Original Article

Sociological Methods & Research
1–31

DOI: 10.1177/00491241241234871
journals.sagepub.com/home/smr


How Valid Are Trust Survey Measures? New Insights From Open-Ended Probing Data and Supervised Machine Learning

Camille Landesvatter[1] **and Paul C. Bauer**[2,3]

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

[1] Mannheim Centre for European Social Research, University of Mannheim, Mannheim, Baden-Württemberg, Germany
[2] Department of Politics, University of Freiburg, Freiburg, Germany
[3] Department of Statistics, Ludwig Maximilian University of Munich, Munich, Germany

Corresponding Authors:
Camille Landesvatter, Mannheim Centre for European Social Research, University of Mannheim, A5, 6, Mannheim, Baden-Wuerttemberg 68159, Germany.
Email: camille.landesvatter@uni-mannheim.de
Paul C. Bauer, Department of Politics, University of Freiburg, Freiburg, Baden-Wurttemberg, Germany; Department of Statistics, Ludwig Maximilian University of Munich, Munich, Germany.
Email: mail@paulcbauer.de

Data Availability Statement included at the end of the article

situation-specific measures should have the desirable property of further reducing variation in associations, that is, 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, bidirectional encoder representations from transformers

Introduction

Generalized social trust is one of the fundamental concepts in contemporary social theory (Coleman 1994; Herreros 2004; Putnam, Leonardi, and Nanetti 1994; Schilke, Reimann, and Cook 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 and Vickstrom 2011; Van Deth 2003). Consequently, empirical research investigating the causes and consequences of trust has multiplied (Buskens and Weesie 2000; Cook and Cooper 2003; Dinesen 2012; Dinesen, Sonne Nørgaard, and Klemmensen 2013; Dinesen and Sønderskov 2015; 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 and Freitag 2018; Delhey and Newton 2005; Delhey, Newton, and Welzel 2011; Ermisch et al. 2009; Nannestad 2008; Robbins 2019; Sturgis and Smith 2010; Torpe and 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 United Kingdom (Sturgis, Brunton-Smith, and Jackson 2019; Sturgis and Smith 2010). 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 and 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 2019, 2021), 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 and Freitag 2018; Sturgis and 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 and Kunz 2022; Neuert, Meitinger, and Behr 2021) and automated text analysis (e.g., Schonlau and Couper 2016). The data we labeled and the resulting supervised classifiers we built are suitable for future applications.

Theory, Hypotheses, and Previous Research

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 and Smith 2010; Uslaner 2002:52). Stolle (2015) for example emphasizes the need to distinguish the scope of generalized trust from trust toward people one personally knows (Stolle 2015:398). 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 and Yamagishi 1994:146), or as trust in the "average person [one] meets" (Coleman 1994:104). 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 and Traunmüller 2009:784), including family members, friends, neighbors and co-workers (Freitag and Traunmüller 2009:784) (i.e., known others), whereas generalized trust encompasses "[…] those beyond immediate familiarity, including strangers" (Freitag and Traunmüller 2009:784) (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 et al. 2009; Ermisch and Gambetta 2010; Fehr et al. 2002; Naef and 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 and Freitag 2018) regarding various aspects, such as scale length or balance (Lundmark, Gilljam, and

Dahlberg 2016), and the frames of reference employed by respondents when answering it (Delhey, Newton, and Welzel 2014; Nannestad 2008; Sturgis and 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 1 depicts our main argument regarding these associations.

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 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 and 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 Figure 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 six higher-order topics they found respondents to associate with the term "most people." The two largest categories they found by

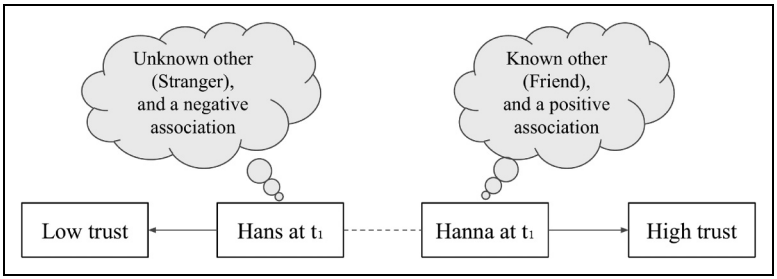

Figure 1. Variation in associations and trust measurement values.

manually classifying responses to their probing question were "known others" (42 percent) and "unknown others" (22 percent).[1] 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 and Freitag 2018:9). Lastly, Uslaner (2002:72-4), as part of the 2000 ANES Pilot Survey, investigated the most people question via think-aloud techniques and showed that 58 percent of the respondents referred to a "general worldview" while 23 percent 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 percent of respondents referred to family members, 11 percent to co-workers, and 12 percent 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 2019, 2021). Our particular interest for each of these measures lies in the proportion of 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 → people you meet for the first time → 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 (H_1). Evidence for H_1 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 (H_2) 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.

