OXFORD

Interacting with Computers, 2026, **38**, 127–148
https://doi.org/10.1093/iwc/iwae062
Advance access publication date 7 January 2025
Article

Unboxing Manipulation Checks for Voice UX

Katie Seaborn[1,*], Katja Rogers[2], Maximilian Altmeyer[3], Mizuki Watanabe[1], Yuto Sawa[1], Somang Nam[4], Tatsuya Itagaki[1] and Ge 'Rikaku' Li[1]

[1]School of Engineering, Institute of Science Tokyo, Tokyo 152-8550, Japan
[2]Informatics Institute, University of Amsterdam, Amsterdam 94323, The Netherlands
[3]Engineering Sciences, Saarland University of Applied Sciences (htw saar), Saarbrücken 66117, Germany
[4]School of Design, York University, Toronto, Ontario M3J 1P3, Canada
[*]*Corresponding author: seaborn.k.aa@m.titech.ac.jp*

Voice-based interaction is experiencing a second wind through the advent of machine learning (ML) techniques, affordable consumer products and renewed work on natural language processing (NLP) and large language models (LLMs). A growing body of work is exploring how users perceive new forms of computer-generated voices from qualitative and quantitative angles. However, critical voices have called for greater rigour, especially in confirming the voice as a manipulated variable, i.e. manipulation checks. We present three case studies that highlight the value of investing in rigorous manipulation checks for HCI researchers and designers. We demonstrate the importance of testing assumptions, the need for care and reflection in the design of response options and measurement and the advantages of more exploratory approaches to understanding user perceptions of and user experiences (UX) with voice phenomena. Through these case studies, we raise awareness, empirically justify and critically assess the value of manipulation checks for voice UX research and beyond.

RESEARCH HIGHLIGHTS

- Manipulation checks—quantitative, qualitative or mixed—are essential for voice UX projects.
- Expansive response options can reduce response biases and reveal new user perceptions.
- Qualitative and exploratory methods can offer deeper insights and uncover latent voice perceptions.

Keywords: *human-centered computing; voice UX; manipulation checks; voice evaluation; user experience*.

1 Introduction

Technology is rediscovering its voice. User interfaces (UI) and agents in human-computer interaction (HCI) not only understand human speech, but also vocalize feedback and output in ways that are informational and even conversational (Sutton *et al.*, 2019). With the growing popularity of conversational systems (e.g. voice-based virtual assistants like Apple's Siri and Amazon's Alexa) alongside the uptake of voice agents (VAs) in interactive systems (e.g. daily assistants, learning environments, games), how we interact with computers and the resulting user experience (UX) now increasingly features voice components and voice user interfaces (VUIs) of some kind.

Against this backdrop, the design and user perceptions of these voices are of substantial importance. People can unreflexively respond to VAs as if they are people (Druga *et al.*, 2017, Lopatovska & Williams, 2018). For example, users have been found to personify the Amazon Alexa by responding with 'thank you' and gendering it as a 'she' (Lopatovska & Williams, 2018). Yet, existing voice systems are 'limited and homogenised' (Sutton *et al.*, 2019,p. 1], even as voices can take on a dizzying array of attributes like pitch, volume, speed, rhythm/cadence, inflection/intonation and timbre (Seaborn *et al.*, 2021). On the one hand, this variety of vocal characteristics can work in concert to convey social characteristics from the human domain, from personality traits (Nass & Lee, 2000) to age and gender (Iseli *et al.*, 2006, Seaborn & Frank, 2022) to geographic source of origin and social class (McEnaney, 2019). On the other, only a small selection of this diversity tends to be implemented in VAs and VUIs (Sutton *et al.*, 2019). This has been highlighted in critical analyses as a pressing concern for work on voice UX (Seaborn *et al.*, 2021, 2024). Stereotypes abound, for instance, with voice-based virtual assistants cast as young women who chuckle at verbal abuse from end-users (Bergen, 2016). Voice agents that take on the persona of a regional doctor through the use of the local language or dialect and verbalized identity cues ('Welcome to Dr ABC's clinic') may build trust but rely on a certain level of deceit among users not aware that the agent is not a real person, let alone *that* doctor (Joshi, 2014, Seaborn *et al.*, 2024). These and other consequences of employing humanlike voices as design material (Sutton *et al.*, 2019) raise questions of ethics and true human-centredness in the use of voices for machines.

Two challenges underlie this state of affairs. First, the broad design space within voice stimuli creation and selection makes it a *complex domain* for practical implementation and empirical user studies. This may be countered by ascertaining how a given voice stimulus is perceived by users and adjusting specific attributes in a targeted fashion, as needed. However, rather than empirically confirm these assumptions, we tend to *rely on expert selection*

Received: June 7, 2024. **Revised:** November 7, 2024. **Accepted:** December 12, 2024

alone. Specifically, we attribute characteristics to the voice with little or no reported reasoning. We may select voices based on the labels given by the source. We may presume that the social characteristics of the voice actor will transfer to the text-to-speech (TTS) version. We may assume that these characteristics will be perceived by the user as we expect (Jung *et al.*, 2019). We may assume that the identity gestalt that arises from the attribution of these characteristics will fit the user and the task (Stigall *et al.*, 2019). We may overlook related characteristics, such as voice naturalness and favourability, based on common use or our own expert appraisals. This can actualize *researcher bias*: unintentional or unconscious assumptions influencing the research design and results (Carducci *et al.*, 2020). As potential confounds, these intertwined challenges prompt the need for a way to ensure that the assumed perceptions and impact on the UX actually hold.

A *manipulation check* is a rigorous way for us to gain insight into construct validity and, broadly, whether an experimental manipulation has worked as intended (Ejelöv & Luke, 2020, Sigall & Mills, 1998). Manipulation checks are typically run (i) pre-study, separate from pilot testing, (ii) within pilot tests of the main study and/or (iii) within the main study itself. Manipulation checks are a 'fundamental technique' in disciplines like social psychology (Ejelöv & Luke, 2020) and are also on the rise in HCI, e.g. Yurrita *et al.* (2023), and specifically work on voice UX, e.g. Jestin *et al.* (2022), Kao *et al.* (2021). In fields like social psychology, manipulation checks are mostly verbal self-reports, e.g. Likert items and scales, analysed via frequentist significance testing (Ejelöv & Luke, 2020). We are not aware of work on how manipulation checks are generally approached within HCI. Still, given our field's longstanding ties to psychology (Carroll, 1997), it is likely true for us, as well. However, a multidisciplinary field like ours may require or find additional value in more exploratory, qualitative and mixed-method approaches.

In this paper, we present and discuss three case studies through the lens of potential benefits and pitfalls of a variety of manipulation checks for voice UX research. We make a combined *opinion* and *methodological* contribution to voice UX as a field of HCI research (Wobbrock & Kientz, 2016). We present actionable implications for designers and researchers in *the HCI community* and their use of *methodology* (Berkel & Hornbæk, 2023). We make a case for (i) testing even established assumptions about voice stimuli, (ii) the need for care in the design of voice perception response options and (iii) the advantages of exploratory methods to inform our understanding of how users conceptualize specific voice stimuli. For this special issue, we characterize our work as *reflective insights into existing methods* through *case studies and practical applications*. We raise awareness of the role that rigorous designs of manipulation checks can take on in the context of voice UX research. Finally, we discuss the implications for audio studies more broadly and critically assess the need for manipulation checks in this space.

2 Background

2.1 Social Perceptions of Voice Phenomena

Voice, as an expressive medium, has inherent characteristics like volume, pitch and cadence (Sutton *et al.*, 2019). These in turn give rise to a broader set of characteristics when we perceive them. For voices based on human models or otherwise designed to be humanlike, many of these characteristics relate to social percepts (Seaborn *et al.*, 2021, 2024). These can be associated with attitudes and perceptions, such as trustworthiness (Behrens *et al.*, 2018, Seymour & Van Kleek, 2021) and ease of understanding (Dubiel *et al.*, 2024), as well as tied to social identities and constructs, including sex and gender, race and ethnicity, nationality and culture and many more (Seaborn *et al.*, 2021). Voice embodiment, from the visual 'body' to the context and role (Dubiel *et al.*, 2024, Seaborn *et al.*, 2021, Torre *et al.*, 2020), can also be influential. For example, Torre *et al.* (2023) found that a gender ambiguous voice reduced gender-based stereotyping towards robots framed as having stereotypical occupational roles. As they note, this offers promise for shifting known attitudes that many laypeople have towards humanoid robots, especially those with feminine presentations, which are often sexualized (Strait *et al.*, 2017). In short, voice stimuli have technical sonic properties that are linked to social and attitudinal percepts.

Social identities, per social identity theory (SIT), are internal self-concepts that reflect or influence social organizing and group membership (Hogg *et al.*, 1995, Tajfel & Turner, 2004, Tajfel *et al.*, 1979, Turner *et al.*, 1987). The implications for voice-based systems are theoretically broad and yet under-explored. This is due in part to current technical limitations. While theorized as a possibility for future agents, active participation in social identity co-creation is presently the domain of people and possibly other animals (Seaborn, 2022). In human-agent interactions, social identification is one-sided, with the person attributing social identities based on social cues perceivable in the agent that are tied to mental models of social constructs like gender and age (Seaborn, 2022). When machines are said to do 'social identification' work, it is shallow, with identification of visual and other data-derived patterns contingent on the decisions of the human developers—and their biases (Buolamwini & Gebru, 2018, Mehrabi *et al.*, 2021). For example, Buolamwini & Gebru (2018) discovered biases in a camera computer vision algorithm, where a Black woman was only recognized when she wore a white mask. This was traced back to decisions on training the algorithm with a large number of white, male faces, and was subsequently corrected once the creators became aware.

Most agentic machines and their voices are designed to carry social cues or social identities of varying complexity. Experts make decisions about which voice is used for which machine (Jung *et al.*, 2019, Stigall *et al.*, 2019). This decision-making process is often inaccessible and under-reported (Seaborn *et al.*, 2021). Yet, it is crucial to understand how machine voice is perceived and whether the assumptions made by selectors of the voice reflect that of others. A plethora of work has discovered a range of effects, expected and unexpected, related to the social cues embedded in computer voice, from anthropomorphism (Desai *et al.*, 2024, Pradhan *et al.*, 2019, Seymour & Van Kleek, 2021) to stereotyped responses (Bergen, 2016, Hwang *et al.*, 2019, Seaborn *et al.*, 2021, 2024). For example, Hwang *et al.* (2019) discovered stereotypes associated with and sexualization of VAs with feminine voices. We may agree that systems should not be designed to reinforce negative societal attitudes towards human gender, i.e. sexism, but other cases call for more nuance. Chang *et al.* (2018), for instance, explored how likeability of a feminine, extraverted robot voice affected acceptance among an older adult cohort. When removing stereotypes negatively affects the outcome, such as healthcare adherence, we may need to make difficult choices for the sake of the user.

The implications of social and attitudinal characteristics for voice-based computers are theoretically broad and under-explored. A crucial next step will be faithful and complete reporting of voice UX design decisions alongside evidence of alignment with user expectations. For this purpose, we argue that manipulation checks will be a key tool for the HCI community going forward.

2.2 Manipulation Checks

A *manipulation check* generally refers to a method for gaining insight into construct validity and, broadly, whether an experimental manipulation worked as intended (Ejelöv & Luke, 2020, Sigall & Mills, 1998). While the term is often used ambiguously or synonymously with 'attention check,' Ejelöv & Luke (2020) draw a clear distinction between the two terms: manipulation checks focus on the change to a given stimulus, while attention checks assess the degree to which the stimulus is attended to. Manipulation checks are crucial to ensuring that the intended change is accurately and unambiguously achieved, within or among user groups, and potentially reveal why not. For example, a pre-study could present a range of similar voice stimuli, as in Torre *et al.* (2023), to pinpoint the voice that best represents the intended manipulation, in this case gender ambiguity. Thus, manipulation checks are confirmatory and possibly enlightening for the (re)design of studies and stimuli.

