June 2026

# The Impact Of Voice Interfaces On Strategic Misreporting During Self-Disclosure

Marc Christopher Grau
*University of St. Gallen*, marcchristopher.grau@unisg.ch

Naim Zierau
*University of St. Gallen*, naim.zierau@gmail.com

Ivo Blohm
*University of St. Gallen*, ivo.blohm@unisg.ch

# THE IMPACT OF VOICE INTERFACES ON STRATEGIC MISREPORTING DURING SELF-DISCLOSURE

*Completed Research Paper*

**Marc Grau, University of St. Gallen, Switzerland, marcchristopher.grau@unisg.ch**

**Naim Zierau, University of St. Gallen, Switzerland, naim.zierau@unisg.ch**

**Ivo Blohm, University of St. Gallen, Switzerland, ivo.blohm@unisg.ch**

## Abstract

*Voice interfaces are increasingly used for large-scale data collection, such as virtual health checkers, where users disclose personal information. Yet, they may shape self-reports differently than text. Heightened social presence and immediacy of the interaction can foster impression management while reducing time for deliberation. We test whether speaking versus typing changes strategic misreporting in a between-subjects experiment (N = 155) where participants disclosed health behaviors via voice or text interfaces. Voice reduced self-enhancing reports on positively valenced items, suggesting potential strategic misreporting in text conditions. For negatively valenced items, differences were not significant, suggesting more careful deliberation in reporting. However, mediation analyses suggest that lower perceived control in voice can indirectly increase reporting of undesirable behaviors, with signs of gender heterogeneity. Guilt as a potential factor did not show signs of mediation. Based on our results, we discuss design implications and theoretical contributions on modality, control, and strategic misreporting.*

*Keywords: Strategic Misreporting, Voice Interface, Reporting, Modality, Self-Disclosure*

## 1 Introduction

Voice-enabled technologies have become one of the most consequential shifts in human–computer interaction in recent times (Cohen & Oviatt, 1995; Melumad, 2023; Seaborn et al., 2021). From reporting symptoms to a virtual health checker (Sezgin et al., 2020) to filing an insurance claim through a bot (Zierau et al., 2023), speech is increasingly replacing the keyboard as an input modality. In fact, WhatsApp users now send on the order of 7 billion voice messages each day (Singh, 2022) with large user groups using voice as their main interface (Bellman, 2017). As voice interfaces proliferate, an ever-growing share of personal and often sensitive data information is disclosed by speaking rather than typing (Melzner et al., 2023). This shift matters because many downstream decisions rest on the accuracy of self-reported data. Virtual triage assistants, for example, can only perform as well as the symptoms they hear, and an insurance bot's risk assessment is only as good as the truthfulness of the claims it receives. As voice becomes the conduit for these disclosures, the integrity of spoken self-reports becomes mission-critical. Unfortunately, self-reported data are famously fallible, subject to social desirability and pronounced in sensitive situations (Rudnik et al., 2024). Almost 12 percent of Americans misstate their private insurance status (Nguyen et al., 2023), and 20 percent are misclassified into the wrong BMI category when height and weight are self-reported (Preston et al., 2015). Such errors, whether accidental or strategic, occur through medical diagnoses, underwriting, and personalized services, sometimes with grave consequences. When incentives favor distortion, such as understating risky behaviors or inflating damages, the stakes rise further. This naturally leads to the following question: If human self-reports are already prone to error, could the choice of input modality, voice versus text, further shape the likelihood of misreporting?

Although prior work offers important insights into how voice differs from text, it provides only fragments when it comes to understanding the truthfulness of spoken self-reports. On the one hand, voice has been shown to amplify social presence, which can increase self-awareness and the desire to manage impressions (Berger et al., 2022; Melumad, 2023). On the other hand, voice can foster a sense of immediacy that encourages spontaneous, even vulnerable, disclosures (Des Jarlais et al., 1999; Moon, 2000). These factors could increase internal pressure to present oneself favorably or heighten the guilt of telling a lie out loud. In other words, it is theoretically ambiguous whether speaking to a computer makes users more honest or more strategic in their self-reports than when typing. This duality raises a critical question: Does speaking make users more honest or more strategic?

Even though prior studies have compared written and spoken input in self-reports (e.g., Moon, 1998; Wambsganss et al., 2022; Wei et al., 2022) they rarely focus on truthfulness nor examine situations where users stand to gain from deception or rely on benign, artificial tasks. Moreover, much research builds on legacy systems such as Interactive Voice Response (IVR), which lack the conversational quality of today's voice assistants (Wei et al., 2022). Finally, much of this work treats voice as a monolith, overlooking how psychological mechanisms, interaction content, and user characteristics shape the outcomes of human-computer conversations. Hence, as voice increasingly becomes the medium through which sensitive data is disclosed, we urgently need to understand whether it helps or hinders the truthfulness of what people say. We address these questions through an online experiment that compares spoken and typed self-reports across scenarios where participants can benefit from misreporting. Specifically, we conducted a between-subject experiment with 155 participants who reported health-related information to receive an insurance premium, which was paid out in the form of a monetary bonus. Moreover, we examined user-reported perceptions to identify key design aspects of voice-enabled systems that could influence the truthfulness and experience of self-reporting. We find that users report strategically depending on the interface modality and the valence of the question. Reporting to positively valenced questions, where higher values would yield a better premium outcome, users talking to a voicebot report significantly lower health status than users typing the answers. For negative answers, the identified effect diminishes. However, in negatively valenced questions, we identify perceived control as a driving mechanism leading to more honest reporting behavior, which is moderated by gender. To summarize, our contributions are threefold. First, we collect empirical evidence that disentangles modality-specific effects on strategic misreporting across differently valenced questions in an incentive-compatible task. Second, we provide theoretical insight into how social and cognitive properties of voice interact to influence self-reporting behavior. Third, we suggest actionable design guidance for developers of high-stakes voice systems, aiming to safeguard data integrity and user trust.

# 2 Conceptual and Theoretical Foundations

## 2.1 Strategic Misreporting in Self-Disclosure

Many organizations rely on self-reported information to guide consequential decisions in domains such as healthcare and insurance (Thomaz et al., 2020). However, self-reports are traditionally vulnerable to strategic misreporting, defined as deliberate distortion of personal information to gain advantage (Cohn et al., 2022; Weinmann et al., 2022). Common examples include under-reporting smoking habits on health forms or inflating damages in insurance claims (Nguyen et al., 2023; Preston et al., 2015). This vulnerability is particularly acute in computerized interactions, where self-reports are increasingly collected through web forms or AI agents rather than human interlocutors (Kim et al., 2023; LaMothe & Bobek, 2020; Petisca et al., 2020). Compared to face-to-face communication, these digital exchanges often lack social and moral cues, such as empathy, eye contact, or the possibility of judgment, that ordinarily discourage dishonesty (LaMothe & Bobek, 2020). Consequently, users are more inclined to mislead machines than humans, particularly when interactions feel impersonal, anonymous, or morally neutral (Lei et al., 2025).

To counter this tendency, research has begun to explore how the design of particular AI agents can encourage more honest and socially appropriate behavior (Lei et al., 2025). Prior work shows that anthropomorphic cues, conversational style, and social roles can shape users' perceptions of AI agents and, in turn, their ethical choices (Kim et al., 2023; Lei et al., 2025). However, much of this research has emphasized how AI agents appear and behave, rather than the underlying structure of interaction. One such foundational but often overlooked element is the communication modality, that is, the channel through which users disclose information, such as voice or text. Modality is not a neutral technical choice; rather, it fundamentally shapes how users engage cognitively and socially with another actor (Oba & Berger, 2024). In this sense, modality constitutes a core property of an IS artifact (Orlikowski & Iacono, 2001) that system designers actively select and configure. Unlike anthropomorphic cues or conversational styles, which have received considerable attention as design variables (Kim et al., 2023; Lei et al., 2025), the choice between voice and text is typically treated as a technical implementation detail rather than as a design decision with behavioral consequences. Modality influences the ease of revision, the level of deliberation, and the associations users draw about audience and context (Cohen & Oviatt, 1995; Oba & Berger, 2024). These dynamics suggest that modality may play a critical role in determining whether and how individuals strategically misreport.