Negative Associations

While trust research regularly discusses the impact of experiences on trust (Brehm and Rahn 1997; Cao, Galinsky, and Maddux 2014; Dinesen 2010; Freitag and Traunmüller 2009; Glanville, Andersson, and Paxton 2013; Glanville and Paxton 2007; Uslaner 2002), studies about trust measurement have neglected this dimension. On average, trust in known others is higher (Bauer and Freitag 2018; Sturgis and Smith 2010; Vollan 2011)—as is also evidenced by measures that directly gauge trust in family members, neighbors, etc. (Freitag and 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 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, that is, 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).[2]

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 to teach children to refuse offers from strangers (Moran et al. 1997:11). Postulating H_1, 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 → people you meet for the first time → a total stranger you meet for the first time) increases the proportion of respondents who have negative associations (p_n) (H_3). Again, evidence for H_3 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 (H_4) 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.

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 × [expected behavior] enters explicitly" (Nannestad 2008:415). 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, Hardin, and Levi 2005; Schilke, Reimann, and Cook 2021). Ermisch et al. (2009) criticized 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" (Ermisch et al. 2009:750). 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 et al. 2009:751; Ermisch and Gambetta 2010:4).

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 situation. Importantly, the decision to trust in situation A may not carry over to situation B (Ermisch and Gambetta 2010:4) 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 vaguer 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 (2019, 2021). These novel measures are based on the stranger question (M3) because they specify the trustee to be a stranger (cf. M3) (see Buskens and Weesie 2000; Yamagishi and Yamagishi 1994; Yuki etal. 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 1). Analogous to H_1, 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 (H_5). 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.

Data, Experimental Design, and Methods

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 and 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.[3] 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, i.e., overall response rate of 95 percent) 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 Online 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/hour on average while the mean duration was 6.8 minutes.

Experimental Design and Measures

Our questionnaire design is depicted in Table 1. Respondents provided their data via an online self-administered survey (created using formR, cf. Arslan, Walther, and Tata 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 Blocks #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 Blocks #1 and #2 in a randomized

Table 1. Experimental Design.

Structure of the survey (from left to right)			
	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 questions #1 and #4	Additional questions
Information and 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)

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' 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.[4] The specific phrasing as well as summary statistics of these questions can be found in Online Appendices A.1 and 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.[5]

Block #2: Situative trust measures and probing questions. Block #2 included four situative measures that represent the Imaginary Stranger Trust (IST) scale developed by Robbins (2019, 2021, 2022). 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,[6] (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 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.

Methods

Table 2 illustrates the structure of our data. Due to the intra-person design, there are multiple (i.e., seven) measures of trust (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`).[7] Online Appendix A.3 provides a detailed description of the open-ended text answers. Table 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.

Both classifications (i.e., known–unknown and sentiment) were achieved using automated text analysis, which in survey data research has become a

Table 2. Illustration of Exemplary Data.

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)
…	…	…	…	…	…

Note: The table displays different exemplary respondents. In the actual dataset each respondent/ID (cf. column 1) appears seven times, because each respondent received all seven trust items (for five of these questions the respondents received a respective probing question).

popular alternative to manual coding (Esuli and Sebastiani 2010; Giorgetti and Sebastiani 2003; Gweon and 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 this number to $n = 1{,}500$.[8] Both samples were a random selection of text answers from the generalized trust measures (see Online Appendix A.5.2 for further details). Based on previous implementations in the literature, we argue that these sample sizes are sufficiently large.[9]

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 falls into category 0).[10] 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 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). Online 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 bidirectional encoder representations from transformers (BERT) models (BERT base model uncased version), using the manually coded data ($n = 1{,}000/1{,}500$) as training data. BERT (Devlin et al. 2019) is an empirically powerful machine learning technique that can be used for various natural language processing tasks (Devlin et al. 2019:1). 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," for example, 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 percent) have negative associations. BERT achieves higher class-wise accuracy in the presence of class imbalance than other ngram-based machine learning techniques (Gweon and Schonlau 2023), and is further demonstrated to remove the need to use data augmentation techniques to mitigate problems of imbalanced data (Madabushi, Kochkina, and Castelle 2020).[11] Importantly, the imbalanced data structure and its consequences does not call into question the effects we found but may have resulted in their slight underestimation. Online Appendix A.5.2 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 is shown in Table 3.