Ejelöv & Luke (2020) identify several types of manipulation checks: ones that focus on *(direct) independent variable* measures, *correlates* (indirect measures of the independent variable) and *discriminants* (variables to assess alternative explanations or confounds), while distinguishing attention checks to measure attention in *general* (during an experimental task), focused on *the experimental treatment* (stimulus checks or treatment checks) or on *instructions* (usually to questions after the stimulus).

Manipulation checks are not a panacea and each type has its weaknesses. Gruijters (2022) point out that in-experiment manipulation checks can be subject to measurement error, test invalidity or simply a difference in manipulation effect size sensitivity, making the comparison between a manipulation check and the study's main outcome limited in meaningfulness—and these issues can apply to pre-experiment checks as well.

Recommendations by Ejelöv & Luke (2020) and others (Abbey & Meloy, 2017, Mutz & Pemantle, 2015, Straub & Gefen, 2004, Waltz *et al.*, 1993) emphasize the need to assess the *size* and *meaningfulness* of manipulation effects, test in *new contexts* and comprehensively *report* results. Others warn against viewing non-significant results as evidence of a lack of an effect (Gruijters, 2022). Yet even unsuccessful manipulations can potentially inform the field—if nothing else, to stop others from repeating mistakes (Ejelöv & Luke, 2020). This is a useful but quantitative-leaning starting point for HCI, e.g. in the use of validated scales (Ejelöv & Luke, 2020). Nevertheless, a multidisciplinary field like ours may require or find value in more exploratory, qualitative and mixed-method approaches to manipulation checks. This guided our selection of case studies, which we turn to next.

3 Overview of Case Studies

We now introduce the three case studies we consider in our quest to advocate for manipulation checks when conducting voice UX research. Case Study 1 covers two related pre-studies that reveal the importance of checking assumptions about even clear-cut manipulations. Case Study 2 demonstrates the value of reflecting on and questioning standard response formats. The final case study presents a novel way of conducting manipulation checks with drawing methodology that, through projection rather than prescription, allows for the emergence of unexpected but relevant factors affecting the manipulation. Figure 1 shows how the case studies map onto these separate studies and the corresponding data. 1A results from a larger study that also collected the data used in Case Study 2. 1B presents a follow-up study to 1A. Case Study 3 covers a separate study with a novel methodology. We now summarize each case:

- ***Case Study 1:*** *Pre-Study Assumption Checks About Social Identity and Favourability* (section 4) Pre-study manipulation checks can reveal a mismatch between *researcher assumptions* and *theoretical predictions* on the connection between social characteristics perceived in voice stimuli—here, agedness—and presumed attitudes related to those social features, notably favourability. 'Older' and 'younger' voice stimuli were assessed by older and younger cohorts in an online perceptions study (1A) and pilot study (1B). We discovered that our assumptions about voice agedness and favourability were not consistently met, leading us to invest in the new voice stimuli.
- ***Case Study 2:*** *Expanding Quantitative Manipulation Checks: Response Options and Measurement* (section 5) A staple of manipulation checks in voice UX is using rating scales to assess perceptions of a range of voice qualities and characteristics, including those linked to social identities (Seaborn *et al.*, 2021). While attitudes are shifting, especially in HCI spaces (Bellini *et al.*, 2018, Spiel *et al.*, 2019), many of us still take certain social identity categories for granted, notably gender (Tannenbaum *et al.*, 2019). Gender is a complex social construct that varies by culture, over time and among individuals (Hyde *et al.*, 2019, Tannenbaum *et al.*, 2019). In rating scales prescribed by researchers, participants may be *selecting the closest, or even just any, response option* provided to them, *because they have to choose something* (Seaborn & Frank, 2022, Seaborn *et al.*, 2022, Spiel *et al.*, 2019), resulting in a participant response bias (Furnham, 1986). This case study highlights *response design biases on the researcher side*, a form of measurement bias, through a novel approach to asking about perceptions of gender in voice stimuli that featured expansive rating scale categories (Seaborn *et al.*, 2022). Doing so allowed us to find novel, theory-expanding results on voices perceived as cute: not only were these voices deemed girlish (as expected), but they were also, unexpectedly, 'gender ambiguous'—carrying both feminine and masculine features.
- ***Case Study 3:*** *Extending Qualitative Methods: Drawing Studies as Pre-Study Manipulation Checks* (section 6) The connection between the voice and the 'body' of voiced but potentially 'bodiless' agents, environments and interfaces remains fertile ground for study (Seaborn *et al.*, 2021). Certain methods can provide a way of understanding the mental models of the voice and any 'body,' i.e. *body schema*, it may have. This case study explored what we call *sonic embodiments* through a drawing study (Fleury, 2011, Kearney & Hyle, 2004, Lee *et al.*, 2019b). Sonic embodiment refers to how voice phenomena embedded in interactive media give rise to an imagined body in people's minds, including its form factor, interactive potential and situatedness (Overend, 2022), or social, environmental and historical context. We asked younger and older adults to draw their visions of younger and older voices. The drawing study methodology made clear latent factors in the resulting 'sonic bodies' that we would have missed in traditional manipulation checks.

All studies reported in this paper were approved by the first author's research ethics board (#2023058).

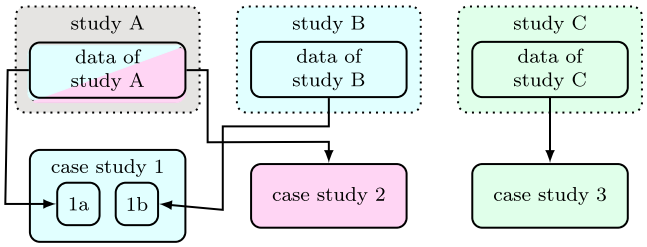

FIGURE 1. Case studies mapped to studies and data sets.

4 Case Study 1: Pre-Study Assumption Checks About Social Identity and Favourability

The first case study comprises two pre-study manipulation checks—a manipulation check study (1A) and a pilot study (1B)—for a VA-based ageism reduction intervention. We aimed to carry out a longitudinal field study, where people of all ages would use the intervention at home every day. The intervention would be designed as a storytelling app with an 'older adult' VA who crafts and reads aloud novel, customized, short tales each night before bed. The goal was to shift negative attitudes about age unconsciously via a mere exposure effect (Zajonc, 1968): repeated, positive experiences with an 'older adult' voice that would build up unconscious, positive associations with age. Critically, this requires (i) the VA to have a perceivable 'older adult' voice and (ii) favourable attitudes towards that voice. This guided our manipulation checks. We also sampled a cross-section of age groups—'younger' and 'older'–to evaluate whether impressions of the voice differed by age. This is because ageism can be internalized by older adults and contingent upon age-group identification (Kornadt & Rothermund, 2012).

We discovered through our manipulation checks in 1A and 1B that our assumptions about voice agedness and voice favourability did not hold true. We had recruited older adults to produce the raw voice material needed for creating the TTS powering the VA's voice. Despite this, our expertise, and 'common knowledge' about 'real' age embedded in voice phenomena, we found that the agedness of voice varied, as did favourability, but not by participant age. We also found that isolating the voice (1A) and conducting the check in the intended interactive scenario (1B) were both valuable to understanding whether and how the manipulations were achieved. We report on the individual pre-studies and then discuss the implications for the case study.

4.1 Pre-Study 1A: A Typical (Isolated) Manipulation Check

We carried out an online user perceptions study following research designs on visual (Nittono *et al.*, 2012) and voice-based (Baird *et al.*, 2018) phenomena. We compared a range of stimuli and recruited a general sample to capture general, relative impressions. Our study protocol was registered before data collection via OSF.[1]

Note that some results were included in the non-archival report (Seaborn *et al.*, 2023) related to Case Study 2 (refer to Figure 1 and section 5).

4.1.1 *Participants and Recruitment*

Participants (N=94; women n=53, men n=37, another gender or N/A n=4) were recruited through Yahoo! Crowdsourcing Japan on December 23rd, 2022. All were Japanese. Most were aged 45–54 (n=34) or 35–44 (n=29), with some younger (18–34 n=15) and older (55–74 n=13). Most were non-users of VAs (n=58). Many were daily or weekly users (n=25), and four used VAs once a month. Another four used to use VAs but did not anymore. Although 100 responses were collected, six incomplete responses were removed. Participants were paid in accordance with the participant pool at 1200 yen per hour, equating to ~300 yen for 20 minutes.

4.1.2 *Procedure*

In a SurveyMonkey questionnaire, participants listened to short (10–15 sec.) clips of computer voices simulating utterances by VAs. They rated the vocal and social qualities of these clips. The clips were presented in a random order to counter novelty and order effects (Schuman & Presser, 1996a). Participants provided demographics on the last page to avoid priming effects (Head *et al.*, 1988). The study took ~20 min.

4.1.3 *Materials*

We used eleven voices from CoeFont,[2] a Japanese TTS provider, and three novel Japanese older adult TTSs (Table 1). A pilot test ($N = 8$) where lab members freely listened to the voice clips and provided qualitative perceptions indicated that the voices carried attributes related to the expected range of ages and genders. For the speech content, we translated the scripts by Baird *et al.* (2018) to Japanese: 'Thank you' as「ありがとうございます」'How are you?' as「おげんきですか」and 'I love you' as「あなたを愛しています」. Each voice 'spoke' all three sentences but in random order.

4.1.4 *Measures*

We cover only the items relevant to this case here (the full questionnaire is in Appendix A). The item order was randomized to curtail order effects (Schuman & Presser, 1996b). All items were translated into Japanese by a native speaker, then back-translated into English and checked by an advanced English speaker and native English speaker. These items were then pilot-tested in-lab with six native Japanese speakers. We measured:

- *Age Perceptions:* Our intervention relied on the voice being perceived as 'old.' and not just 'adult.' Agedness was captured in a nominal scale: infant/baby (0–2 years), child (3–12 years), teen-aged (13–19 years), adult (20–39 years), middle-aged (40–64 years), older adult (65+ years) and ageless.
- *Favourable Impressions:* The intervention required positive impressions of the voice. Favourability of the voice is often used to evaluate VA performance (Seaborn *et al.*, 2021). We used an unvalidated item rated on a 7-point Likert scale ranging from 1=strongly disagree to 7=strongly agree.
- *Attitudes Toward Age:* One way of assessing voice favourability when age/ism may be a factor is through a measure of ageism. This was assessed using the Japanese version (Sawa & Seaborn, 2022) of the Ambivalent Ageism Scale (AAS-JP) (Cary *et al.*, 2017). The 13-item AAS-JP is divided into two modules that measure the level of *benevolence* (nine items) and *hostility* (four items) on a 7-point Likert scale (1=strongly disagree to 7=strongly agree). The original AAS was validated in English while the AAS-JP was localized but not yet validated. Internal scale reliability via Cronbach's α was acceptable ($\alpha \geq 0.7$) (McCrae *et al.*, 2011) for the benevolence factor ($\alpha = .80$) and hostility factor ($\alpha = .73$).

[1] Registered on May 11th, 2023; https://osf.io/49jz5

[2] https://coefont.cloud

TABLE 1. Voices used in Case Study 1A. Average pitch was calculated with the Pitch Detect plug-in for Audacity.

Label	Name	Description	Source	Average Pitch
CF_W_Ana	林アナ	Adult Woman	CoeFont	251 Hz
CF_MM_Yoro	よろこびおじさん	Middle-Aged Man	CoeFont	173 Hz
CF_MM_Oka	岡田斗司夫	Middle-Aged Man	CoeFont	258 Hz
CF_G_Sayo	小夜 SAYO[β]	Girl	CoeFont	314 Hz
CF_G_Nana	なな 9さい	Girl	CoeFont	309 Hz
CF_MM_Oko	怒るおじさん	Middle-Aged Man	CoeFont	247 Hz
CF_M_Taka	高橋 俊輔	Adult Man	CoeFont	216 Hz
CF_W_Asa	あさのゆき	Adult Woman	CoeFont	390 Hz
CF_MW_Yosh	淑江おばあちゃん	Middle-Aged Woman	CoeFont	149 Hz
CF_B_Ken	けんしん	Boy	CoeFont	313 Hz
CF_G_Saku	さくら	Girl	CoeFont	297 Hz
TTS_OW	おばあちゃん	Older Woman	Novel TTS	251 Hz
TTS_OM_Hi		lder Man (Higher-Pitched)	Novel TTS	172 Hz
TTS_OM_Lo		Older Man (Lower-Pitched)	Novel TTS	157 Hz

- *Demographics:* We collected gender (man, woman, non-binary or X-gender, transgender, prefer not to say and/or qualitative input), age (ranges in 5-year chunks from 18 until 74, then 75+), education (from high school to postgraduate, according to the Japanese system) and VA use frequency (every day, several times a week, weekly, monthly, never, used to but stopped, with a reason input, prefer not to say).