A related but conceptually distinct construct is Socially Desirable Responding (SDR). SDR refers to a general tendency to answer in ways that conform to perceived social norms and avoid social sanction, even in the absence of direct incentives (Paulhus, 1991). In contrast, we focus on strategic misreporting as deliberate, incentive-sensitive distortion of self-reports to obtain a better outcome without the risk of getting caught (Cohn et al., 2022). Whereas SDR is primarily driven by social-evaluative concerns and impression management, strategic misreporting is instrumental and goal-directed. It requires individuals to evaluate the payoff structure, plan how to adjust their answers, and implement those adjustments.

## 2.2 A Dual-Pathway Model: Social vs. Cognitive Mechanisms

Why might spoken self-reports differ from written ones? Discourse research points to several key distinctions between the two modalities (Chafe & Tannen, 1987; Redeker, 1984). First, speech is generally more spontaneous and intuitive than writing. Voice interactions unfold rapidly through turn-taking (Levinson, 2016), leaving little room for planning or revision compared to written text (Rubin et al., 2000). Second, speech is also inherently more social (Akinnaso, 1982): Paralinguistic cues such as intonation, pauses, or loudness convey affective information (Hildebrand et al., 2020; Melzner et al., 2023; Rogers et al., 2016; Sutton et al., 2019), while turn-taking fosters a sense of social presence, even in the absence of a real listener (Clark-Gordon et al., 2019). To that end, speakers are more likely to adapt their phrasing to their perceived audience, adjusting for familiarity or formality (Bell, 1984). Taken together, these characteristics suggest that spoken reports are often more immediate, less filtered, and more audience-sensitive than written ones. Although prior work clearly documents differences between spoken and written communication, it remains less clear whether these distinctions extend to disclosing information truthfully to machines. This is a critical gap, as users often adapt their speech to accommodate limitations in voice recognition (Cohn et al., 2021; Luger & Sellen, 2016). At the same time, they appear to transfer social norms into these interactions. Moreover, IVR surveys have elicited more extreme and less socially desirable responses compared to traditional formats (Dillman & Christian, 2005; Roster et al., 2004). However, these legacy systems differ markedly from today's voice interfaces, which interact with customers in an increasingly human-like manner (Wang et al., 2023) raising new cognitive and social considerations. We therefore investigate two distinct pathways through which interface modality may influence strategic misreporting: A social-affective mechanism and a cognitive-process mechanism. For each mechanism we build upon existing theory to propose two distinct pathways through which interface modality may influence strategic misreporting.

The social-affective pathway emphasizes that voice, as a rich and inherently social modality, amplifies social presence and audience awareness (Akinnaso, 1982; Melzner et al., 2023) and can be considered a social actor (Nass et al., 1994). Prior IS work shows that more humanlike and socially present agents can increase impression management and shape ethical behavior, in part by making moral norms and anticipated guilt more salient (Kim et al., 2023; Lei et al., 2025; Schuetzler et al., 2018). In this view,

speaking "out loud" to an apparently social partner should heighten the anticipated guilt of misreporting, particularly for socially valued behaviors, and thereby curb self-enhancing distortion.

The cognitive-process pathway, in contrast, emphasizes interactional constraints and cognitive resources. Interpersonal Deception Theory (IDT) conceptualizes deception as a strategic, effortful process that requires planning, monitoring, and adaptation to the interaction environment (Buller & Burgoon, 1996). Text-based interfaces typically afford high deliberation and revisability: users can pause, scan all questions, and edit or restructure their answers (Moffett et al., 2021; Zierau et al., 2023), which facilitates deliberate misreporting. Voice interfaces are synchronous, immediate, and offer low revisability; studies of speech interfaces document that users often feel constrained in pacing and repair, reporting reduced perceived control over the interaction (Luger & Sellen, 2016; Maharjan et al., 2021; Porcheron et al., 2018). From an IDT perspective, these process constraints limit the capacity to plan and execute deceptive responses, especially when concealing negative behaviors requires careful deliberation. We therefore conceptualize perceived control as the subjective manifestation of these modality-specific affordances: lower perceived control in voice should make strategic misreporting harder to implement, even when incentives and SDR motives are present.

## 2.3 Hypotheses

Based on our integrated framework, we evaluate two primary pathways through which modality may affect strategic misreporting: a social-affective pathway (via guilt) and a cognitive-process pathway (via perceived control). IDT posits that deception is a strategic, deliberate process (Buller & Burgoon, 1996). Text-based interfaces afford high deliberation and revisability, which supports the planning required for strategic misreporting (Bereby-Meyer & Shalvi, 2015). Conversely, voice interfaces are synchronous, immediate, and have low revisability (Moffett et al., 2021), constraining this planning. At the same time, the Computers Are Social Actors (CASA) effect, amplified by voice, creates a sense of social presence that text lacks (Melumad & Meyer, 2020; Nass et al., 1994). This social presence can increase self-awareness and a "self-audience" effect (Reinfeldt et al., 2010), raising the anticipated guilt of lying out loud. We thus conceptualize strategic misreporting as shaped by (a) social-affective pressures that make deception feel morally costly, and (b) cognitive process constraints that make deception harder to execute. However, strategic misreporting is not monolithic. Effects are dependent on the direction, it includes self-enhancement (inflating positive traits) and self-protection (concealing negative behaviors) (Tourangeau & Yan, 2007). We argue these are driven by different paths. Self-enhancement is primarily an act of impression management. We expect this to be most sensitive to the social-affective pressure of voice. In contrast, concealing negative behaviors is an act of deliberation, which we expect to be most sensitive to the cognitive process constraints (IDT) of voice. A key consideration is that item valence is not inherent to a behavior but depends on the reporting context and incentive structure surrounding disclosure (Tourangeau et al., 2000; Tourangeau & Yan, 2007). How sensitive a given item feels to respondents, and thus how strongly it invites strategic distortion, is shaped by the design of the reporting task rather than by the behavior alone. Because this sensitivity varies across item classes even within the same interaction, we distinguish two forms of strategic distortion that map onto differently valenced items.

*H1a: For positively valenced items, participants in the voice condition will report lower (i.e., less self-enhancing) values than participants in the text condition.*

*H1b: For negatively valenced items, participants in the voice condition will report higher (i.e., more self-disclosing) values than participants in the text condition.*

We propose two distinct but complementary mechanisms to explain valence-dependent effects, and we explicitly allow that both can, in principle, operate for positively and negatively valenced behaviors. The social-affective path focuses on guilt: moral emotions such as guilt link internalized moral standards to (un)ethical behavior and help regulate norm-violating choices (Tangney et al., 2007). Anticipated guilt has been shown to deter dishonest or self-serving behavior in high-stakes interactions, including those involving AI agents (Kim et al., 2023). At the same time, research on social desirability and self-presentation distinguishes between different components of biased responding, such as egoistic versus

moralistic biases and self-deceptive enhancement versus impression management (Paulhus, 1991; Paulhus & John, 1998). While this work typically treats self-enhancement as part of socially desirable responding, our theorizing does not frame self-enhancement primarily as generic SDR. Instead, we conceptualize self-enhancing distortion as a moralized act of "claiming unearned virtues," in which misrepresentation threatens one's moral self-image and is regulated by anticipated guilt rather than by a broad desire to appear socially acceptable. On this view, guilt can in principle constrain both exaggeration of positive, socially valued behaviors and denial of negative, norm-violating behaviors, but it should be especially salient when endorsing strongly positive self-descriptions that are closely tied to moral identity.