Alternative approaches with which we classified our data (i.e., regular expressions and random forest) can be found in Online Appendix A.6.

Results

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). Regardless of the subsample, there is a gradual decline in trust from Measure 1 (most people

Table 3. Accuracy, Precision, Recall, and F1-Score.

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

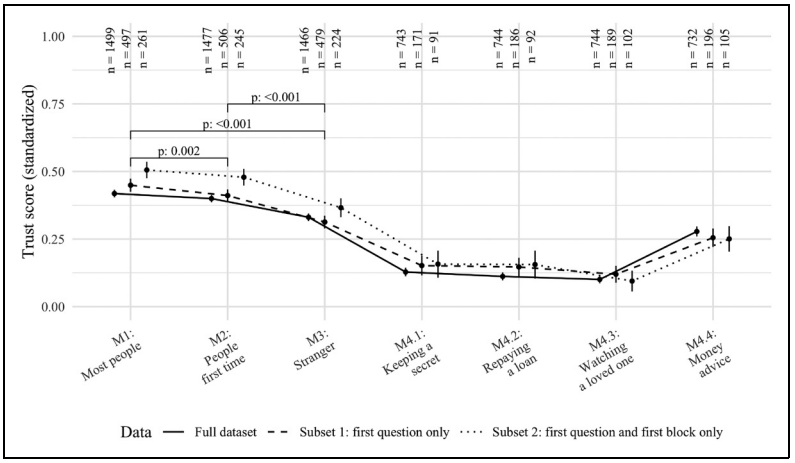

Figure 2. Standardized trust scores across different trust measures and respondent subsets. *Note*: The figure shows point estimates for average trust scores and 95 percent confidence intervals. Details on the respondent subsets are provided in the "Methods" section. *P*-values are derived from *t*-tests for the full dataset, for details see footnote 12. Data for M4.1–4.4 include the "stranger" wording only (see footnote 6).

question) to Measure 2 (people first time question), and finally, to Measure 3 (stranger question).

Within-subjects ANOVA reveals that the generalized trust scores differed statistically significantly for the same individual for the three question wordings ($F(1.7, 2{,}505) = 129$, $p < 0.001$).[12]

Additionally, situative trust measures M4.1–4.4 consistently exhibit lower trust levels likely owing to their emphasis on trust decisions where the truster has a lot to lose.[13] It is crucial to note that Figure 2 provides a descriptive overview 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.

Associations Across Standard and Situative Measures

We start by examining the known–unknown dimension. Figure 3 displays the share of respondents who described associations of either known or unknown

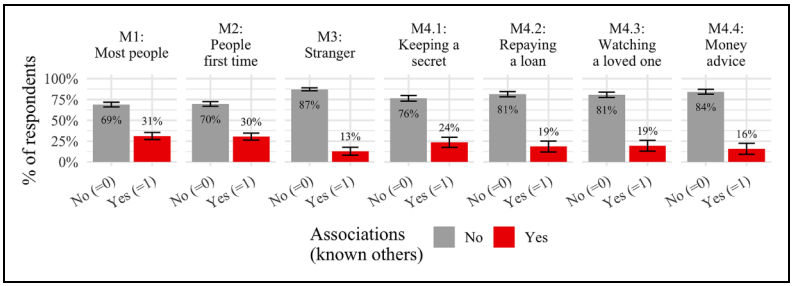

Figure 3. Distribution of associations with known people across trust measures. *Note*: Error bars represent 95 percent confidence intervals (lower cutoff at 0). Data is the full dataset irrespective of the question or block randomization (details are provided in the "Methods" section). Results for different subsets of the data can be found in Online Appendices A.5.2 and A.5.3.

others across our seven measures.[14] In line with our expectation (H_1), the share of respondents referring to a known other statistically significantly decreases for M3 (i.e., 13 percent) while shares for M1 and M2 are similar (31 percent and 30 percent, 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.