4.1.5 *Data Analysis*

We generated descriptive statistics—means (M), medians (MD), standard deviations (SD) and interquartile ranges (IQR)—for all quantitative data. Where possible, we generated counts and percentages. We grouped voices by perceived age (baby, child, teen, adult, ageless) and gender (masculine, feminine, gender ambiguous, gender neutral) based on counts and percent agreement across participants. We conducted normality checks based on skewness, kurtosis and Shapiro–Wilks tests. When these indicated non-normal distributions, we used non-parametric statistics, e.g. Kendall's tau correlations. Otherwise, we used parametric statistics, e.g. Pearson's correlations with Bonferroni corrections.

4.1.6 *Results of 1A*

All clips generated from the three 'older adult' TTSs were generally deemed 'older adult' in age, with agreement scores ranging from 76.% for the woman (TTS_OW), 85.1% for the higher-pitched man (TTS_OM_Hi) and 68.1% for the lower-pitched man (TTS_OM_Lo). Two middle-aged CoeFont voices were categorized, to some degree, as 'older adults,' namely CF_MM_Oko and CF_MW_Yosh, with ratings of 19.5% and 90.4%, respectively. Still, given that the male voice (CF_MM_Oko) was generally rated as 'middle-aged' (69.1%), it was not perceived as 'older.' However, the female voice (CF_MW_Yosh) was, which had only 7.4% 'middle-aged' ratings. We include both as a point of comparison to the novel TTS voices.

Favourability of all voices was middling (Table 2). Weak, positive correlations were found between benevolent ageism (M=3.4, SD=0.8, MD=3.3, IQR=1.1) ratings and (favorable) impressions towards CF_MM_Oko (r =.219, p ¡.05), CF_MW_Yosh (r =.157, p ¡.05), TTS_OW (r =.234, p ¡.05), TTS_OM_Hi (r =.220, p ¡.05) and TTS_OM_Lo (r =.233, p ¡.05). No correlations were found between hostile ageism (M=4.1, SD=1.0, MD=4.0, IQR=1.4) and these impressions.

TABLE 2. Descriptive statistics for favourability of the voices (Case Study 1A).

Voice	Source	*M*	*SD*	*MD*	*IQR*
CF_MM_Oko	CoeFont	2.9	1	3	2
CF_MW_Yosh	CoeFont	2.5	0.8	3	1
TTS_OW	TTS	2.8	0.9	3	1
TTS_OM_Hi	TTS	2.8	0.9	3	1
TTS_OM_Lo	TTS	2.7	0.8	3	1

4.1.7 *Discussion of 1A*

Voice agedness varied for the 'older adult' voices. This raised questions about our desired voice manipulation: even the most highest-performing voice —the 'higher-pitched' older man (CF_OM_Hi)—was rated by 85.1% as 'older.' While this represents majority opinion, some still rated the voice as belonging to a younger (albeit adult) age group. Moreover, favourability was middling. Still, negative attitudes towards age did not seem to affect these impressions, contradicting the literature (Kornadt & Rothermund, 2012, Palmore, 1999). Moreover, the one result connecting these impressions to benevolent ageist attitudes was weak. We wonder if this was because the voices were experienced in isolation, divorced from the sort of embodied interactions representative of typical human–human experiences and indeed many human–machine experiences. Context matters, including for agent voice perceptions (Mendelson & Aylett, 2017, Torre *et al.*, 2020). As Mendelson & Aylett (2017) argued, conducting voice perception studies with the voice in isolation (outside of the intended context of use) reduces ecological validity and may lead to spurious results. We surmised that direct interaction with the voice in agent form could more readily tap into mental models of age and older folks. We thus conducted the next pre-study with an interactive version.

4.2 Pre-Study 1B with the Interactive Versions of the Voices: Assumption Check Mayhem

We evaluated two 'younger' and 'older' VAs in an in-person interactive scenario: personalized storytelling. Our goal was to evaluate the interactive version of the most stable of the novel 'older adult' voices—the woman—in a realistic context of use. We also wished to explore specific age groups—younger and older cohorts—to determine if participant age influenced perceptions

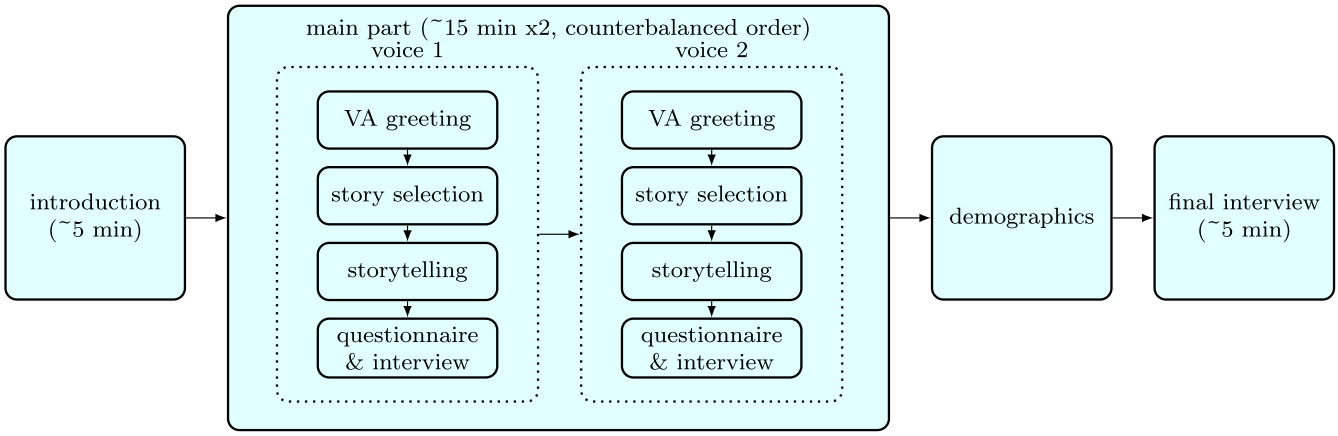

FIGURE 2. Study flow for the interactive voice agent user study (Case Study 1B).

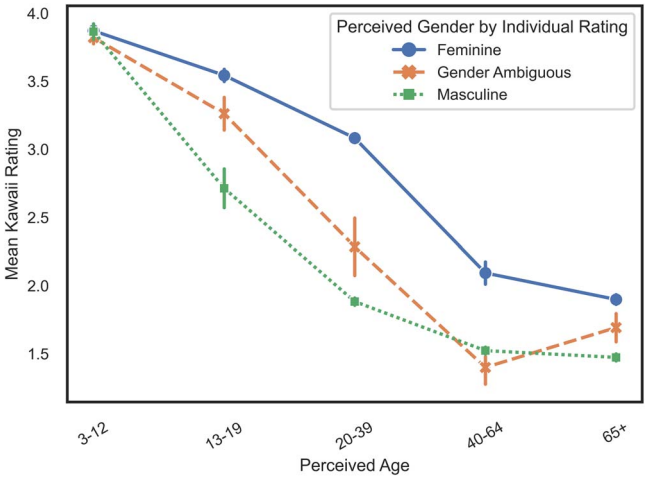

FIGURE 3. Individual perceptions of gender linked to kawaii ratings (Case Study 2).

of voice age (Kornadt & Rothermund, 2012). The methods were the same as for 1A, except as noted.

This case study was initially published in non-archival form (Sawa *et al.*, 2023).

4.2.1 *Participants and Recruitment*

Of the 34 people, 16 were younger (9 men, 7 women, none of another gender; all university students and working adults in their 20s; recruited by word-of-mouth) and 18 were older (aged 60+; 9 men, 9 women, none of another gender; recruited from a local silver society).

4.2.2 *Procedure*

Participants were greeted and informed about the study, and then signed consent. Each participant had two sessions with the VA to experience two voices (Figure 2). They sat at a table and viewed a display with the VA system (refer to 4.2.3). One researcher hosted the session while another controlled the VA remotely via Wizard of Oz (WoZ). The voice of the VA was switched from 'younger' to 'older' in a counterbalanced fashion across participants. The VA read a short story, then asked the participant to fill out the questionnaire. A brief interview with a researcher took place after each session (5 min) and at the end of the study (5 min). The study took 40–60 mins. with a ~10-min. break between voices.

4.2.3 *System Design*

We chose the feminine 'older adult' Japanese TTS (`TTS_OW`) for the 'older' version of the interactive agent based on its favourable reception and deemed agedness in 1A, as well as its technical stability as a TTS (the other voices did not generate consistent output). For the 'younger' one, we chose the default Japanese feminine voice used by Google Assistant (`ja-JP-Wavenet-B` feminine TTS,[3] with an average pitch of 333 HZ). A web-based WoZ app was created using clips generated by the TTSs for the activity . The VAs carried out the activity in the same way, with the only difference being the TTS used, i.e. the agedness of the voice.

4.2.4 *Activity*

The VA experience consisted of a storytelling activity to inspire dreams before bed. The 'dream' stories were selected based on an online survey with young and old Japanese people (N=49; 25 younger and 24 older adult participants) that covered recent, most memorable and most desired dreams.[4] Based on these, we selected six 5-minute stories from an online library[5] containing elements identified in the survey findings (e.g. the presence of family or an animal). Participants met the VA and selected a story in a conversational format. The story was read aloud by the VA. After a questionnaire, interview and break, the VA explained that it wished to try on a new 'storyteller' voice and asked the participant to select another story, which it then read aloud in the second voice.

4.2.5 *Measures*

After each story, the VA asked participants to fill out a questionnaire. They were also interviewed by a researcher about their impressions. We built upon the measures in 1A, introducing novel and validated scales, unless noted.

- *Agedness:* We used the same items as in 1A (4.1.4).
- *Satisfaction, Trust, Familiarity:* All were measured with single-item 7-point Likert scales (1=strongly disagree to 7=strongly agree) used in previous VA work (Lee *et al.*, 2019a).
- *Likability and Anthropomorphism:* The 5-item likability (α =.982) and anthropomorphism (α =.937) subscales of the Godspeed questionnaire were used (Bartneck *et al.*, 2009), each measured on a scale of 1=strongly disagree to 5=strongly agree. We modified the items to align with VAs, as per previous research

[3] https://cloud.google.com/text-to-speech/docs/voices
[4] Details provided in the supplementary materials (Appendix B).
[5] Aozora Bunko (Blue Sky Library): https://www.aozora.gr.jp

(Seymour & Van Kleek, 2021); a lab pilot test (N=7) confirmed the linguistic changes.

- *Attitudes Toward Age*: We used the AAS-JP (refer to 4.1.4). In line with 1A, the internal reliability of the translated AAS-JP scale was acceptable for the benevolence factor (α =.70), although it did not reach the recommended threshold of ($\alpha \geq$ 0.7) (McCrae *et al.*, 2011) for the hostility factor (α =.37).
- *Demographics*: We used gender, age, education and VA use frequency (refer to 4.1.4).
- *Post-Story Interviews:* Participants were asked: 'What is your general impression of the VA?'
- *Final Interview:* After the second post-story interview, we asked four more questions: 'What voice qualities differentiated the two VAs?,' 'Did your impression of the VA change between the first and second sessions?,' 'If you had to choose one of the VAs to read you a bedtime story, which would you prefer?' and 'Please tell us the reason for this preference.'