The IDT-based cognitive-process path emphasizes deliberate response editing and concealment. Work on sensitive questions in surveys shows that underreporting socially undesirable or stigmatized behaviors often arises at a deliberate response stage, where respondents strategically sanitize or omit negative information, a process that is cognitively effortful and depends on planning and monitoring resources (Gnambs & Kaspar, 2015; Tourangeau & Yan, 2007). Studies of voice and conversational interfaces further show that turn-by-turn, audio-only interaction, lack of visual overview, and low revisability can leave users feeling they have less control over how to formulate, review, and correct their answers than in traditional text-based or web interfaces (Luger & Sellen, 2016; Maharjan et al., 2021, 2022). Reduced perceived control in such settings thus reflects a loss of the very planning and editing capacities that strategic concealment of negative behaviors relies on. Accordingly, although process constraints could also dampen self-enhancing exaggeration, we expect the control pathway to be most diagnostic for self-protective misreporting on negatively valenced behaviors, where deliberate concealment is especially demanding.

*H2a: The effect of voice on positively valenced items (H2a) will be mediated by increased anticipated guilt.*

*H2b: The effect of voice on negatively valenced items (H2b) will be mediated by reduced perceived control.*

Finally, communication norms and deception propensities are not uniform. Research suggests gender differences in both deception for financial gain (Capraro, 2018; Dreber & Johannesson, 2008) and the perception of agents (Clark et al., 2019). We therefore hypothesize a moderating effect:

*H3: The effects of modality on strategic misreporting will be moderated by participant gender.*

# 3 Methodology

## 3.1 Experimental Design and Procedure

The study employed a between-subjects experimental design (n = 155). Participants were randomly assigned to one of two groups: the treatment group, which interacted with a voice bot, and the control group, which answered the same set of questions using a web survey. Both groups were presented with identical questions regarding their personal habits. A list of the questions is provided in Table 1.

We derived the questions similarly to Schuetzler et al. (2018), where they study social presence in the context of sensitive information disclosure. To extend the scope of data collection towards positive and negative features, we added five questions and adjusted them to be on similar time scales. These questions are particularly suited for the context of our paper because we want to study the effects of interacting with a voice bot or the web survey in a natural conversational setting instead of a stylized task such as a coin-flipping experiment (Cohn et al., 2022). People naturally know their actual behavior and reporting inflated or deflated numbers requires a conscious effort for strategic misreporting. Furthermore, people act differently depending on whether they act in a simulated or an authentic setting (Y. Wang & Shah, 2022). Lastly, collecting data in health and healthcare is a widely adopted practice by various businesses and public institutions (Dorsey et al., 2014; Hutchinson & Sutherland, 2019; Maharjan et al., 2021). Therefore, it can be assumed that users are familiar with this kind of survey task.

| Question | Association |
|---|---|
| How many fruits or vegetables do you eat per day? | Positive |
| On how many days do you consume fast food or junk food in a week? | Negative |
| How many glasses of water do you drink per day? | Positive |
| How many glasses of sugary beverages do you drink per day? | Negative |
| On how many days do you consume alcohol in a week? | Negative |
| How many days per week do you exercise? | Positive |
| On average, in minutes, how long is a typical exercise session when you train? | Positive |
| How many hours of sleep do you get on average per night? | Positive |
| On how many days do you feel stressed in a week? | Negative |

*Table 1. Experimental questions with corresponding health association labels*

In our insurance premium setting, behaviors such as exercise frequency or water intake are positively valenced because higher values signal healthier habits and thus a better premium outcome. Conversely, behaviors such as alcohol or fast food consumption are negatively valenced because higher values indicate health risks. The only difference between the groups was the interface used to collect their responses. Participants in the treatment group interacted with the voice bot in a question-answer style conversation, while participants in the control group entered their responses to the questions into text fields provided by the survey. To minimize potential grouping effects, the questions were presented in a randomized order. The overall setup, the questions, and the information provided to participants were consistent across both groups, ensuring that the only variable was the mode of interaction - voice versus text.

Participants conducted the study remotely, having been recruited through the online platform Prolific. To ensure the quality of the data, participants were required to have a working microphone and audio output. The study was framed as an experiment to improve the communication of bonus programs in a health insurance context. Participants were informed that their information would be used to determine a bonus for their health insurance premium. Therefore, participants were incentivised to report better health metrics, as they were informed that the top 10% of respondents with the healthiest reported behaviors would receive an additional bonus payment. The entire experiment took around 10 minutes per participant. After completing the experimental part, all participants were asked to complete a post-experiment survey.

## 3.2 Participants

A total of 172 participants were recruited through Prolific, all English-speaking UK or US residents. The participant pool was restricted to the UK and US to ensure a baseline familiarity with modern conversational AI and to maintain a consistent linguistic and cultural context for the health-related questions. Seven participants failed an attention check and dropped out of the study. Another seven participants did not finish the experiment and, thus, dropped out. Lastly, three participants did not consent to the study and were rejected. This leads to a total of 155 participants finishing the study.

The average age of the participants was 35.2 years (SD = 10.5), with a gender distribution of 55\% male, 43\% female, and 2\% identifying as other genders. Participants were randomly assigned to the voice bot interface or text interface conditions. All participants provided informed consent before starting the experiment and received £1.50 for their participation, with the potential bonus amounting to an additional £1.00. In the voice bot condition, data from 73 participants was successfully recorded, while the text-based survey group yielded 82 complete trials.

To assess the effectiveness of the randomization procedure, we conducted t-tests for continuous demographic variables and chi-square tests for categorical demographic variables between the voice bot and text survey conditions. The t-tests for continuous variables (age and income) suggest no significant difference in age between the voice and text conditions ($p = 0.49$), while the income variable amounts to a *p-value* of *0.07*. The chi-square tests for categorical variables show no significant differences in

gender, ethnicity, education, or parents' education between the groups, with *p-values* of *0.44*, *0.12*, *0.29*, and *0.72*, respectively. Overall, the randomization procedure effectively balanced the demographic variables between both groups. There were no statistically significant differences, suggesting that the randomization was generally successful.

## 3.3 Interaction Interfaces

The voicebot interface used in the study was developed using Next.js. It was designed to simulate a human-like conversation about participants' health habits. The minimalistic graphical user interface ensured participants focused on the spoken dialogue and restricted usage to the voice modality without additional design elements. Participants recorded their answers by pressing and holding a button. When the bot spoke, the record button was replaced with a speaking animation. A transcript or any other visual elements besides the experiment description and the button were not shown to participants. An example of the UI is depicted in Figure 1.