With regards to the sentiment dimension, we expected to find different shares of negative sentiment for each question wording (see Figure 4). In line with our expectations (H_3), the share of negative associations is higher for M3 (i.e., 8.7 percent) compared to M2 (7 percent). Not in line with our hypothesis, the share for M1 is higher (10 percent). 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."

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 ($r(7, 490) = -0.08$, $p = <0.001$).

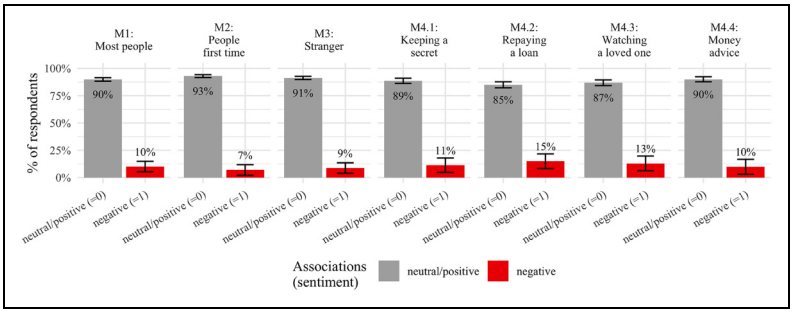

Figure 4. Distribution of associations and their sentiment across trust measures. *Note:* Error bars represent 95 percent confidence intervals (lower cutoff at 0). Data is the full dataset irrespective of the question or block randomization (details are provided in the "Methods" section).

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 5 visualizes the coefficients for a series of regression models (see Online 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., Models #1 and #2 in Figure 5). We subsequently add covariates to these bivariate regressions (e.g., Models #3 and #4 in Figure 5).[15] Finally, the fifth model includes both dummies in one model and adds covariates.

In accordance with our expectations (H_2), we observe that associations with known others have a positive effect on trust for all of our three 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 ($p < .001$), 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 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 5) does not mitigate the effect of the known–unknown dummy on trust.

In line with our expectation (H_4), we find that negative associations have a negative effect on trust for all of our three generalized trust

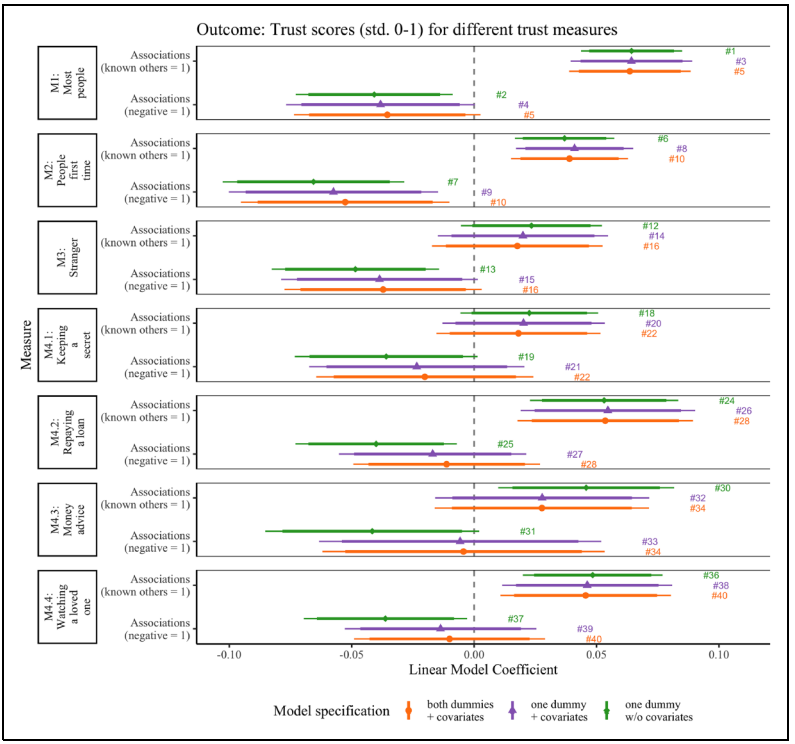

Figure 5. Associations and trust scores across different measures. *Note:* The figure shows point estimates for coefficients of our dummy variables of interest namely having associations with known others or negative associations. Bars represent 90 percent (thicker) and 95 percent (thinner) confidence intervals. Data is the full dataset irrespective of the question or block randomization (details are provided in the "Methods" section).