4.2.6 *Data Analysis*

We followed the quantitative analysis methods in 1A (4.1.5). Due to recording errors, the data from two younger participants was lost, resulting in $N = 32$ for this analysis.

For the qualitative data analysis, one Japanese native researcher conducted a thematic analysis of the open-response interview data. They drew on several aspects of the reflexive thematic analysis approach by Clarke & Braun (2021), notably flexible coding, no predefined codebook, coding and theme development alone and drawing on their reflexive knowledge of the Japanese context, as well as aspects of a codebook approach, notably use of frequencies (quantification), discussion of themes with the first author for refinement and clarification and associating themes with specific voices and age groups to explore quantitative differences.

After familiarization with the data, they applied open coding to create a set of potential themes. These were examined, narrowed down, combined and redefined iteratively. For this, they followed Patton (1990), using the criteria of 'internal homogeneity' and 'external heterogeneity' to ensure that the data within each theme were semantically homogeneous and that themes were clearly distinguished. The resulting coarser categories were designated as 'themes' and the finer categories as 'subthemes.' Theme definitions and names were devised. Then, frequencies for themes and sub-themes were generated. For this, the participant age group (older and younger) and VA age (older and younger) were used as grouping variables. We also used Chi-squared tests of independence to evaluate differences in frequencies, following examples in HCI (Poeller *et al.*, 2023) and other fields (Bajaj & Reed, 2022, Hochard *et al.*, 2017), with Bonferroni corrections applied due to multiple comparators (voice agedness and participant age group). We note that these Chi-squared tests are exploratory and should not be considered objective or generalizable.

4.2.7 *Results of 1B*

Descriptive statistics for the quantitative voice rating measures for all participants are presented in Table 3, and by voice age category and participant age group in Table 4.

This time, the 'older woman' voice (`TTS_OW`) was deemed less old, with only 16 (47.1%) attributing an age of 60+. Most other attributions were 'adult' (11 counts or 32.4%). The younger voice (`ja-JP-Wavenet-B`) was deemed young (in the 20–30s, with 26 counts or 76.5%).

Measures for the 'young' and 'old' voices did not significantly differ statistically within or across younger and older participants. However, satisfaction with the 'younger' voice ($M = 5.1, SD = 1.4$) was statistically significantly higher than the 'older' voice ($M = 4.24, SD = 1.71$) , $t(34) = -2.508, p = .017, d = .578$.

Descriptive statistics for ageism are in Table 5; no statistically significant relationships between measures or differences by VA or participant age group were found.

Qualitative findings (Table 6) revealed a range of impressions on the quality of the voices, the characteristics of the VA and its varying agedness and the storytelling activity. Comparisons showed statistically significant differences by voice age for themes of loudness, smoothness, clarity, coldness, ease of listening and compatibility with the activity. The 'younger' voice was favoured in terms of loudness, clarity and smoothness. However, the 'older' voice was better received in terms of warmth (the younger voice was perceived as cold), relaxation and compatibility with the activity. Younger participants found the voice slow but also familiar.

4.2.8 *Discussion of 1B*

We explored the role of age perceptions when the voice was given an agent form and embedded within an interactive scenario. Yet, the novel 'older woman' voice (`TTS_OW`) was deemed less old than in 1A, even though the context was deemed more suited to this voice than the younger voice (`ja-JP-Wavenet-B`). As such, we conclude that we cannot rely on contextual interpretations of voice age (Torre *et al.*, 2020). This led us to develop a new set of voices with new voice actors that sounded older and were favourable to all ages for the main study.

Despite what we expected based on similarity-attraction theory (Byrne, 1969), impressions of the younger and older VAs were not linked to participant age or ageism attitudes. Also, despite the suitability-to-task results, all tended to prefer the 'younger' voice in terms of TTS quality. For younger people, this preference was linked to familiarity with the voice (the default for Google Assistant) and its perceived higher technical quality as a TTS. Still, we did not capture confirmatory fluency measures, which we recognize as a limitation of this study.

Participants deemed the 'older' storyteller (`TTS_OW`) more appropriate for the activity, a relaxing voice that was not cold, unlike the younger voice. Yet, these impressions were not linked to benevolent ageism, against expectations. Also, as noted, they did not relate to voice agedness in the expected way. These results also contrast with other work[6] suggesting that older adults may strongly prefer younger-sounding VAs. One reason could be that the ageism scale we used was only checked for reliability with older adults (Sawa & Seaborn, 2022) and remains unvalidated. Notably, the original scale was not tested with older adults, only young and middle-aged people (Cary *et al.*, 2017). Our results may then be a matter of localization validation or generational differences in ageism.

Another possibility is complementary-attraction, where those with different but compatible traits are preferred (Gurtman, 2009). Future work may explore what traits these could be. A further possibility is positive ageism, defined by positive stereotypes about the capabilities and roles of older adults in society (Chonody, 2016): grandparents are known to pass down wisdom through stories or read aloud to grandchildren at bedtime. Future work may include other measures of attitudes toward age and interview

[6] https://doi.org/10.21203/rs.3.rs-1905540/v1 (preprint)

TABLE 3. Descriptive statistics of measures by voice age for *both* groups overall (Case Study 1B). *$p < .05$.

	'Young' Voice (`ja-JP-Wavenet-B`)				'Old' Voice (`TTS_OW`)			
Measure	**M**	**SD**	**MD**	**IQR**	**M**	**SD**	**MD**	**IQR**
Satisfaction	**5.1***	1.4	5	2	4.2	1.7	4	2
Trust	5.3	1.4	5	3	4.9	1.5	5	2
Familiarity	4.4	1.8	5	2.8	4.4	1.9	4.5	3
Likability	3.6	0.9	4	1	3.2	1	4	1
Anthropomorphism	2.7	0.8	2.8	1.4	2.7	0.9	3.6	1.4

TABLE 4. Descriptive statistics of measures by voice age and participant age group (Case Study 1B). 'Young' voice: `ja-JP-Wavenet-B`. 'Old' voice: `TTS_OW`.

	Young Adults								Older Adults							
	'Young' Voice				'Old' Voice				'Young' Voice				'Old' Voice			
Measure	**M**	**SD**	**MD**	**IQR**	**M**	**SD**	**MD**	**IQR**	**M**	**SD**	**MD**	**IQR**	**M**	**SD**	**MD**	**IQR**
Satisfaction	5.4	1.2	6	1.3	4.8	1.8	5	3	4.7	1.6	5	2.8	3.7	1.5	4	2
Trust	5.6	1.3	6	2.3	5	1.4	5	1.3	5.1	1.6	5	2.8	4.9	1.7	5	2.8
Familiarity	4.2	1.9	4	2.5	4.8	2	5.5	2.3	4.6	1.8	5	1.8	4	1.8	4	2.8
Likability	3.5	1	3.6	1.5	3.4	1	3.7	1.4	3.6	0.9	3.5	1.3	3.1	1	3.1	1.4
Anthropomorphism	2.6	1	2.4	1.3	2.9	0.8	3.4	1.4	2.7	1.1	2.8	1.7	2.4	0.8	2.6	1.2

TABLE 5. Descriptive statistics for the ageism measure via the AAS-JP (Case Study 1B).

	Benevolent				Hostile			
Group	**M**	**SD**	**MD**	**IQR**	**M**	**SD**	**MD**	**IQR**
Both	3.1	0.8	3.2	0.9	3.4	0.9	3.5	1.2
Younger	3.4	0.6	3.3	0.8	3.4	0.8	3.6	0.8
Older	2.9	0.8	3.1	1	3.4	1	3.5	1.5

questions on the connection between VA role, agedness and the activity.

4.3 Discussion Synthesis for Case Study 1

We explored whether certain generalizations would hold—that real age would link to age perceptions in voice—and certain stereotypes would be true—that people with ageist attitudes would be less favourable to older-sounding voices compared to those less ageist—based on previous research and established theory. Yet, this was not the case in 1A and 1B. Whether the voice was isolated (1A) or placed in an interactive context with a VA form (1B), its agedness and favourability were mixed. Notably, the theorized expectations relating to ageism were weak or absent. Instead, voice perceptions seemed linked with perceived fit to the VA task; at the same time, this did not boost perceptions of voice agedness. This shows that social identities are dynamic, pluralistic and socially constructed (Hogg *et al.*, 1995, Tajfel & Turner, 2004, Tajfel *et al.*, 1979, Turner *et al.*, 1987).

Without these checks, we would have proceeded with a broken manipulation. This would have failed to produce a result for the intervention, which rested on a perceptible 'older adult' voice stimulus plus high favourability for the voice. This case cements the need for care—to test even obvious assumptions, like voice actor age directing the perceived age of the TTS-derived voice—when voice is used as a manipulation. Moreover, checks should not just involve voice perception tests but also the interactive context, VA role and the measures upon which we may later rely to explore the effects of or on these perceptions.

4.3.1 *Implications for Manipulation Checks*

Case Study 1 reveals several upsides and downsides:

Merits:

- We can identify issues with assumptions about voice material and how these may be measured before the study. This can lower the risk of failure for the main study or platform and provide valuable information on how (and how much) to modify the voice stimuli or measurement. Notably, the labels attributed to voice material, such as age, may not hold and need to be tested in user perception studies.
- Comparing isolated voice perception of the stimuli alone to contextual results (i.e. embedded in the VA and interaction scenario) in separate pre-studies can highlight the (in)consistency of percepts across settings and elicit attitudinal responses. The role of context (or its lack) can be verified.

Demerits:

- Running one or more pre-studies requires more resources, time and energy than a single study alone. Still, quicker and cheaper online studies can be fruitful (as in 1A) and pre-studies may be crucial for determining next steps (such as redoing the voice).

5 Case Study 2: Expanding Quantitative Manipulation Checks: Response Options and Measurement

A staple of manipulation checks in voice UX is the rating scale (Seaborn *et al.*, 2021). A host of voice characteristics are presumed important and commonly assessed, including gender and sex, age, personality, emotion, accent and dialect and a complex factor variably known as anthropomorphism, humanlikeness, machine-likeness and artificiality. In general, these factors can be linked

TABLE 6. Themes and subthemes by voice age and participant age group with Bonferroni-corrected Chi-squared test results (Case Study 1B). *p ¡.05. **p ¡.01. Y or 'Young' voice: `ja-JP-Wavenet-B`. O or 'Old' voice: `TTS_OW`.

			By Voice Age				By Age Group			
Themes	**Subthemes**	**Characteristics**	**Y**	**O**	$\chi 2$	p	**Y**	**O**	$\chi 2$	p
Voice Qualities	Talking Speed	Slow	0	4	4.27	.078	**4**	0	5.49	**.038***
	Pitch	High	**6**	0	6.62	**.020***	2	4	0.29	1.178
		Low	0	4	4.27	.078	2	2	0.07	1.589
	Inflection	Natural	5	1	2.94	.173	2	4	0.29	1.178
		Unnatural	3	10	4.73	.059	5	8	0.19	1.334
	Pronunciation	Clear	**9**	1	7.59	**.012***	5	5	0.19	1.329
	Speaking Style	Smooth	**5**	0	5.42	**.040***	2	3	0.03	1.721
		Stilted	1	5	2.94	.173	5	1	4.22	.080
	Hoarseness	Hoarse	0	3	3.15	.152	1	2	0.14	1.419
VA Characteristics	Unreliability	Unreliability	1	3	1.07	.603	2	2	0.07	1.589
	Affect	Emotional	1	4	1.95	.325	2	3	0.03	1.721
		Unemotional	7	1	5.14	.047	5	3	1.31	.506
	Anthropomorphism	Humanlike	4	8	1.64	.200	6	6	0.68	.822
		Machinelike	13	11	0.37	1.085	10	14	0.31	1.154
	Warmth	Warm	1	2	.35	1.109	1	2	0.14	1.419
		Cold	**5**	0	5.42	**.040***	2	3	0.03	1.721
	Friendliness	Friendly	4	7	0.99	.640	5	6	0.02	1.801
Activity UX	Relaxation	Relaxed	1	**9**	6.34	**.024***	4	6	0.46	.994
	Compatibility with the Activity		0	**7**	7.00	**.008****	5	2	1.23	.533
	Familiarity Compared to Known VAs		4	0	4.27	.078	**4**	0	5.49	**.038***
	Ease of Listening	Easy to Listen	**14**	3	9.69	**.004****	7	10	.529	.467
		Hard to Listen	1	5	2.94	.173	2	4	0.06	1.606

Y: Younger Voice/Age Group. O: Older Voice/Age Group.

to demographics and social identities as well as special characteristics predicted for computer-generated or computer-based voice phenomena. The work of Baird *et al.* (2017, 2018) represents rigorous and typical investigations of the extent to which such basic social characteristics are conveyed by TTS voices.