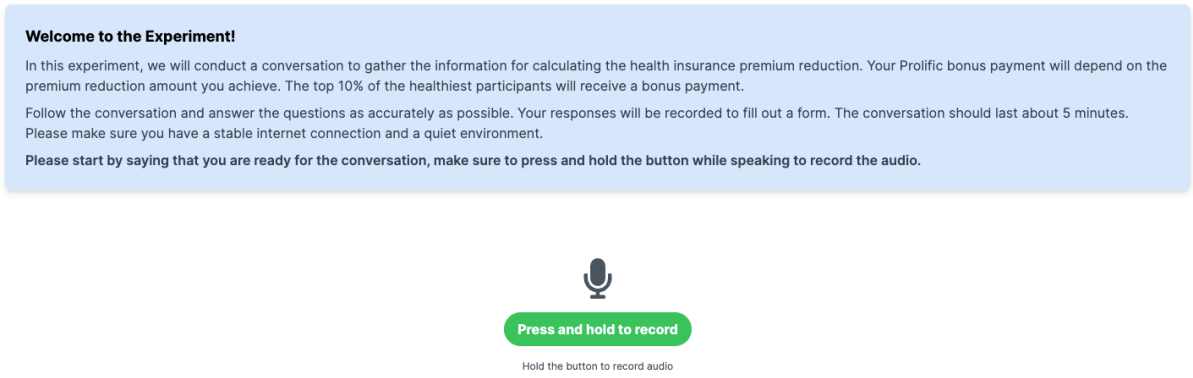


*Figure 1. Voice User Interface with Task Description and Interaction Button*

The voicebot used Whisper large-v3 (Radford et al., 2023) for speech-to-text conversion. Two LLMs were used in parallel for text processing, with one instance of GPT-4o-mini handling the conversation and a second instance of GPT-4o-mini handling information extraction and validation. The second LLM checked the answers and stored them in a database. To handle errors, the conversational LLM re-prompted the user if an answer was unintelligible. Any unanswered questions were stored in a list and fed randomly to the conversational LLM instance for questioning. When a question was answered, it was removed from the list. If the answer was not understood or was unrelated to the context, the question was not removed and asked again. Re-prompting was triggered only by speech recognition failures, not by content evaluation. A clearly understood response was accepted regardless of its content, which differs from the deliberate revisability of text interfaces where users can compare and re-craft answers across items. Once all questions were answered, the conversational LLM thanked the user for participating. The text messages generated by the conversation handling LLM were synthesized into audio responses using TTS-1 from the OpenAI API. Only the audio response was returned to the user.

Participants in the control group answered the questions using a traditional web survey interface provided by Qualtrics, as seen in Figure 2. Each question was presented as a separate item on the survey, with text fields for participants to input their responses. The interface provided no interactive feedback, aiming to simulate a static online survey experience. This setup, while straightforward, lacked the conversational dynamics of the voice bot. The layout of the Qualtrics survey was standardized across all participants to ensure consistency.

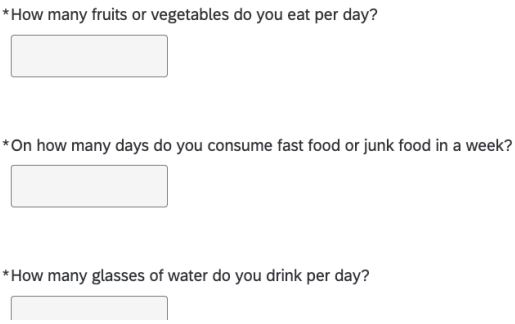


*Figure 2. Text User Interface with Exemplary Questions*

## 3.4 Data Processing

Responses and complete conversation scripts collected from the voice bot interactions were automatically transcribed and stored in a database. Responses to the web survey were collected via the Qualtrics platform. Before analysis, all data underwent preprocessing. While survey answers were all restricted to be numerical, answers to the voice bot could not be restricted to numerical values due to the inherent conversational nature of the process. Thus, we applied different processing steps for vaguely answered questions based on observations across all answers. We calculated the mean of the mentioned numbers for the answers provided as intervals. Answers on the wrong time scale were adjusted to fit the question's time scale. For example, "four times a month" as an answer was corrected to once a week. Comparative descriptions of behavior, such as "less than two," were reduced to the next lower integer following the mathematical description of the statement. Statements like "Never," "None," or "very rarely" were all set to zero to represent the user's intention numerically. These processing steps were designed to conservatively translate ambiguous natural language into quantitative data. Finally, all experiment and questionnaire answers were standardized using z-score normalization for scale-invariant analysis and better comparability.

## 3.5 Constructs

We summarized self-reported health behaviors along three constructs: (i) positively valenced behaviors (POS), (ii) negatively valenced behaviors (NEG), and (iii) an aggregate score that combines all questions into a single score. The POS score contains question items for which higher values indicate healthier behavior; the NEG score contains question items for which higher values indicate unhealthier behavior (see Table 1). To ensure comparability across item scales, all item responses were z-standardized within the analytic sample prior to score construction.

*Latent factor scores.* For each construct, we fit a single-factor confirmatory factor analysis (CFA) using the z-standardized responses to the questions: one CFA for the POS score, one for the NEG score, and one for the aggregate score. For the aggregate constructs, all negatively valenced items were negated (so that higher values consistently indicate “healthier” behavior) before entering the model. CFAs were estimated with full-information maximum likelihood for missing data and latent variances fixed to 1. Individual-level factor scores were extracted using the regression method and then standardized to aid interpretation. Higher POS factor scores indicate more health-promoting behavior; higher NEG factor scores indicate more health-risk behavior; higher aggregate factor scores indicate overall “healthier” responding.

*Equal-weight composite indices*. As a complementary summary, we also computed equal-weight indices by averaging the z-standardized items: a POS average score and a NEG average score, and an aggregate average score formed by combining POS items with reverse-scored NEG items. CFA-based factor scores or the z-average indices were used as outcomes were applicable in the main and robustness analyses; patterns are interpreted on a standardized scale.

*Self-report measures.* All interface perception measures have been collected using 7-point Likert scales and are derived based on multiple items. First, to assess the feeling of control, we relied on items to

measure control in computer-mediated communication. Items assess the feeling of control over the interaction, how much users think the interface allowed them to control the interaction, and their feeling of control. The items have been adapted from Wang et al. (2007). Second, we use the subscale of state guilt, adapted from the state shame and guilt scale (Marschall et al., 1994), to measure the current state of guilt as a self-reported post-test measure. Items assess feelings of tension, remorse and feeling wrongful about just conducted actions.

# 4 Results

We organize the results around our dual-pathway model. First, we establish validity of potential constructs based on the question items and whether we can build latent factors for positive items, negative items and a potential aggregate score. Second, we test whether modality directly affects reporting behavior across positively and negatively valenced items (H1a, H1b). Third, we evaluate the two proposed mechanisms: the social-affective pathway via guilt (H2a) and the cognitive-process pathway via perceived control (H2b). Finally, we examine whether these processes differ by gender (H3).

We assessed the construct validity of our POS and NEG scores using CFA. The NEG score demonstrated an acceptable fit. However, the POS score model was unstable (e.g., negative residual variances, Heywood cases), and the factor was nearly synonymous with a single item (exercise frequency). Given this measurement instability, and to guard against over-interpreting this latent score, we proceeded by using the latent factor score for the NEG construct and the more robust, simple equal-weight composite index (z-average) for the POS construct.

## 4.1 Differences between Conditions

We first tested our hypotheses regarding the direct effects of modality on reporting. As shown in Table 2, our results provide strong support for *H1a*. Participants in the Text condition reported significantly higher (i.e., more self-enhancing) POS composite scores than those in the Voice condition ($p<.001$). This main effect was also reflected in the aggregate composite score, which was significantly higher in the Text condition ($p<.001$).

However, we did not find support for a direct modality effect on negatively valenced items, as predicted by *H1b*. There was no significant difference between conditions for either the NEG average score ($p=.439$) or the NEG latent factor ($p=.720$). Finally, we examined the main effect of modality on our proposed mediators. Perceived control was, as expected, significantly higher in the Text condition ($p<.001$). In contrast, perceived guilt showed no reliable difference between conditions ($p=.663$). (All comparisons were made using Student t-tests or Mann-Whitney U tests following checks for normality and homoscedasticity).

| Outcome | Text Mean (SD) | Voice Mean (SD) | p-value | Conclusion |
|---|---|---|---|---|
| POS Composite (z) | 0.163 (0.478) | -0.184 (0.485) | <.001*** | Text > Voice |
| NEG Composite (z) | -0.056 (0.577) | 0.063 (0.688) | 0.439 | n.s. |
| NEG Factor | -0.062 (0.883) | 0.069 (1.120) | 0.720 | n.s. |
| Aggregate Composite (z) | 0.116 (0.407) | -0.130 (0.413) | <.001*** | Text > Voice |
| Guilt (z) | 0.039 (0.947) | -0.044 (1.061) | 0.663 | n.s. |
| Perceived Control (z) | 0.331 (0.820) | -0.371 (1.057) | <.001*** | Text > Voice |

*Table 2. Differences between Conditions in Reporting Measures and Constructs*

## 4.2 Effects of Voice on Reporting Behavior

We now turn to the proposed mechanisms and ask which of the two pathways, social-affective (guilt) or cognitive-process (control), helps explain the observed modality effects. Consistent with the group differences, the voice condition is associated with reliably lower POS scores and aggregate scores via a

direct (non–mediated) path. We then tested whether any effects were mediated via perceived guilt (Table 3) or perceived control (Table 4). We report the average causal mediation effect (ACME), average direct effect (ADE) and total effect.