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 H_2. 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, that is, trusting someone to repay a loan ($\beta_{\#24} = 0.053$,

$p < 0.001$). For the situative measures, however, while consistent with H_4, 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 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.

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, Newton, and Welzel 2011; Ermisch et al. 2009; Nannestad 2008; Robbins 2019; Sturgis and 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 and Kunz 2022). Open-ended probing (Neuert, Meitinger, and Behr 2021) is still a novelty in trust research, and similar data has so far only been collected in interviewer-administered settings (Sturgis and 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). 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 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 percent to 31 percent, cf. Figure 3) incorporated thoughts of known individuals in their responses while answering classic trust questions, which is in line with previous research (e.g., Sturgis and 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 and 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 2019, 2022). Interestingly, compared to M3, the situative measures (M4.1–4.4) showed an increase in respondents thinking about known individuals (but still considerably smaller than in M1 and M2) (cf. Figure 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 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.[16] 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 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 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 (M3 and M4.1–4.4) 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.[17] 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 3). However, for those interested in interpersonal comparability, situative measures like the IST scale 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.[18] Generally, future studies could make use of additional, situative measures by 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 and Freitag 2018; Robbins 2022).[19] 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 the United Kingdom (Sturgis and 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 lengths (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 (7×2) 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 2019, 2021)), and considering other emerging measures, such as the Risk Aversion question in the GSOEP and the UK Household Longitudinal Study,[20] 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. 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 Online 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.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Paul C. Bauer and Camille Landesvatter (Project No. 449946260) gratefully acknowledge support by the German Research Foundation (DFG).

ORCID iDs

Camille Landesvatter https://orcid.org/0000-0003-0156-5364
Paul C. Bauer https://orcid.org/0000-0002-8382-9724

Data Availability Statement

Data and code required to reproduce the findings presented in this study are available in a public repository on Harvard Dataverse (doi:10.7910/DVN/FJXH5G). To access the data and code, please visit the following link: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FJXH5G. For any inquiries or assistance related to accessing the materials, readers are encouraged to contact the corresponding author listed in this manuscript.

Supplemental Material

The supplemental material for this article is available online.

Notes

1. Smaller categories they found refer to "local community" (e.g., people in their town) (3 percent), "job/profession" (e.g., politicians and salesmen) (4 percent), "other" (e.g., "trusting is naive") (5 percent), and "don't know/no answer" (6 percent).
2. Where the latter comprises both neutral and positive associations.
3. Gender: two groups, namely males and females; Age: five groups in year-year brackets; Ethnicity: five groups, namely White, Mixed, Asian, Black, and Other.
4. 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, 1972) described 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.
5. 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.
6. 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 Online Appendix A.2).
7. Each respondent was probed for each generalized trust measure (M1–M3), resulting in 3×1, 500 entries, as well as for two out of four situative trust measures

(M4.1–4.4), resulting in additional 2×1, 500 entries. Out of 10,500 answers to trust questions, 3,000 responses were not probed.

8. Detecting sentiment proves more complex than spotting mentions of known and unknown others due to several factors, such as ambiguous word meanings.
9. Schonlau and Couper (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 Couper (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 and Couper 2016).
10. Coding of the n=1,000 training data observations shows that circa 9 percent of the answers include mentions of "groups of people," these instances were all coded as "unknown others."
11. Still, we attempted oversampling (see e.g., Gosain and 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.
12. 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), 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, $t(1{,}455) = 15.15$, $p < 0.001$.
13. 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. This approach could help mitigate the impact of strong outliers from specific situations.
14. Online Appendix A.5.2 shows these results using data from the manually coded share of data only ($n = 1{,}000/1{,}500$). Online Appendix A.5.3 shows these results using data for Subset 2 only ($n = 1{,}500$).
15. Age (catgeorical), sex, ethnicity, socioeconomic status, income, and education.
16. Analogous to Sturgis and Smith (2010), we randomized respondents to trust measures in Blocks #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.