Yet, these factors are more complex than we might expect. While attitudes are shifting (Bellini *et al.*, 2018, Spiel *et al.*, 2019), many of us still take certain social identity categories for granted. Sex and gender and the complex relationship between the two is a prominent example (Tannenbaum *et al.*, 2019). Everyone has a mental model of sex/gender that feels obvious and unquestionable. Nevertheless, sex and gender are complex biological and social constructs that vary by culture and over time, as well as among individuals (Hyde *et al.*, 2019, Tannenbaum *et al.*, 2019). Moreover, we often unreflexively rely on simple models for these constructs, i.e. the gender or sex binary of masculine/male and feminine/female, and associated stereotypes (Seaborn & Frank, 2022, Seaborn *et al.*, 2022, Spiel *et al.*, 2019).

A range of research on computer voices has found that participants do tend to *rely on simple models and stereotypes* of sex/gender and other variables, e.g. (Behrens *et al.*, 2018, Lee *et al.*, 2000, 2019a, Nass *et al.*, 1997, Sundar & Nass, 2000). Yet, participants may simply be *selecting the closest, or even just any, response option* provided to them, *because they have to choose something*. Much has been written about participant response biases (Furnham, 1986). For example, we may overly rely on agreement statements, fueling the issue of acquiescence, a well-known phenomenon in self-reports where participants tend to agree to statements independent of their content (Knowles & Nathan, 1997). Subsequently, they may agree to contradictory statements, ultimately leading to uninterpretable results. Other examples include loaded questions (with emotional wording), answer options in surveys that do not fully represent the range of possible answers, or dependent variables measured in the same order, inducing order effects (Blankenship, 1942). What we highlight in this case study is *response design biases on the researcher side*, a form of measurement bias.

This case study was previously published in non-archival form (Seaborn *et al.*, 2023).

5.1 Theoretical Background for Case Study 2

Kawaii is a Japanese term that can roughly be translated as 'cute,' yet covers a broader set of cultural meanings and associations. Nittono and colleagues were instrumental in establishing the science of kawaii (Nittono, 2010, Nittono *et al.*, 2012). Kawaii is a sociocultural phenomenon with psychological and behavioural implications (Nittono, 2016, Nittono & Ihara, 2017). Kawaii has been treated as a visual and/or physical property, i.e. a matter of appearance that can be visually perceived. Yet, there could be kawaii *vocalics*: the paralinguistic or 'meta' qualities of voice, such as pitch, volume, rate of speech, verbal fillers and timbre convey information beyond speech content—notably emotion, personality and social identity qualities, such as gender and age (Poyatos, 1993). This led us to ask: **RQ1.** *Can voice alone evoke kawaii?*

Shiokawa (1999) argued that kawaii is a dynamic and vague construct, malleable to individual modes of expression. Although some have characterized it as gender-neutral in modern times (Nittono, 2016, Urakami *et al.*, 2021), kawaii is stereotyped as 'girlish' (young and feminine). Notably, Shiokawa (1999) criticized the use of kawaii as infantilizing when applied to adult women. We thus wished to check: **RQ2.** *Do perceptions of voice age and gender relate to voice kawaii?* Whether the notion of kawaii as young and feminine holds true for voice remained unknown. This led us to hypothesize:

> **H1.** *Perceptions of voice gender will be linked to kawaii in terms of femininity, i.e. kawaii is gendered feminine.*

Research on photos of older adults smiling (Nittono & Ihara, 2017) and trends like 'otona-kawaii' (adult-kawaii) (Lieber-Milo, 2021) suggest that kawaii could be an ageless phenomenon for voice, as well. This led us hypothesize:

> **H2.** *Perceptions of voice age will not be linked to kawaii, i.e. kawaii voice is an ageless phenomenon.*

Still, the notion of kawaii as 'girlish' implies an intersection between gender and age, i.e. feminine and young. This is not unprecedented: social identities typically are intersectional (Crenshaw, 2017). If the stereotype holds true for voice, then:

> **H3.** *Perceptions of voice age and gender will intersect such that young and feminine voice perceptions will be linked to perceptions of kawaii, e.g. the stereotype of kawaii as girlish.*

5.2 Methods for Case Study 2

We conducted an online voice perceptions study. This was the same study as Case Study 1A, but here we cover the measures related to *kawaii*. The protocol was registered via OSF.[7] For details on participants (N=94), procedure and materials (set of fourteen voices), refer to the description of this study in Case Study 1A (in 4.1).

5.3 Measures

In addition to the below, we used the *age perceptions* data described in Case Study 1A (refer to 4.1).

5.3.1 *Perceived Gender*

Genderedness was captured in an expansive nominal scale (Seaborn *et al.*, 2022): feminine, masculine, aspects of both femininity and masculinity (hereafter: ambiguous), neither femininity nor masculinity (hereafter: neutral) and an open-ended field for another option.

5.3.2 *Kawaii Perceptions*

Since there was no validated instrument, we used a one-item 5-point scale of agreement (1=strongly disagree to 5=strongly agree) on voice kawaii-ness. Given that it was a one-item scale, we operationalized it across participants as a mean greater than 3.5 (skewed towards agreement on kawaii-ness) *and* a median of 4 or above (nominal agreement). We will note marginal cases, where only one or the other of these metrics were met.

5.4 Results of Case Study 2

Descriptive statistics indicated that several voices were deemed kawaii by participants (**RQ1**): the 'teenaged girl' `CF_G_Sayo` (M=3.6, SD=1, MD=4, IQR=1); the 'young girl' `CF_G_Nana` (M=3.9, SD=0.9, MD=4, IQR=0); the 'young boy' `CF_B_Ken` (M=3.7, SD=0.9, MD=4, IQR=1); and the 'young girl' `CF_G_Saku` (M=3.7, SD=0.8, MD=4, IQR=1).

We now consider the **RQ2** hypotheses and qualitative findings. Note that voice gender was classified in two ways: by consensus per sample, and by individual ratings.

[7] Registered on January 20th, 2023; https://osf.io/49jz5

5.4.1 *H1. Perceptions of voice gender will be linked to kawaii in terms of femininity*

Based on consensus ratings for each voice, most voices were deemed either feminine or masculine by a majority of participants. A Mann-Whitney U test indicated that the voices rated as feminine (by consensus) (MD=3, IQR=2) were more kawaii than voices rated as masculine (by consensus) (MD=2, IQR=1), U = 73926, p ¡.05.

For individual ratings regardless of voice, a Chi-Squared test found a statistically significant association between Perceived Gender and Kawaii, $\chi^2(8, 1307) = 337.19$, $p < .05$, $\phi = .359$. The gender neutral and other options were removed due to a violation of the assumption for the Chi-Squared test with insufficient data points. Tests were calculated with adjusted significance level: $\alpha_{\text{adjusted}} = \frac{0.05}{3} = .017$. Posthoc Chi-Squared tests (Bonferroni corrected) showed relationships between Perceived Gender (by individual ratings) and Kawaii ratings for all categories: ambiguous (M=3.4, SD=1.2, MD=4, IQR=1.8), feminine (M=3, SD=1.2, MD=3, IQR=2) and masculine (M=1.9, SD=1, MD=2, IQR=1); refer to Figure 1. Specifically, the Chi-squared tests results were: (ambiguous, feminine) = $\chi^2(4, 700) = 21.60$, $p < \alpha_{\text{adjusted}}$, (ambiguous, masculine) = $\chi^2(4, 741) = 198.08$, $p < \alpha_{\text{adjusted}}$ and (feminine, masculine) = $\chi^2(4, 1173) = 256.89$, $p < \alpha_{\text{adjusted}}$.

Based on the consensus ratings, we can partially accept the hypothesis: voices rated as feminine (by consensus) tended to be rated kawaii. However, based on the descriptive individual ratings, when voices were rated as gender ambiguous (by individuals) then those individuals also tended to rate them highly for kawaii.

5.4.2 *H2. Perceptions of voice age will not be linked to kawaii*

A moderate, negative correlation was found between Perceived Age and Kawaii, $\tau b = -.547$, $p < .01$. A Chi-Squared test found a significant relationship between Perceived Age and Kawaii ratings, $\chi^2(16, 1308 = 720.91$, p ¡.05, $\theta = .372$. A Kruskal–Wallis test indicated a significant difference by Age, $\chi^2(5 = 635.16$, p ¡.05, with a Dunn's test (Bonferroni corrected) revealing significant differences across all categories except for two combinations: middle aged–older adult, and child–teen: child (M=3.8, SD=.9, MD=4, IQR=1), teen (M=3.6, SD=1, MD=4, IQR=1), adult (M=2.5, SD=1, MD=2, IQR=1), middle-aged (M=1.6, SD=.7, MD=1, IQR=1) and older adult (M=1.7, SD=.8, MD=1, IQR=1). Thus, we cannot accept the hypothesis; kawaii appears to be an age-based phenomenon, favouring youth.

5.4.3 *H3. Perceptions of voice age and gender will intersect: younger-sounding feminine voices and kawaii*

Chi-Squared tests found a significant relationship between Perceived Age and Kawaii rating for voices (by individual ratings), feminine, $\chi^2(16, 561 = 273.01$, p ¡.05), masculine voices, $\chi^2(16, 607 = 423.11$, p ¡.05), and ambiguous voices, $\chi^2(16, 134 = 81.05$, p ¡.05). A Kruskal–Wallis test revealed the same for Perceived Gender (by individual ratings), $\chi^2(2 = 308.078$, p ¡ 0.05, with a Dunn's test (Bonferroni corrected) showing significant differences for all Perceived Genders (by individual ratings). Similarly, a Kruskal–Wallis test for voice and gender classifications (by consensus) revealed significant differences in Kawaii ratings, $\chi^2(6 = 723.255$, p ¡.05. A Dunn's test using a Bonferroni correction indicated this for all pairs except for masculine adult–feminine older adult and feminine child–feminine teen: feminine child (M=3.8, SD=.9, MD=4, IQR=1), feminine teen (M=3.6, SD=1.0, MD=4, IQR=1), feminine adult (M=3.0, SD=.9, MD=3, IQR=1), masculine adult (M=2.0, SD=.8,

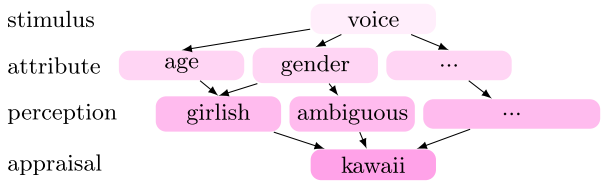

FIGURE 4. Preliminary model of kawaii vocalics with computer voice and social identity extensions, as it pertains to age and gender (Case Study 2). We note that this is a partial model, as other attributes and perceptions may also apply.

MD=2, IQR=1), feminine older adult (M=1.9, SD=.8, MD=2, IQR=1.3), masculine middle-aged (M=1.6, SD=.7, MD=1, IQR=1), masculine older adult (M=1.5, SD=.7, MD=1, IQR=1). We can partly accept the hypothesis: the stereotype of 'girlishness' with an element of gender ambiguity.

5.5 Discussion of Case Study 2

Our expansive approach to gender and consideration of individuality led to novel discoveries for kawaii and an extension of the existing model. Without this manipulation check *and* this approach to the check, we would have missed these key findings.