Guilt did not mediate the condition effect for any outcome. For the POS score and the aggregate score, the direct (non–mediated) effect of voice remained robustly negative. As guilt did not mediate any outcome, *H2a* is rejected.

| **Outcome** | **ACME** | ***p-value*** | **ADE** | ***p-value*** | ***Total*** | ***p-value*** |
|---|---|---|---|---|---|---|
| POS Composite (z) | -0.005 | *0.641* | -0.343 | ***<.001**** | -0.347 | ***<.001**** |
| NEG Factor | 0.019 | *0.616* | 0.110 | *0.502* | 0.129 | *0.440* |
| NEG Composite (z) | 0.008 | *0.650* | 0.110 | *0.296* | 0.118 | *0.296* |
| Aggregate Composite (z) | -0.006 | *0.637* | -0.240 | ***<.001**** | -0.246 | ***<.001**** |

*Table 3. Results of Mediation Analysis with Guilt as a Mediator. ACME = average causal mediation effect (indirect path via mediator); ADE = average direct effect of voice vs. text; Total = combined effect.*

Perceived control showed a moderate but statistically significant indirect path for the NEG latent factor. Indirect paths for the POS score, NEG score, and aggregate score were not significant. For the POS score and aggregate score, the voice effect was again direct (negative ADE). This finding of a significant indirect path for the NEG latent factor provides support for *H2b*.

| **Outcome** | **ACME** | ***p-value*** | **ADE** | ***p-value*** | ***Total*** | ***p-value*** |
|---|---|---|---|---|---|---|
| POS Composite (z) | -0.006 | *0.850* | -0.342 | ***<.001**** | -0.348 | ***<.001**** |
| NEG Factor | 0.161 | ***0.034**** | -0.028 | *0.854* | 0.133 | *0.432* |
| NEG Composite (z) | 0.085 | *0.061* | 0.033 | *0.737* | 0.118 | *0.252* |
| Aggregate Composite (z) | -0.041 | *0.127* | -0.205 | ***<.003**** | -0.246 | ***<.001**** |

*Table 4. Results of Mediation Analysis with Control as a Mediator*

### 4.2.1 Gender as a Moderator in Perceived Control

Given prior evidence that deception propensities and partner models differ by gender (Capraro, 2018; Dreber & Johannesson, 2008), we next examine whether the cognitive-process pathway via control operates uniformly across participants or is moderated by gender (H3). Gender was tested as a moderator of the indirect path from modality condition to reporting behavior through perceived control. Among male participants, the ACME was significant for the NEG latent factor (*p=.011*), the NEG score (*p=.026*), and the aggregate score (*p=.027*); among female and other participants, the ACME was not significant. The direct and total effects for the POS score and aggregate score were negative and significant in both gender groups. This observed difference in the indirect path provides tentative support for H3, though the subgroup sample sizes limit the statistical power of this comparison.

| **Male** | | | | | | |
|---|---|---|---|---|---|---|
| **Outcome** | **ACME** | ***p-value*** | **ADE** | ***p-value*** | ***Total*** | ***p-value*** |
| POS Composite (z) | -0.018 | *0.603* | -0.348 | ***<.001**** | -0.366 | ***<.001**** |
| NEG Factor | 0.196 | ***0.011**** | -0.014 | *0.936* | 0.182 | *0.333* |
| NEG Composite (z) | 0.117 | ***0.026**** | 0.045 | *0.666* | 0.162 | *0.192* |
| Aggregate Composite (z) | -0.062 | ***0.027**** | -0.211 | ***<.002**** | -0.272 | ***<.001**** |
| ***Non-Male*** | | | | | | |
| **Outcome** | **ACME** | ***p-value*** | **ADE** | ***p-value*** | ***Total*** | ***p-value*** |

| | | | | | | |
|---|---|---|---|---|---|---|
| POS Composite (z) | 0.016 | *0.771* | -0.346 | ***<.001**** | -0.330 | ***<.001**** |
| NEG Factor | 0.085 | *0.614* | -0.015 | *0.938* | 0.070 | *0.723* |
| NEG Composite (z) | 0.018 | *0.837* | 0.041 | *0.701* | 0.060 | *0.609* |
| Aggregate Composite (z) | 0.001 | *0.994* | -0.212 | ***<.003*** | -0.210 | ***<.010*** |

*Table 5. Results of Mediation Analysis with Control as Mediator moderated by Gender*

To summarize, our findings provide a nuanced picture of our hypotheses. The voice interface had a strong, direct impact on positively valenced health questions, causing participants to report less favorably than in the text condition; this offers strong support for *H1a*. In contrast, the direct effect on negatively valenced items was not significant, leading us to reject *H1b*. Our mechanism hypotheses showed a similar split. We found no significant differences in perceived guilt, indicating that guilt does not mediate reporting behavior for any items; therefore, *H2a* is rejected. However, we found that voice significantly decreased perceived control, and this reduction in control did indirectly associate with higher (more honest) reporting of negatively valenced behaviors, thus supporting *H2b*. Finally, this indirect effect via control was observed almost exclusively among male participants, providing support for the moderation predicted in *H3*.

# 5 Discussion

We examined how input modality shapes strategic self-reporting in an incentive-compatible setting, with a particular focus on two theoretically motivated pathways: a social-affective pathway via guilt and a cognitive-process pathway via perceived control. Three findings stand out. First, for positively valenced items (where higher values signal "healthier" behavior), speaking yielded reliably lower, less self-enhancing reports than typing. Second, for negatively valenced items (where higher values indicate more "unhealthy" behavior), average differences by modality were small and not statistically reliable; however, reduced perceived control in the voice condition was indirectly associated with higher reporting of negative behaviors. This control pathway was concentrated among male participants, pointing to potential gendered heterogeneity in how process constraints influence disclosure. Third, guilt neither differed by modality nor mediated any effects, suggesting that, in our setting, internal moral emotions did not drive strategic misreporting.

## 5.1 Theoretical Contributions

Our work advances theory on modality and strategic misreporting in three ways. First, by explicitly investigating a social-affective pathway (anticipated guilt) and a cognitive-process pathway (perceived control), we show that the impact of modality is not uniform. We suggest that cognitive constraints, rather than moral emotions, offer potential to explain when voice affects reporting behavior. These findings contribute to growing research on content-driven disclosure to conversational agents (Cox et al., 2025). Voice reduced self-enhancing responding on positively valenced items through a direct pathway, while we found no comparable direct effect for negatively valenced items. Context (e.g. by setting incentives) and sensitive domains lead to different perceptions of conversational agents when disclosing information (Zhang et al., 2024), which our results underscore in the context of voicebots and differently valenced questions. This reframes strategic reporting and disclosure by modality as content-sensitive rather than a single main effect. It also answers long-standing calls in speech interaction research to move beyond broad modality comparisons and theorize when and why speech changes behavior depending on task and content (Clark et al., 2019; Porcheron et al., 2018). For researchers, this implies modeling modality effects at the level of item classes or domains (e.g., health-promoting vs. health-risk behaviors) rather than aggregating everything into a single construct. This item-level modeling also helps distinguish strategic adjustment from modality-induced imprecision. Voice naturally invites approximate language, and small categorical shifts could reflect rounding rather than deliberate distortion. However, the directional consistency of effects on positively valenced items and the asymmetry across valence classes argue against a pure noise explanation, as undirected imprecision would affect both item classes similarly.