17. 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 (cf. Hardin 2002:9).
18. For more detailed considerations between shorter and longer versions of IST, we refer readers to Robbins (2022).
19. 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.
20. "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

ANES. 2000. "ANES: Codebook Variable Documentation (Version 04), 2000 Pilot Study (2000.PN)."

Arslan, Ruben C., Matthias P. Walther, and Cyril S. Tata 2020. "Formr: A Study Framework Allowing for Automated Feedback Generation and Complex Longitudinal Experience-Sampling Studies Using R." *Behavior Research Methods* 52(1): 376-87.

Barr, Abigail. 2003. "Trust and Expected Trustworthiness: Experimental Evidence From Zimbabwean Villages." *The Economic Journal* 113(489): 614-30.

Bauer, Paul C. and Markus Freitag. 2018. "Measuring Trust." Pp. 1-27 in *The oxford handbook of social and political trust*, edited by E. M. Uslaner. Oxford University Press.

Behr, Dorothée, L. Kaczmirek, W. Bandilla, and Michael Braun. 2012. "Asking Probing Questions in Web Surveys: Which Factors Have an Impact on the Quality of Responses?" *Social Science Computer Review* 30(4): 487-98.

Behr, Dorothée, Katharina Meitinger, Michael Braun, and Lars Kaczmirek. 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, Norman. 2003. *Analyzing Quantitative Data*. London: SAGE Publications Ltd.

Brehm, John and Wendy Rahn. 1997. "Individual-Level Evidence for the Causes and Consequences of Social Capital." *American Journal of Political Science* 41(3): 999.

Buskens, Vincent and Jeroen Weesie. 2000. "An Experiment on the Effects of Embeddedness in Trust Situations: Buying a Used Car." *Rationality and Society* 12(2): 227-53.

Cao, Jiyin, Adam D. Galinsky, and William W. Maddux. 2014. "Does Travel Broaden the Mind? Breadth of Foreign Experiences Increases Generalized Trust." *Social Psychological and Personality Science* 5(5): 517-25.

Carifio, James and J. Rocco Perla. 2007. "Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends About Likert Scales and Likert Response Formats and Their Antidotes." *Journal of Social Sciences* 3(3): 106-16.

Coleman, James S. 1994. *Foundations of Social Theory*. Boston, MA: Harvard University Press.

Cook, Karen S. and Robin M. Cooper (2003) "Experimental Studies of Cooperation, Trust, and Social Exchange." Pp. 209-44 in *Trust and reciprocity: Interdisciplinary lessons for experimental research*, edited by E. Ostrom and J. Walker. New York: Russell Sage Foundation.

Cook, Karen S., Russell Hardin, and Margaret Levi. 2005. *Cooperation Without Trust?* New York: Russell Sage Foundation.

Delhey, Jan and Kenneth Newton. 2005. "Predicting Cross-National Levels of Social Trust: Global Pattern or Nordic Exceptionalism?" *European Sociological Review* 21(4): 311-27.

Delhey, Jan, Kenneth Newton, and Christian Welzel. 2011. "How General Is Trust in 'Most People'? Solving the Radius of Trust Problem." *American Sociological Review* 76(5): 786-807.

Delhey, Jan, Kenneth Newton, and Christian Welzel. 2014. "The Radius of Trust Problem Remains Resolved." *American Sociological Review* 79(6): 1260-65.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019) "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." Pp. 4171–86 in *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, edited by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics.

Dinesen, Peter Thisted. 2010. "Upbringing, Early Experiences of Discrimination and Social Identity: Explaining Generalised Trust Among Immigrants in Denmark." *Scandinavian Political Studies* 33(1): 93-111.

Dinesen, Peter Thisted. 2012. "Does Generalized (Dis)Trust Travel? Examining the Impact of Cultural Heritage and Destination-Country Environment on Trust of Immigrants." *Political Psychology* 33(4): 495-511.

Dinesen, Peter Thisted, Asbjørn Sonne Nørgaard, and Robert Klemmensen. 2013. "The Civic Personality: Personality and Democratic Citizenship." *Politische Studien* 62(1_suppl): 134-52.

Dinesen, Peter Thisted and Kim Mannemar Sønderskov. 2015. "Ethnic Diversity and Social Trust: Evidence From the Micro-Context." *American Sociological Review* 80(3): 550-73.