The voices deemed kawaii tended to be perceived as 'girlish,' confirming stereotypes (Inuhiko, 2006, Shiokawa, 1999). However, some of the highest kawaii ratings were attributed by individuals to voices they also deemed *gender-ambiguous*: having a mix of feminine and masculine qualities. This may be explained by Shiokawa's (Shiokawa, 1999) characterization of kawaii as ambiguous and individualistic. In contrast to previous work involving images of older folks (Nittono & Ihara, 2017) and the otona-kawaii trend (Lieber-Milo, 2021), 'older adult' voices were not considered kawaii, and the most kawaii voices were perceived as young. Taken together, these findings can be assembled into a preliminary model of kawaii vocalics that includes the social identity characteristics of age and gender (Figure 4).

Key to our purpose is the notion that an expansive approach to response options, especially gender (Seaborn *et al.*, 2022), can reveal unexpected but valid user perceptions of voice phenomena. Rather than replicating the binary response options of research past, we asked about gender more broadly, including ambiguity and neutrality. The result was a novel finding for kawaii voice perceptions, showing that voices are perceived as kawaii when considered feminine but also sometimes when considered gender ambiguous. An important aspect of this result is that it is based in individual differences, where a selection of individuals participating in the study perceived different voices as kawaii and gender-ambiguous. HCI has long grappled with the rather amorphous concept of 'individual differences' (Dillon & Watson, 1996, Egan, 1988). Here we have a case of a shared concept—gender-ambiguous kawaii vocalics—among distinct people across different voice stimuli. This shows how choice of measurement can engender perceptions that may have implications for voice more broadly. Expanding our approach to conducting voice perception work may inform nearby disciplines, such as sociolinguistics, and prompt interdisciplinary collaborations.

We would be remiss not to mention our reliance on several single-item and unvalidated measures, however justified or well-used in previous research. The means of voice measurement need to be considered carefully (Seaborn & Urakami, 2021)—this also holds for manipulation checks. Yet in this example, standardized measures for kawaii did not exist at the time; and we must also balance the length of the questionnaire. Future work should seek to produce robust, validated, internally consistent instruments for computer voice features.

Here, we outline the merits and demerits revealed by this case study for the use of manipulation check studies generally, and the careful design of response options in manipulation checks specifically.

Merits:

- Rich data in manipulation check studies can already lead to novel insights about voice phenomena.
- Expansive response options allow us to more precisely operationalize real-world phenomena, leading to potentially more accurate results.
- When conducted through an online survey, manipulation checks are simple and pose low burden for researchers and participants alike, and can be replicable in other questionnaires.

Demerits:

- Although expansive compared to prior work, the response options are still prescriptive by nature and potentially limited; determining the most appropriate response options can be tricky. As part of the expansive approach (Seaborn *et al.*, 2022), we included an open-ended item, but this data was small and limited. Exploratory qualitative research or participatory methods to generate response options and larger sample sizes could help to with this.
- More expansive response options may not always be feasible or practical because increasing the number of items or response options can also lead to confusion or overwhelm participants.
- Caution: Participants may also disagree with the provided response options, such as for political reasons, and submit troll responses in open-ended fields (Rogers & Weber, 2019).

6 Case Study 3: Extending Qualitative Methods: Drawing Studies as Pre-Study Manipulation Checks

The connection between the voice and the 'body' of voiced but potentially 'bodiless' computers remains fertile ground for study (Seaborn *et al.*, 2021). Experts make choices about what voice goes with what body. But what do participants imagine, if anything, when they hear the voice? What is the source orientation (Guzman, 2019, Sundar & Nass, 2000): the room? The speaker? The air? Somehow inside or outside of these contexts? A social being or merely a computer agent? Grasping how users imagine the voice that they hear is an exercise in understanding user mental models of the voice and any 'body,' i.e. *body schema*, it may have. This can require exploratory methods, including in manipulation check studies—e.g. through a drawing study.

While rare in HCI (Fleury, 2011, Lee *et al.*, 2019a), drawing studies have a long history elsewhere, including art therapy (Guillemin, 2004, Malchiodi, 2006, Naumburg, 1966), child psychology (Allen & Butler, 2020, Chan, 2006, Jolley & Thomas, 1995, McNeal & Ji, 2003, Piaget & Inhelder, 1969) and personality and emotion research (Kearney & Hyle, 2004, Machover, 1949). Drawings allow people of all ages to express ideas regardless of language ability, providing insight into state of mind, visual memory, cognitive ability (Jolley & Thomas, 1995, Piaget & Inhelder, 1969), mental representations, perceptions, values and preferences (McNeal & Ji, 2003). Notably, Seaborn *et al.*

(2024) highlight drawing studies as an emerging exemplar for voice UX. In one example, Lee *et al.* (2019a) found that the envisioned personas of bodiless VAs were affected more by voice characteristics than conversational scripts. Hence, drawing studies can unearth latent percepts that users may not directly raise and experts may not expect or think to evaluate in typical manipulation checks.

Here, we explored the drawing study method as a manipulation check of *sonic embodiment*: the imagined body to which voice phenomena embedded in interactive media give rise, including form factor, interactive potential and situatedness (Overend, 2022), or the social, environmental and historical context. We extended the work of Lee *et al.* (2019a) by explicitly considering social characteristics and identities. We asked younger and older adults to draw their visions of younger and older voices. The goal was to ensure that a novel 'older adult' TTS was imagined as an 'older adult.' One's own self-image and body schema may influence visions of 'others' (Westen, 1988) based on social identities such as age, based on one's own situatedness (Overend, 2022). Indeed, the drawing study made clear other latent factors in these 'sonic bodies' that we would have otherwise missed in traditional manipulation check or voice perception studies.

6.1 Methods for Case Study 3

Our drawing study had a within-participants design, where all participants drew all voices. Our protocol was registered in advance on OSF.[8]

6.1.1 *Participants*

We recruited 34 Japanese people. 18 identified as men and 16 as women; no one of another gender identity was recruited. There were younger people in their 20s and people aged 65+: 11 people were 18–24 years old, four were 25–34 years old, 12 were 65–74 years old and seven were 75+. We used Jikken Baito to recruit younger people[9] and recruited older adults through a local silver society. Six had used voice assistants for more than one year, while the other 28 had never used any. Participants were paid 1200 yen in cash or an Amazon Gift Card.

6.1.2 *Materials*

We used four TTS-generated voices representing younger and older voice cohorts. Two were the default Japanese Google Assistant voices: the `ja-JP-Wavenet-B` feminine Google Assistant voice used in 1B and the default masculine voice, `ja-JP-Wavenet-D` (with an average pitch of 196 HZ). Two were the feminine (`TTS_OW`) and higher-pitched masculine (`TTS_OM_Hi`) older adult voices used in the other case studies.

6.1.3 *Procedure and Data Collection*

The procedure is shown in Figure 5. Participants were greeted and informed about the study, and then signed the consent form. Participants sat at a table and put on headphones. They listened to one of the voices telling a story for ~2 minutes. Then they were asked to draw the body of the voice using a black pen on white paper. There were no time restrictions. This was repeated for all four voices, each reading a different story in omniscient form. We randomized the order of the voices and what story each voice spoke. After experiencing each voice, we conducted a debriefing interview about the drawings to ensure that we fully understood what was drawn (Lee *et al.*, 2019a). We asked, 'What did you draw?' and 'Why did you draw it?' At the end, they provided their demographics.

6.1.4 *Data Analysis*

Two researchers conducted a hybrid (inductive and deductive) applied thematic analysis (Guest *et al.*, 2011) of the drawings. The analysis was led by a Japanese native. After looking through the drawings several times, this rater developed an initial set of codes *inductively*, using their knowledge of voice UX factors, notably the frameworks developed by Seaborn & Urakami (2021), e.g. gender, accent, personality, social identity theory (Hogg *et al.*, 1995, Tajfel & Turner, 2004, Tajfel *et al.*, 1979, Turner *et al.*, 1987) and impression formation linked to stereotypes (Baird *et al.*, 2018, Nass *et al.*, 1997), as well as commentary in the debrief interview. They went through several rounds of coding to produce an initial codebook. The rater then grouped these codes into higher-level themes. They also *deductively* added the themes originally developed by Lee *et al.* (2019a): 'human,' 'speaker,' 'system' and 'space object' (e.g. satellites). Next, the two raters separately coded ~30% of the drawings using the inductive and deductive themes. The raters first took a *semantic* approach, focusing on the clear, descriptive characteristics in the drawings. Then, they considered *latent* patterns by taking in the whole picture and considering the interview comments. Edge cases were discussed to determine inclusion or exclusion.

6.2 Results of Case Study 3

A total of 136 drawings were made. We now describe and show the sonic embodiments imagined by participants.

6.2.1 *'Body' Anthropomorphism*

97 drawings (71%) were coded as 'human,' 19 (14%) 'speaker,' 10 (7%) 'robot' and seven (5%) 'system' (Figure 6). There were no drawings for 'space object,' and none of headphones, even though participants experienced the voices through headphones.

6.2.2 *Agedness*

Examples of the range of agedness found in the drawings are shown in Figure 7. The older voices tended to be drawn as old and the younger voices young. Nevertheless, a range of age-ambiguous and ageless embodiments were also drawn. Most of these were non-humanoid but some held humanlike features; through the interviews, we understood that these were anthropomorphic but indeterminable or flexible in the details.

6.2.3 *Genderedness*

Feminine voices tended to be given feminine form factors, and vice versa for masculine voices (Figure 8). Yet, some gender and age perceptions intersected in complex and ambiguous ways. Two examples are shown for gender ambiguity in Figure 8 (third column): the top being for the older feminine voice (`TTS_OW`) and the bottom being for the older masculine voice (`TTS_OM_Hi`), though both suggest the features of a girl or young woman.

6.2.4 *Cultural Identity (Latent Factor)*

An unexpected latent factor was voice perceptions that were clearly linked to certain cultural identities. Cues were present in the drawings, but we confirmed our understanding in the interviews. In the end, we found that 68 (50%) were drawn to be 'Japanese,' 21 (15%) were one of 'Chinese,' 'Asian' or 'not Japanese,' and 47 (35%) were of no culture or national identity. Notably, all 21 drawings coded as foreign to the (Japanese) participants were from

[8] Registered on July 4th, 2022 at https://osf.io/xes23
[9] https://www.jikken-baito.com

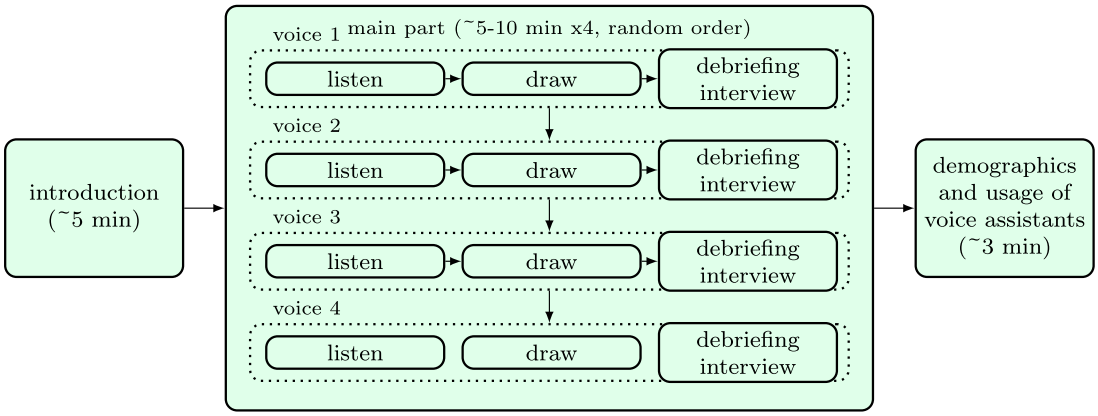

FIGURE 5. Study flow for the drawing study (Case Study 3).