Second, we clarify mechanisms by separating social–affective accounts (e.g., audience awareness) from cognitive-process accounts (e.g., immediacy, revisability) manifest in the defined pathways of guilt and control. Contrary to expectations from social presence or moral salience, guilt did not carry modality effects. Instead, perceived control emerged as a pathway for negatively valenced items only. This suggests that cognitive-process constraints can shape reporting even when social–affective cues do not (Clark et al., 2019; Luger & Sellen, 2016; Mahmood et al., 2022). Theoretically, we add a mechanism-level distinction: the curbing of self-enhancing distortion through guilt or audience cues and the constraining of deliberate concealment through process limitations (immediacy, low revisability, and low control) are separable routes that may differentially apply by content valence. This helps connect behavioral outcomes to concrete design levers that research has identified for speech interaction (Mahmood et al., 2022; Porcheron et al., 2018).

Third, we surface heterogeneity in who is most sensitive to process control. The moderated mediation indicates that the control pathway concentrates among male participants. We do not claim essential differences, but our data align with evidence that users carry different characteristics and expectations into speech interactions, which shape evaluation and behavior (Clark et al., 2019; Doyle et al., 2021; Dreber & Johannesson, 2008). However, our theorizing does not specify why male participants would be more sensitive to control constraints than other groups. The observed pattern is consistent with several possible explanations, including differences in deception propensity (Capraro, 2018). Without isolating the mechanism behind this moderation, the findings remain descriptive and should motivate targeted follow-up rather than strong theoretical claims. Still, taking multi-culturality, stereotypes, and varying personality of users into account when designing voice interactions represents a growing need (Chin et al., 2024; Duan et al., 2025; Wenzel & Kaufman, 2024). Therefore, we encourage research models of strategic misreporting that treat users' perceptions of the agent and the user's background as moderators of modality effects, rather than assuming a uniform audience.

## 5.2 Practical Implications

Our results show that modality is not a neutral implementation detail but a consequential design choice in self-report systems that shapes behavioral outcomes in predictable, valence-dependent ways. Voice and text systematically shift how people answer when they have incentives to distort. Organizations that currently treat web forms, chat, and voice as interchangeable channels for collecting self-reported data should therefore reassess these practices.

First, voice interfaces appear particularly useful when organizations want to reduce overclaiming on positively associated behaviors (e.g., healthy habits, compliance, performance). In our setting, voice reduced self-enhancing reports relative to text. This suggests that in domains where inflated self-reports are costly, such as health insurance bonus programs, wellness incentives, or other benefit schemes, assessing positively valenced items through voice can improve data integrity and yield more conservative, and arguably more realistic, inputs into decision processes. Which behaviors feel sensitive to respondents is itself a function of how the reporting task is framed (Tourangeau & Yan, 2007). Designers can therefore influence not only the modality but also the perceived sensitivity of individual items, for instance through normative anchoring or contextual reframing.

Second, the cognitive-process pathway via perceived control indicates a powerful but delicate lever. Lower perceived control in voice was associated with increased reporting of negatively valenced behaviors yet deliberately suppressing control to elicit more candid reports risks undermining user satisfaction and acceptance. Practically, designers should favor lightweight control affordances, such as read-back and confirm prompts, a "repeat or correct last answer" option, or single-step corrections, rather than full transcript editing. These features can restore a sense of agency while still limiting extensive planning and re-crafting that support strategic misreporting in text.

Finally, the preliminary evidence of gendered heterogeneity in the control pathway suggests that one one-size-fits-all modality decisions may not be optimal. Modality choices should be tuned to domain, stakes, and user population, and organizations should monitor both data patterns and user experience across segments. In line with a socio-technical perspective, decisions about when to deploy voice versus

text should be treated as part of the governance of AI-mediated self-reporting, balancing honesty and user satisfaction.

### 5.3 Limitations and Future Work

Although our findings provide valuable insights, several limitations and new research directions warrant further investigation. First, individual ground truth cannot be established in our design. As in most self-report studies, we infer dishonest reporting at the aggregate level rather than verifying each participant's truthfulness. To directly assess accuracy, future work should incorporate verifiable outcomes—e.g., linking to administrative records (with consent), receipt/medical document uploads, or sensor/trace data that can serve as external benchmarks. Complementarily, researchers could adapt well-validated behavioral paradigms that induce opportunities to lie and ensure incentive compatibility to isolate modality effects on strategic misreporting (Weinmann et al., 2022).

Second, we did not model full conversational dynamics. Our interface used voice primarily as an input channel rather than a multi-turn, adaptive conversation. Prior work suggests that backchannels, timing, and prosodic feedback shape social presence; multi-turn voice agents might therefore alter both perceived audience and strategic behavior. Future studies should manipulate conversational features such as agent self-disclosure, empathic acknowledgments, adaptive follow-ups, and timing/latency to test whether responsive dialogue attenuates or amplifies misreporting in incentivized contexts.

Third, the voice interface re-prompted participants when answers were unintelligible, which may have provided incidental revision opportunities and contributed to lower perceived control. Future implementations should log re-prompting frequency per item to assess whether repeated attempts correlate with reported values.

Fourth, we did not leverage behavioral and paralinguistic cues for lie inference. Voice affords rich signals (e.g., pauses, speech rate, pitch variability, disfluencies) that may correlate with deception, but such cues are noisy and context dependent. Future work should (i) pre-register feature sets and analytic plans to avoid fishing, (ii) validate models out-of-sample, (iii) compare acoustic markers with text-linguistic markers from ASR transcripts (hedges, pronouns, negations), and (iv) treat cues as risk indicators rather than ground-truth lie detectors, with careful attention to fairness across accents and speaking styles.

Fifth, we grouped items by a simple positive/negative valence, but the degree of social desirability or stigma likely varies (e.g., *sugary beverages* are less stigmatized than *alcohol use*). Future work should pre-test disclosure items to create a continuous sensitivity score and model its interaction with modality. Because perceived sensitivity is shaped by context, framing and incentive structures (Tourangeau & Yan, 2007), such work should also examine how deliberate design choices shift item sensitivity and thereby moderate modality effects.

Sixth, our design cannot fully separate strategic misreporting from modality-induced imprecision, as voice responses are inherently less exact than typed numerical entries. The valence asymmetry in our results suggests a motivated process, but future work should incorporate objective benchmarks or within-subject designs to quantify precision and strategic components independently.

Finally, domain and stake specificity limit generalizability. We studied health-related reporting for an insurance-like outcome with a modest sample (N=155). Effects may differ in other high-stakes domains (e.g., credit applications, tax claims) or under stronger, real-world incentives and enforcement. Field experiments with organizational partners could test whether modality effects scale with stakes and institutional context. Additionally, our relatively young sample may not capture variation in voice interface familiarity that comes with age. Future work should examine age as a potential moderator of modality effects, as older users may hold different expectations and comfort levels with voice-based interaction.

## 6 Acknowledgements

This work was supported by Innosuisse and is part of the project 104.954 IP-SBM.

