2023; Latif et al., 2023; Rytting et al., 2023; Zhu et al., 2023) represent only the very latest methodologies in this regard, and are the culmination of many years of methodological development. Young fields like Speech Emotion Recognition bring new possibilities and tools to the social sciences. The evaluation of these comparatively new methods might be of high priority in the next few years.

Furthermore, the analysis of open-ended answers requires the initial collection of

such data, an aspect subject to ongoing debate according to Popping (2015). For example, survey participation via smartphones has increased over the last years (Gummer et al., 2023), but comes with some challenges for open-ended questions due to the small typing screen (Beuthner et al., 2022). The amount of research devoted on the correct design and usage of open-ended questions in different modes and devices (Denscombe, 2008; Keusch, 2014; Kunz et al., 2021; Peytchev & Hill, 2010; Schmidt et al., 2020; Smyth et al., 2009; Zuell et al., 2015), will be a crucial determinant in the future prevalence of open-ended questions. This could, for example, include research on the correct use of incentives in the context of smartphone-based studies (Wenz & Keusch, 2023). Generally, recent advances in web survey methodology will influence the future landscape of open-ended questions in surveys.

References

Beuthner, C., Silber, H., & Stark, T. H. (2022). Effects of smartphone use and recall aids on network name generator questions. *Soc. Networks*, *69*, 45–54.

Denscombe, M. (2008). The length of responses to Open-Ended questions: A comparison of online and paper questionnaires in terms of a mode effect. *Soc. Sci. Comput. Rev.*, *26*(3), 359–368.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms Crowd-Workers for Text-Annotation tasks. *arXiv preprint:2303.15056*.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint:2203.05794*.

Gummer, T., Höhne, J. K., Rettig, T., Roßmann, J., & Kummerow, M. (2023). Is there a growing use of mobile devices in web surveys? evidence from 128 web surveys in germany. *Qual. Quant.*, *57*(6), 5333–5353.

Gweon, H., & Schonlau, M. (2023). Automated classification for open-ended questions with bert. *Journal of Survey Statistics and Methodology*.

Keusch, F. (2014). The influence of answer box format on response behavior on List-Style Open-Ended questions. *J Surv Stat Methodol*, *2*(3), 305–322.

Kunz, T., Quoß, F., & Gummer, T. (2021). Using placeholder text in narrative Open-Ended questions in web surveys. *J Surv Stat Methodol*, *9*(5), 992–1012.

Latif, S., Usama, M., Malik, M. I., & Schuller, B. W. (2023). Can large language models aid in annotating speech emotional data? uncovering new frontiers. *arXiv preprint:2307.06090*.

Mosca, E., Harmann, K., Eder, T., & Groh, G. (2022). Explaining neural NLP models for the joint analysis of Open-and-Closed-Ended survey answers. In A. Verma, Y. Pruksachatkun, K.-W. Chang, A. Galstyan, J. Dhamala, & Y. T. Cao (Eds.), *Proceedings of the 2nd workshop on trustworthy natural language processing (TrustNLP 2022)* (pp. 49–63). Association for Computational Linguistics.

Peytchev, A., & Hill, C. A. (2010). Experiments in mobile web survey design. *Soc. Sci. Comput. Rev.*, *28*(3), 319–335.

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., & Wingate, D. (2023). Towards coding social science datasets with language models. *arXiv preprint: 2306.02177*.

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.

Schonlau, M., Weiß, J., & Marquardt, J. (2023). Multi-label classification of open-ended questions with BERT. *arXiv preprint: 2304.02945*.

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 Opin. Q.*, *73*(2), 325–337.

Wenz, A., & Keusch, F. (2023). Increasing the acceptance of Smartphone-Based data collection. *Public Opin. Q.*, *87*(2), 357–388.

Zhang, R., Gong, J., Ma, S., Firdaus, A., & Xu, J. (2022). Automatic coding mechanisms for Open-Ended questions in journalism surveys: An application guide. *Digital Journalism*, 1–22.

Zhu, Y., Zhang, P., Haq, E.-U., Hui, P., & Tyson, G. (2023). Can ChatGPT reproduce Human-Generated labels? a study of social computing tasks.

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. *Soc. Sci. Comput. Rev.*, *33*(1), 115–122.