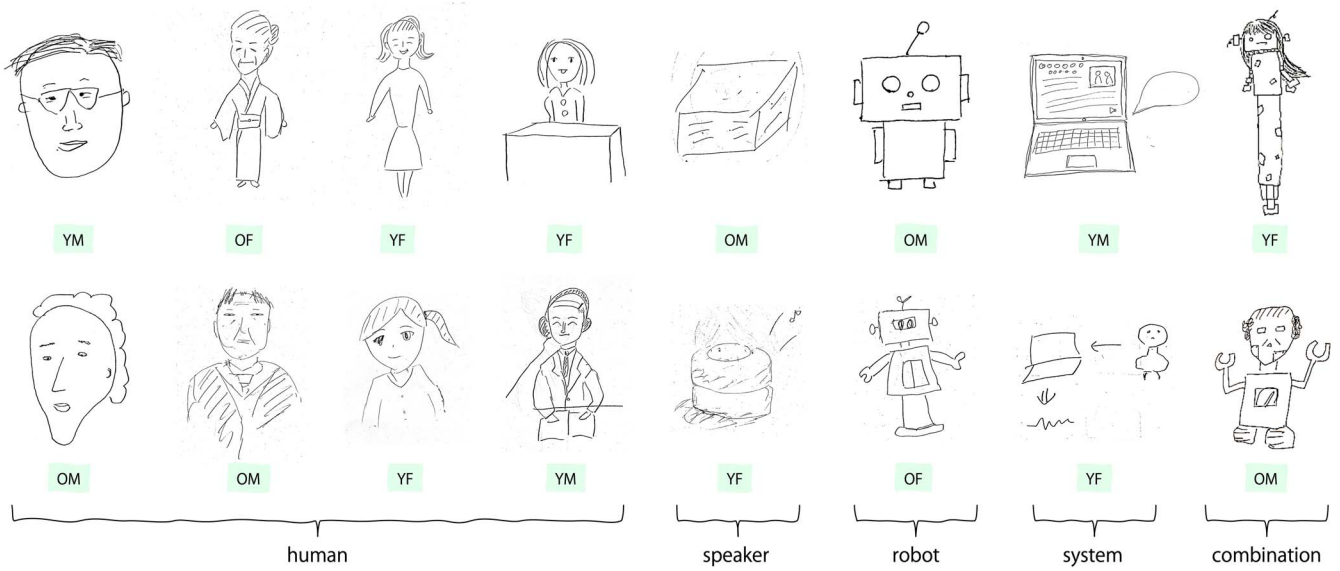

FIGURE 6. Sonic embodiments drawn by participants (Case Study 3). YM: Young masculine TTS (`ja-JP-Wavenet-D`). YF: Young feminine TTS (`ja-JP-Wavenet-B`). OM: Old masculine TTS (`TTS_OM_Hi`). OF: Old feminine TTS (`TTS_OW`).

older voices. The interviews indicated that this traces back to the gender ambiguity of the older voices but also to the differences in TTS quality between our novel TTS and the commercial Google Assistant product.

6.2.5 *Familiarity (Latent Factor)*

The interviews indicated that the drawings were influenced by preexisting models and relationships known only to the illustrator (Figure 9). One participant reported recognizing one younger voice as a Google voice (`ja-JP-Wavenet-B`), so he drew a smart speaker, a clear case of legacy bias (Morris *et al.*, 2010). Another felt that the older feminine voice (`TTS_OW`) was similar to the voice of a person he knew, so he drew that person.

6.2.6 *Environment (Latent Factor)*

Although presented as belonging to virtual agents, some participants drew environmental or spatial embodiments (Figure 9). The interviews suggested that they perceived the voice as *in* something rather than the thing itself. Still, rather than a device, as in prior work (Guzman, 2019), they placed the voices in space somewhere. Notably, some spaces were natural and some were constructed, e.g. a map of the planet or a physical building.

6.3 Discussion of Case Study 3

Critical analysis of the drawings revealed patterns both expected and unexpected. Some people imagined very different ages or genders for the voices: deviating from the impressions of the majority, the identity of the source of the voice (i.e. the voice actor's age and gender) and researcher expectations (i.e. individual differences). We also discovered an unexpected result: perceptions of 'cultural identity' in certain voices, apparently linked to TTS quality (i.e. noise and verbal fluency), different from accent and dialect. This has various implications. Designers who wish to create a foreign-sounding voice may do so simply by reducing the quality of the TTS.

Importantly, we would not have found this result without the 'blank slate' of drawing methods, absent of all preconceptions except what mental models the participant brought to the table. Typical quantitative forms of measurement for manipulation checks would not have captured such perceptions, nor would we have thought to include custom items to capture such latent factors before running this study. With drawing methodologies, people use their imaginations freely, whereas prescribed response options and predetermined measures restrict the format of user responses. Although these exploratory aspects generalize to qualitative methods more broadly, drawing as a qualitative research

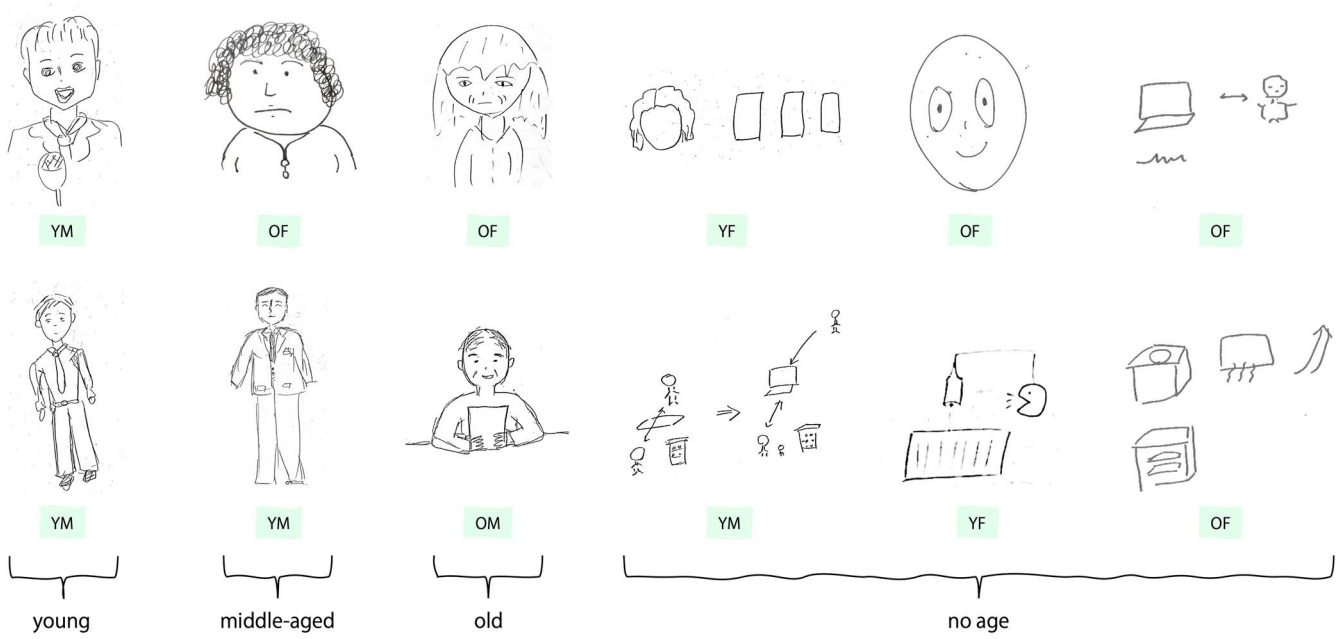

FIGURE 7. Agedness of the voices (Case Study 3). YM: Young masculine TTS (`ja-JP-Wavenet-D`). YF: Young feminine TTS (`ja-JP-Wavenet-B`). OM: Old masculine TTS (`TTS_OM_Hi`). OF: Old feminine TTS (`TTS_OW`).

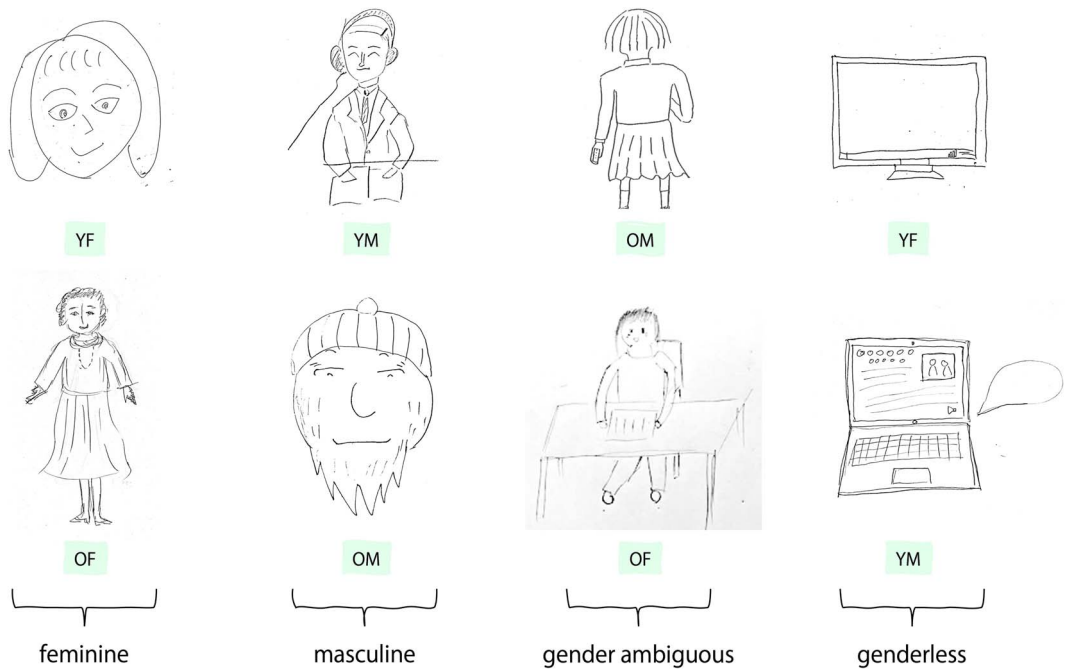

FIGURE 8. Genderedness of the voices based on researcher interpretation and participant explanations (Case Study 3). YM: Young masculine TTS (`ja-JP-Wavenet-D`). YF: Young feminine TTS (`ja-JP-Wavenet-B`). OM: Old masculine TTS (`TTS_OM_Hi`). OF: Old feminine TTS (`TTS_OW`).

method specifically (Fleury, 2011, Guillemin, 2004, Kearney & Hyle, 2004) has been linked to uncovering latent constructs in the imagination, stimulating deeper reflection among participants and increasing participants' comfort during the study.

Our hybrid thematic approach had limitations. The inductive approach, led by a Japanese native versed in voice UX, was fruitful. Induction can affect reliability, but this can be addressed. Here, we used multiple raters, and one was not Japanese. We can suggest using crowdsourcing procedures to assess generalizability at a larger scale or employing quantitative measures—at the same time or in a follow-up manipulation check—to assess the surfaced themes in a complementary way (if generalizability aligns with the epistemic framework). Finally, we relied on one rater to develop and choose all codes. Future work could explore flexible index coding (Deterding & Waters, 2018), which could surface new patterns during the coding process by multiple raters.

We now outline a list of merits and demerits highlighted by this case study for the drawing method.