## References


Akinnaso, F. N. (1982). On The Differences Between Spoken and Written Language. *Language and Speech*, *25*(2), 97–125. https://doi.org/10.1177/002383098202500201

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145–204. https://doi.org/10.1017/S004740450001037X

Bellman, E. (2017, August 7). The End of Typing: The Next Billion Mobile Users Will Rely on Video and Voice. *Wall Street Journal*. https://www.wsj.com/articles/the-end-of-typing-the-internets-next-billion-users-will-use-video-and-voice-1502116070

Bereby-Meyer, Y., & Shalvi, S. (2015). Deliberate honesty. *Current Opinion in Psychology, Morality and Ethics*, *6*, 195–198. https://doi.org/10.1016/j.copsyc.2015.09.004

Berger, J., Rocklage, M. D., & Packard, G. (2022). Expression modalities: How speaking versus writing shapes word of mouth. *Journal of Consumer Research*, *49*(3), 389–408. https://doi.org/10.1093/jcr/ucab076

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal Deception Theory. *Communication Theory*, *6*(3), 203–242. https://doi.org/10.1111/j.1468-2885.1996.tb00127.x

Capraro, V. (2018). Gender differences in lying in sender-receiver games: A meta-analysis. *Judgment and Decision Making*, *13*(4), 345–355. https://doi.org/10.1017/S1930297500009220

Chafe, W., & Tannen, D. (1987). The relation between written and spoken language. *Annual Review of Anthropology*, *16*, 383–407. https://doi.org/10.1146/annurev.an.16.100187.002123

Chin, J., Desai, S., Lin, S. (Cheng-H., & Mejia, S. (2024). Like My Aunt Dorothy: Effects of Conversational Styles on Perceptions, Acceptance and Metaphorical Descriptions of Voice Assistants during Later Adulthood. *Proc. ACM Hum.-Comput. Interact.*, *8*(CSCW1), 88:1-88:21. https://doi.org/10.1145/3637365

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., Wade, V., & Cowan, B. R. (2019). What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, 1–12. https://doi.org/10.1145/3290605.3300705

Clark-Gordon, C. V., Bowman, N. D., Goodboy, A. K., & Wright, A. (2019). Anonymity and Online Self-Disclosure: A Meta-Analysis. *Communication Reports*, *32*(2), 98–111. https://doi.org/10.1080/08934215.2019.1607516

Cohen, P. R., & Oviatt, S. L. (1995). The role of voice input for human-machine communication. *Proceedings of the National Academy of Sciences*, *92*(22), 9921–9927. https://doi.org/10.1073/pnas.92.22.9921

Cohn, A., Gesche, T., & Maréchal, M. A. (2022). Honesty in the Digital Age. *Management Science*, *68*(2), 827–845. https://doi.org/10.1287/mnsc.2021.3985

Cohn, M., Predeck, K., Sarian, M., & Zellou, G. (2021). Prosodic alignment toward emotionally expressive speech: Comparing human and Alexa model talkers. *Speech Communication*, *135*, 66–75. https://doi.org/10.1016/j.specom.2021.10.003

Cox, S. R., Jacobsen, R. M., & van Berkel, N. (2025). The Impact of a Chatbot's Ephemerality-Framing on Self-Disclosure Perceptions. *Proceedings of the 7th ACM Conference on Conversational User Interfaces, CUI '25*, 1–17. https://doi.org/10.1145/3719160.3736617

Des Jarlais, D. C., Paone, D., Milliken, J., Turner, C. F., Miller, H., Gribble, J., Shi, Q., Hagan, H., & Friedman, S. R. (1999). Audio-computer interviewing to measure risk behaviour for HIV among injecting drug users: A quasi-randomised trial. *Lancet (London, England)*, *353*(9165), 1657–1661. https://doi.org/10.1016/s0140-6736(98)07026-3

Dillman, D. A., & Christian, L. M. (2005). Survey Mode as a Source of Instability in Responses across Surveys. *Field Methods*, *17*(1), 30–52. https://doi.org/10.1177/1525822X04269550

Dorsey, R., Graham, G., Glied, S., Meyers, D., Clancy, C., & Koh, H. (2014). Implementing health reform: Improved data collection and the monitoring of health disparities. *Annual Review of Public Health*, *35*, 123–138. https://doi.org/10.1146/annurev-publhealth-032013-182423

Doyle, P. R., Clark, L., & Cowan, B. R. (2021). What Do We See in Them? Identifying Dimensions of Partner Models for Speech Interfaces Using a Psycholexical Approach. *Proceedings of the 2021 CHI*

*Conference on Human Factors in Computing Systems, CHI '21*, 1–14. https://doi.org/10.1145/3411764.3445206

Dreber, A., & Johannesson, M. (2008). Gender differences in deception. *Economics Letters*, *99*(1), 197–199. https://doi.org/10.1016/j.econlet.2007.06.027

Duan, W., Li, L., Freeman, G., & McNeese, N. (2025). A Scoping Review of Gender Stereotypes in Artificial Intelligence. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25*, 1–20. https://doi.org/10.1145/3706598.3713093

Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*, *47*(4), 1237–1259. https://doi.org/10.3758/s13428-014-0533-4

Hildebrand, C., Efthymiou, F., Busquet, F., Hampton, W. H., Hoffman, D. L., & Novak, T. P. (2020). Voice analytics in business research: Conceptual foundations, acoustic feature extraction, and applications. *Journal of Business Research*, *121*, 364–374. https://doi.org/10.1016/j.jbusres.2020.09.020

Hutchinson, M. K., & Sutherland, M. A. (2019). Conducting surveys with multidisciplinary health care providers: Current challenges and creative approaches to sampling, recruitment, and data collection. *Research in Nursing & Health*, *42*(6), 458–466. https://doi.org/10.1002/nur.21976

Kim, T., Lee, H., Kim, M. Y., Kim, S., & Duhachek, A. (2023). AI increases unethical consumer behavior due to reduced anticipatory guilt. *Journal of the Academy of Marketing Science*, *51*(4), 785–801. https://doi.org/10.1007/s11747-021-00832-9

LaMothe, E., & Bobek, D. (2020). Are Individuals More Willing to Lie to a Computer or a Human? Evidence from a Tax Compliance Setting. *Journal of Business Ethics*, *167*(2), 157–180. https://doi.org/10.1007/s10551-019-04408-0

Lei, S., Xie, L., & Peng, J. (2025). Unethical Consumer Behavior Following Artificial Intelligence Agent Encounters: The Differential Effect of AI Agent Roles and its Boundary Conditions. *Journal of Service Research*, *28*(4), 598–613. https://doi.org/10.1177/10946705241278837

Levinson, S. C. (2016). Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, *20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Luger, E., & Sellen, A. (2016). "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16*, 5286–5297. https://doi.org/10.1145/2858036.2858288

Maharjan, R., Rohani, D. A., Bækgaard, P., Bardram, J., & Doherty, K. (2021). Can we talk? Design Implications for the Questionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent. *Proceedings of the 3rd Conference on Conversational User Interfaces, CUI '21*, 1–11. https://doi.org/10.1145/3469595.3469600

Maharjan, R., Rohani, D. A., Doherty, K., Bækgaard, P., & Bardram, J. E. (2022). What Is the Difference? Investigating the Self-Report of Wellbeing via Conversational Agent and Web App. *IEEE Pervasive Computing*, *21*(2), 60–68. https://doi.org/10.1109/MPRV.2022.3147374

Mahmood, A., Fung, J. W., Won, I., & Huang, C.-M. (2022). Owning Mistakes Sincerely: Strategies for Mitigating AI Errors. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22*, 1–11. https://doi.org/10.1145/3491102.3517565

Marschall, D., Sanftner, J. L., & Tangney, J. P. (1994). State Shame and Guilt Scale (SSGS). *George Mason University*.