Ermisch, John and Diego Gambetta. 2010. "Do Strong Family Ties Inhibit Trust?" *Journal of Economic Behavior & Organization* 75(3): 365-76.

Ermisch, John, Diego Gambetta, Heather Laurie, Thomas Siedler, and S.C. Noah Uhrig. 2009. "Measuring People's Trust." *Journal of the Royal Statistical Society* 172(4): 749-69.

Esuli, Andrea and Fabrizio Sebastiani. 2010. "Machines That Learn How to Code Open-Ended Survey Data." *International Journal of Market Research* 52(6): 775-800.

Fehr, Ernst, Urs Fischbacher, Bernhard von Rosenbladt, Jürgen Schupp, and Gert G. Wagner. 2002. "A Nation-Wide Laboratory. Examining Trust and Trustworthiness by Integrating Behavioral Experiments Into Representative Surveys." *Journal of Contextual Economics-Schmollers Jahrbuch* 122(4): 519-42.

Freitag, Markus and Richard Traunmüller. 2009. "Spheres of Trust: An Empirical Analysis of the Foundations of Particularised and Generalised Trust." *European Journal of Political Research* 48(6): 782-803.

Giorgetti, Daniela and Fabrizio Sebastiani. 2003. "Automating Survey Coding by Multiclass Text Categorization Techniques." *Journal of the American Society for Information Science and Technology* 54(14): 1269-77.

Glanville, Jennifer L., Matthew A. Andersson, and Pamela Paxton. 2013. "Do Social Connections Create Trust? An Examination Using New Longitudinal Data." *Social Forces; a Scientific Medium of Social Study and Interpretation* 92(2): 545-62.

Glanville, Jennifer L. and Pamela Paxton. 2007. "How Do We Learn to Trust? A Confirmatory Tetrad Analysis of the Sources of Generalized Trust." *Social Psychology Quarterly* 70(3): 230-42.

Glass, GeneV., Percy D. Peckham, and James R. Sanders. 1972. "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance." *Review of Educational Research* 42(3): 237-88.

Gosain, Anjana and Saanchi Sardana (2017) "Handling Class Imbalance Problem Using Oversampling Techniques: A Review." Pp. 79-85 in *2017 international conference on advances in computing, communications and informatics (ICACCI)*.

Gweon, Hyukjun and Matthias Schonlau. 2023. "Automated Classification for Open-Ended Questions With BERT." *Journal of Survey Statistics and Methodology* (smad015): 1-12.

Hardin, Russell. 2002. *Trust and Trustworthiness*. New York: Russell Sage Foundation.

Herreros, Francisco. 2004. *The Problem of Forming Social Capital. Why Trust*. New York: Springer.

Larsen, Christian Albrekt. 2013. *The Rise and Fall of Social Cohesion: The Construction and De-Construction of Social Trust in the US, the UK, Sweden and Denmark*. Oxford: OUP Oxford.

Lundmark, Sebastian, Mikael Gilljam, and Stefan Dahlberg. 2016. "Measuring Generalized Trust: An Examination of Question Wording and the Number of Scale Points." *Public Opinion Quarterly* 80(1): 26-43.

Madabushi, Harish Tayyar, Elena Kochkina, and Michael Castelle. 2020. "Cost-Sensitive BERT for Generalisable Sentence Classification with Imbalanced Data." *arXiv* 2003.11563.

Meitinger, Katharina and Tanja Kunz. 2022. "Visual Design and Cognition in List-Style Open-Ended Questions in Web Probing." *Sociological Methods & Research* 0(0): 1-28.

Moran, Ellen, David Warden, Lindsey Macleod, Gillian Mayes, and John Gillies. 1997. "Stranger-Danger: What Do Children Know?" *Child Abuse Review: Journal of the British Association for the Study and Prevention of Child Abuse and Neglect* 6(1): 11-23.

Naef, Michael and Jürgen Schupp. 2009. "Measuring Trust: Experiments and Surveys in Contrast and Combination." *SOEPpaper No. 167*.

Nannestad, Peter. 2008. "What Have We Learned About Generalized Trust, If Anything?" *Annual Review of Political Science* 11(1): 413-36.

Neuert, Cornelia E., Katharina Meitinger, and Dorothée Behr. 2021. "Open-Ended Versus Closed Probes: Assessing Different Formats of Web Probing." *Sociological Methods & Research* 52(4): 1981-2015.