Melumad, S. (2023). Vocalizing Search: How Voice Technologies Alter Consumer Search Processes and Satisfaction. *Journal of Consumer Research*, *50*(3), 533–553. https://doi.org/10.1093/jcr/ucad009

Melumad, S., & Meyer, R. (2020). Full Disclosure: How Smartphones Enhance Consumer Self-Disclosure. *Journal of Marketing*, *84*(3), 28–45. https://doi.org/10.1177/0022242920912732

Melzner, J., Bonezzi, A., & Meyvis, T. (2023). Information Disclosure in the Era of Voice Technology. *Journal of Marketing*, *87*(4), 491–509. https://doi.org/10.1177/00222429221138286

Moffett, J. W., Folse, J. A. G., & Palmatier, R. W. (2021). A theory of multiformat communication: Mechanisms, dynamics, and strategies. *Journal of the Academy of Marketing Science*, *49*(3), 441–461. https://doi.org/10.1007/s11747-020-00750-2

Moon, Y. (1998). Impression management in computer-based interviews: The effects of input modality, output modality, and distance. *Public Opinion Quarterly*, *62*(4), 610–622. https://doi.org/10.1086/297862

Moon, Y. (2000). Intimate Exchanges: Using Computers to Elicit Self-Disclosure From Consumers. *Journal of Consumer Research*, *26*(4), 323–339. https://doi.org/10.1086/209566

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '94*, 72–78. https://doi.org/10.1145/191666.191703

Nguyen, H. T., Le, H. T., Connelly, L., & Mitrou, F. (2023). Accuracy of self-reported private health insurance coverage. *Health Economics*, *32*(12), 2709–2729. https://doi.org/10.1002/hec.4748

Oba, D., & Berger, J. (2024). How communication mediums shape the message. *Journal of Consumer Psychology*, *34*(3), 406–424. https://doi.org/10.1002/jcpy.1372

Orlikowski, W. J., & Iacono, C. S. (2001). Research Commentary: Desperately Seeking the "IT" in IT Research—A Call to Theorizing the IT Artifact. *Information Systems Research*, *12*(2), 121–134. https://doi.org/10.1287/isre.12.2.121.9700

Paulhus, D. L. (1991). Measurement and control of response bias. In *Measures of personality and social psychological attitudes* (pp. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, *66*(6), 1025–1060. https://doi.org/10.1111/1467-6494.00041

Petisca, S., Paiva, A., & Esteves, F. (2020). Perceptions of People's Dishonesty Towards Robots. In A. R. Wagner, D. Feil-Seifer, K. S. Haring, S. Rossi, T. Williams, H. He, & S. Sam Ge (Eds.), *Social Robotics* (pp. 132–143). Springer International Publishing. https://doi.org/10.1007/978-3-030-62056-1_12

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice Interfaces in Everyday Life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18*, 1–12. https://doi.org/10.1145/3173574.3174214

Preston, S. H., Fishman, E., & Stokes, A. (2015). Effects of categorization and self-report bias on estimates of the association between obesity and mortality. *Annals of Epidemiology*, *25*(12), 907-911.e2. https://doi.org/10.1016/j.annepidem.2015.07.012

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning, ICML'23*, *202*, 28492–28518.

Redeker, G. (1984). On differences between spoken and written language∗. *Discourse Processes*, *7*(1), 43–55. https://doi.org/10.1080/01638538409544580

Reinfeldt, S., Ostli, P., Håkansson, B., & Stenfelt, S. (2010). Hearing one's own voice during phoneme vocalization—Transmission by air and bone conduction. *The Journal of the Acoustical Society of America*, *128*(2), 751–762. https://doi.org/10.1121/1.3458855

Rogers, T., ten Brinke, L., & Carney, D. R. (2016). Unacquainted callers can predict which citizens will vote over and above citizens' stated self-predictions. *Proceedings of the National Academy of Sciences*, *113*(23), 6449–6453. https://doi.org/10.1073/pnas.1525688113

Roster, C. A., Rogers, R. D., Albaum, G., & Klein, D. (2004). A Comparison of Response Characteristics from Web and Telephone Surveys. *International Journal of Market Research*, *46*(3), 359–373. https://doi.org/10.1177/147078530404600301

Rubin, D. L., Hafer, T., & Arata, K. (2000). Reading and listening to oral-based versus literate-based discourse. *Communication Education*, *49*(2), 121–133. https://doi.org/10.1080/03634520009379200

Rudnik, J., Raghuraj, S., Li, M., & Brewer, R. N. (2024). CareJournal: A Voice-Based Conversational Agent for Supporting Care Communications. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*, 1–22. https://doi.org/10.1145/3613904.3642163

Schuetzler, R. M., Giboney, J. S., Grimes, G. M., & Nunamaker Jr., J. F. (2018). The influence of conversational agent embodiment and conversational relevance on socially desirable responding. *Decision Support Systems*, *114*, 94–102. https://doi.org/10.1016/j.dss.2018.08.011

Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M. (2021). Voice in Human–Agent Interaction: A Survey. *ACM Computing Surveys*, *54*(4), 81:1-81:43. https://doi.org/10.1145/3386867

Sezgin, E., Huang, Y., Ramtekkar, U., & Lin, S. (2020). Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. *Npj Digital Medicine*, *3*(1), 122. https://doi.org/10.1038/s41746-020-00332-0

Singh, M. (2022, March 30). WhatsApp tops 7 billion daily voice messages. *TechCrunch*. https://techcrunch.com/2022/03/30/people-are-sending-7-billion-voice-messages-on-whatsapp-every-day/

Sutton, S. J., Foulkes, P., Kirk, D., & Lawson, S. (2019). Voice as a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, 1–14. https://doi.org/10.1145/3290605.3300833

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral Emotions and Moral Behavior. *Annual Review of Psychology*, *58*, 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145

Thomaz, F., Salge, C., Karahanna, E., & Hulland, J. (2020). Learning from the Dark Web: Leveraging conversational agents in the era of hyper-privacy to enhance marketing. *Journal of the Academy of Marketing Science*, *48*(1), 43–63. https://doi.org/10.1007/s11747-019-00704-3

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Wambsganss, T., Zierau, N., Söllner, M., Käser, T., Koedinger, K. R., & Leimeister, J. M. (2022). Designing Conversational Evaluation Tools: A Comparison of Text and Voice Modalities to Improve Response Quality in Course Evaluations. *Proc. ACM Hum.-Comput. Interact.*, *6*(CSCW2), 506:1-506:27. https://doi.org/10.1145/3555619

Wang, L. C., Baker, J., Wagner, J. A., & Wakefield, K. (2007). Can A Retail Web Site be Social? *Journal of Marketing*, *71*(3), 143–157. https://doi.org/10.1509/jmkg.71.3.143

Wang, L., Huang, N., Hong, Y., Liu, L., Guo, X., & Chen, G. (2023). Voice-based AI in call center customer service: A natural field experiment. *Production and Operations Management*, *32*(4), 1002–1018. https://doi.org/10.1111/poms.13953

Wang, Y., & Shah, C. (2022). Authentic versus synthetic: An investigation of the influences of study settings and task configurations on search behaviors. *Journal of the Association for Information Science and Technology*, *73*(3), 362–375. https://doi.org/10.1002/asi.24554

Wei, J., Jiang, W., Wang, C., Yu, D., Goncalves, J., Dingler, T., & Kostakos, V. (2022). Understanding How to Administer Voice Surveys through Smart Speakers. *Proc. ACM Hum.-Comput. Interact.*, *6*(CSCW2), 548:1-548:32. https://doi.org/10.1145/3555606

Weinmann, M., Valacich, J. S., Schneider, C., Jenkins, J. L., & Hibbeln, M. (2022). The Path of the Righteous: Using Trace Data to Understand Fraud Decisions in Real Time: MIS Quarterly. *MIS Quarterly*, *46*(4), 2317–2336. https://doi.org/10.25300/MISQ/2022/17038

Wenzel, K., & Kaufman, G. (2024). Designing for Harm Reduction: Communication Repair for Multicultural Users' Voice Interactions. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*, 1–17. https://doi.org/10.1145/3613904.3642900

Zhang, Z., Jia, M., Lee, H.-P. (Hank), Yao, B., Das, S., Lerner, A., Wang, D., & Li, T. (2024). "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*, 1–26. https://doi.org/10.1145/3613904.3642385

Zierau, N., Hildebrand, C., Bergner, A., Busquet, F., Schmitt, A., & Leimeister, J. M. (2023). Voice bots on the frontline: Voice-based interfaces enhance flow-like consumer experiences & boost service outcomes. *Journal of the Academy of Marketing Science*, *51*(4), 823–842. https://doi.org/10.1007/s11747-022-00868-5