Palan, Stefan and Christian Schitter. 2018. "Prolific.ac—A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* 17: 22-27.

Portes, Alejandro and Erik Vickstrom. 2011. "Diversity, Social Capital, and Cohesion." *Annual Review of Sociology* 37(1): 461-79.

Putnam, Robert D., Robert Leonardi, and Raffaella Y. Nanetti. 1994. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Robbins, Blaine G. 2019. "Measuring Generalized Trust: Two New Approaches." *Sociological Methods & Research* 51(1): 305-56.

Robbins, Blaine G. 2021. "An Empirical Comparison of Four Generalized Trust Scales: Test–Retest Reliability, Measurement Invariance, Predictive Validity, and Replicability." *Sociological Methods & Research* 0(0): 1-44.

Robbins, Blaine G. 2022. "Valid and Reliable Measures of Generalized Trust: Evidence From a Nationally Representative Survey and Behavioral Experiment." *Socius* 9:1-26.

Rosenberg, Morris. 1956. "Misanthropy and Political Ideology." *American Sociological Review* 21(6): 690-95.

Schilke, Oliver, Martin Reimann, and Karen S. Cook. 2021. "Trust in Social Relations." *Annual Review of Sociology* 47(1): 239-59.

Schonlau, Matthias and Mick P. Couper. 2016. "Semi-Automated Categorization of Open-Ended Questions." *Survey Research Methods* 10(2): 143-52.

Smith and Sandra Susan. 2010. "Race and Trust." *Annual Review of Sociology* 36(1): 453-75.

Sønderskov and Kim Mannemar. 2011. "Does Generalized Social Trust Lead to Associational Membership? Unravelling a Bowl of Well-Tossed Spaghetti." *European Sociological Review* 27(4): 419-34.

Stolle, Dietlind. 2015. "Trusting Strangers—The Concept of Generalized Trust in Perspective." *Österreichische Zeitschrift für Politikwissenschaft* 31(4): 397-412.

Sturgis, Patrick, Ian Brunton-Smith, and Jonathan Jackson. 2019. "Regression-Based Response Probing for Assessing the Validity of Survey Questions." Pp. 571-91 in *Advances in questionnaire design, development, evaluation and testing*. John Wiley & Sons, Inc.

Sturgis, Patrick and Patten Smith. 2010. "Assessing the Validity of Generalized Trust Questions: What Kind of Trust Are We Measuring?." *International Journal of Public Opinion* 22(1): 74-92.

Sztompka, Piotr. 1999. *Trust: A Sociological Theory*. Cambridge: Cambridge University Press.

Torpe, Lars and Henrik Lolle. 2011. "Identifying Social Trust in Cross-Country Analysis: Do We Really Measure the Same?" *Social Indicators Research* 103(3): 481-500.

Uslaner, Eric M. 2002. *The Moral Foundations of Trust*. Cambridge: Cambridge University Press.

Van Deth, Jan W. 2003. "Measuring Social Capital: Orthodoxies and Continuing Controversies." *International Journal of Social Research Methodology* 6(1): 79-92.

Vollan, Björn. 2011. "The Difference Between Kinship and Friendship: (Field-) Experimental Evidence on Trust and Punishment." *The Journal of Socio-Economics* 40(1): 14-25.

Yamagishi, Toshio and Midori Yamagishi. 1994. "Trust and Commitment in the United States and Japan." *Motivation and Emotion* 18(2): 129-66.

Yuki, Masaki, William W. Maddux, Marilynn B. Brewer, and Kosuke Takemura. 2005. "Cross-Cultural Differences in Relationship- and Group-Based Trust." *Personality & Social Psychology Bulletin* 31(1): 48-62.

Author Biographies

Camille Landesvatter is a PhD candidate and Research Associate at the Mannheim Centre for European Social Research, University of Mannheim, Germany. Her research focuses on computational social science, where she is currently working in the project "TrustME: Measurement and Explanation of Trust". Within this project,

she explores generalized social trust, specifically investigating the analysis of open-ended survey responses using automated approaches, such as machine learning techniques.

Paul C. Bauer works at the University of Freiburg and the Ludwig Maximilian University of Munich. His research interests include political sociology, methods and computational social science. His most recent work has been published in the Journal of Big Data, Political Communication, Political Behavior and Socius.