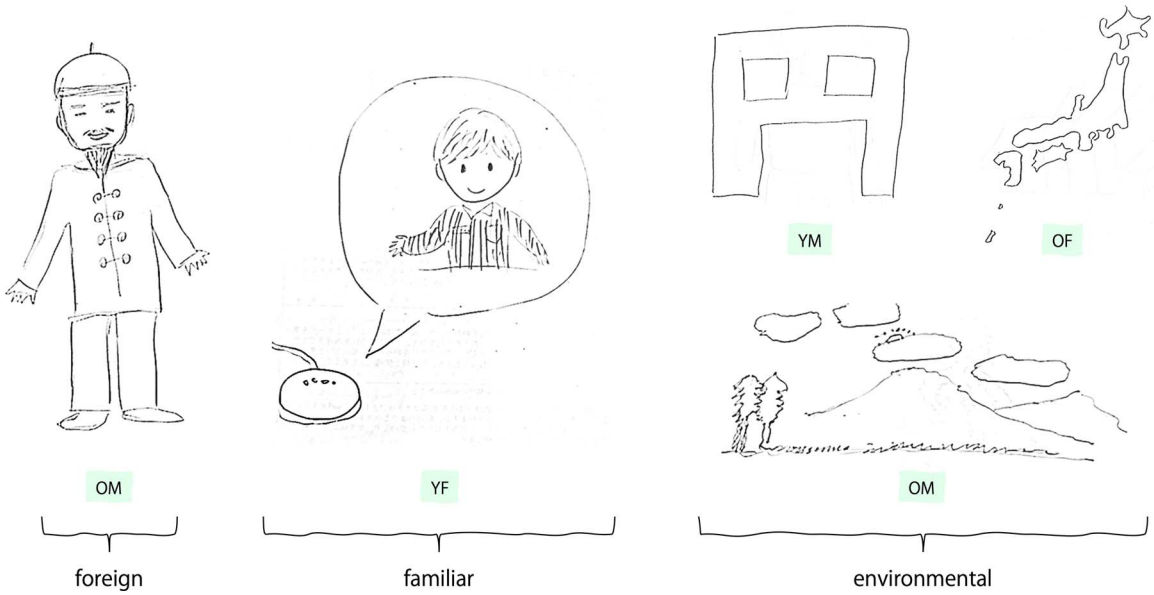

FIGURE 9. Unexpected embodiments of the voices (Case Study 3). YM: Young masculine TTS (`ja-JP-Wavenet-D`). YF: Young feminine TTS (`ja-JP-Wavenet-B`). OM: Old masculine TTS (`TTS_OM_Hi`). OF: Old feminine TTS (`TTS_OW`).

Merits:

- An absence of priming and constraints in materials and instruments reduces the influence of researcher preconception; participants have a literal blank slate.
- Affords detection of implicit or latent factors that may not yet have theoretical or literature grounding, for rigour and the development of new instruments.
- Similar to other methods, allows for identification of the extent to which voice features, known and latent, vary across individuals and populations.

Demerits:

- Data analysis can be quicker than for other qualitative data, but longer than quantitative manipulation checks. Other elicitation methods, such as interviews or annotations, may be required to fully understand the drawings and unearth latent factors that cannot be inferred through the drawings alone, as per established practice for participant-generated visuals (Guillemin, 2004, Guillemin & Drew, 2010). This again adds time and effort.
- Such projective methods may generate unexpected and hard-to-quantify data (Donoghue, 2010). Efficacy depends on its purpose. For example, drawings and annotations could be taken as introspective truths when designing voices for a customizable system.
- The number of drawings per participant and session may be limited due to time constraints. Thus, the check may not address all aspects of the voice UX (e.g. omitting parts of a use case).
- The difficulty of drawing can vary widely based on the complexity of the voice and context of use, as well as participant comfort with drawing.

7 Overall Discussion

Manipulation checks have become recognized as an essential part of experimental rigour in HCI-adjacent research (Fiedler *et al.*, 2021) that may be translatable to HCI and notably voice UX work. At the same time, critical voices have raised several looming issues, such as overly relying on significance testing of single variables (Ejelöv & Luke, 2020), blatantly asking participants about a manipulation (Fiedler *et al.*, 2021) and the checks affecting participant thinking (Hauser *et al.*, 2018). This raises questions of validity while 'wrongly enhanc[ing] subjective confidence' (Fayant *et al.*, 2017,p. 125]. More work and discourse is needed to illuminate the value of manipulation checks and establish best practices. In voice UX, these conversations still need to be held.

Our three case studies form the first step of this conversation. Through these cases, we have demonstrated the importance and potential severity of manipulation checks for voice UX research. We acknowledge that manipulation checks can be problematic, depending on their use (Fayant *et al.*, 2017). Still, simply employing complex voice stimuli as design material or as experimental manipulations without understanding how the voices are perceived is ill-advised. Yet, manipulation checks are rare, or at least rarely reported (Clark, 2019, Seaborn *et al.*, 2021, Seaborn & Urakami, 2021). Voice UX researchers should carefully consider when and how to ***include*** and ***design*** manipulation checks with rigour (Fiedler *et al.*, 2021, Hauser *et al.*, 2018). We have offered several ways in our case studies. We discuss these alongside higher-level considerations next.

7.1 Manipulation Checks for Testing Assumptions

In Case Study 1, we showed how assumptions about the underlying social characteristics and perceptions of voice (in this case, favourability and ageist attitudes) may not hold true. This reflects the concept of researcher bias (Carducci *et al.*, 2020) and the issue of unwarranted assumptions in the context of (missing) manipulation checks (Fayant *et al.*, 2017, Fiedler *et al.*, 2021). We need rigorous assumption testing and to avoid decisions based on stereotypes, anecdotal evidence or simplifications of scientific findings, such as work suggesting that feminine voices are preferred (Tolmeijer *et al.*, 2021). Instead of assuming that the social characteristics of the voice we select will be universally perceived as intended, match the context (Stigall *et al.*, 2019, Torre *et al.*, 2020) or fit the target users well (Jung *et al.*, 2019), we should *evaluate*

these assumptions. While much of the focus in VUI work has been *what is said and how*, i.e. speech content, we need to delve deeper into *how things are being said and by 'who,'* i.e. voice, as well (Cambre & Kulkarni, 2019, Seaborn *et al.*, 2021). Voice perceptions are complex and dynamic, dependent on the interplay of user, device and contextual characteristics (Cambre & Kulkarni, 2019).

The results for the ageism instrument should also give us pause. Despite being deemed valid and reliable for English cohorts (Cary *et al.*, 2017) and showing initial reliability in the Japanese localization (Sawa & Seaborn, 2022), the instrument lacked robustness in our context of use. Whether caused by WEIRDness (Henrich *et al.*, 2010) or something else, such problems can only be discovered and rectified through a process of openness, replication and trial-by-error. Manipulation checks may be instrumental in this. As we showed, care is needed when designing *how* a manipulation is to be tested, as well as in confirming the manipulation itself through such testing.

7.2 Manipulation Checks with Expansive Response Options

Expanding the standard way of assessing manipulation checks quantitatively may lead to novel perspectives on seemingly well-understood phenomena. Case study two is a clear example. Here, voices deemed gender-ambiguous—perceived as having a mix of feminine and masculine qualities—received among the highest kawaii ratings, challenging assumptions about kawaii voices being 'girlish.' This finding was only made possible through adoption of a different, more expansive operationalization of gender rating scales, including options for gender ambiguity and neutrality. Thus, valid and holistic operationalizations of relevant variables, even non-manipulated ones, are essential for manipulation checks, as postulated by Fiedler *et al.* (2021). Importantly, we need to check our assumptions and biases as researchers who decide on, if not design, perception instruments. This will require understanding our own ideas, values and mental models reflexively (Rode, 2011). This may also require changing our instruments or adopting new ones. Resources are emerging, such as the HCI Guidelines for Gender Equity and Inclusivity,[10] which are notably centered on the human identities of participants rather than the humanlike attributes that could be perceived in voices. Methodological research will need to craft new ways of capturing perceptions that keep up with the dynamism of the human experience.

7.3 Manipulation Checks through Holistic and Qualitative Means

The third case study highlights the benefit of using a variety of qualitative methods to better understand the mental models people have about voice agents, especially to unveil latent variables that may play an important role in voice perception. Here, we refrained from 'standard' manipulation checks methods. Instead, we answered demands for more holistic ways of approaching manipulation checks (Fiedler *et al.*, 2021, Hauser *et al.*, 2018) and drew inspiration from less common research methods (Fleury, 2011, Lee *et al.*, 2019a).

The specific method of a drawing study has only rarely been used for the purpose of manipulation checks in past research: the only prior example that we could find was a check for drawing quality, in the main study, through the presence of required patterns (Schleinschok *et al.*, 2017). However, qualitative and mixed methods approaches to manipulation checks are not without precedent. Participants can be prompted to provide open-ended descriptors or accounts that can bring clarity around thought processes and the success of experimental priming (Avey *et al.*, 2011, Rood, 2011). Avey *et al.* (2011) used a combination of quantitative and qualitative checks to ensure that the leadership conditions were distinct; the qualitative check involved prompting 2∼3 open-ended descriptions, which were coded for indicative themes. Such a method could be directly translated to manipulations of voice social identity in group settings. Video analysis (Franco *et al.*, 2016) can also accompany quantitative manipulation checks as a means of double-checking self-reports through a more objective channel that can also include behavioural responses. Hauser *et al.* (2018)—while critical of manipulation checks—do find value in unobtrusive methods, i.e. non-priming verbal or behavioural qualitative manipulation checks. For example, recordings, transcripts or written accounts could be checked for pronoun use as a vector of sense of independence in relation to a voice agent ('we' vs. 'I'), or emotionally-laden words as indicative of affective state, if meant to be manipulated in the voice UX context. All of these have the potential to be taken up in future work.

Quantitative methods and significance testing are important, but should not form the entire basis of validating a manipulation (Ejelöv & Luke, 2020). Likert scales, for instance, are widely employed because of their simplicity and ease of analysis, but can vary in reliability and validity, limiting their ability to capture nuanced emotions and perspectives. Qualitative methods are well posed to assess manipulation checks more holistically, perhaps especially for voice UX (Seaborn *et al.*, 2024). Together with debriefing interviews, drawings can allow participants to richly express their ideas, thoughts and emotions. As such, we consider drawing a promising alternative or complement for manipulation checks for 'bodiless' or 'disembodied' computer voices and VUIs.

For manipulation checks, the primary benefits may be twofold: unveiling unexpected, latent factors and allowing for a more exploratory way of analysing and validating assumptions than typical top-down, quantitative manipulation checks. Drawings also offer more indirect and unobtrusive assessments of assumptions and manipulations, without revealing goals, expectations or hypotheses, a concern for manipulation checks (Fiedler *et al.*, 2021, Hauser *et al.*, 2018). Thus, using drawing methods for voice manipulation checks may help prevent participants from wondering about the underlying goals of the study, reducing the risk of influencing behaviour or responses (Hauser *et al.*, 2018). In addition, given the task demand involved, drawing may decrease the risk of attention loss, another major concern of unintended effects on the manipulation (Ejelöv & Luke, 2020).

In the past, drawing has been utilized as a qualitative method outside of the domain of manipulation checks. We can draw inspiration from these domains for what manipulations may be evaluated. For instance, Guillemin (2004) expanded their operationalization of 'illness' based on patient drawings. Kearney & Hyle (2004) uncovered latent emotions and cognitions that could be then measured quantitatively. Fleury (2011) found that the method elicited unexpected disclosures that could inform the design of mobile phones. For manipulation checks, the inherently stimulating activity of drawing paired with a potentially more comfortable atmosphere (Fleury, 2011) may encourage reflection in a more free-form, non-prescriptive and holistic manner (Guillemin, 2004), which may reveal latent thoughts and uncover hidden aspects of sense-making (Kearney & Hyle, 2004). Indeed, we found that was so in Case Study 3.

Yet, we must acknowledge challenges in drawing study methods. Drawing studies can be more resource-intensive and

[10] https://www.morgan-klaus.com/gender-guidelines.html

costly than quantitative counterparts, with fewer participants and longer data analyses times, as well. Still, we can imagine *crowdsourcing* qualitative data analysis by providing images to mass numbers of raters. This would follow in the footsteps of initiatives like Foldit for protein structures (Cooper *et al.*, 2010). Drawing studies may be conducted online to gather more diverse and larger samples, but, to the best of our knowledge, this has not been explored (although a manuscript is under review).

In general, qualitative inquiries can be conducted in a variety of paradigms (Creswell & Miller, 2000). In this context, manipulation checks are well suited to support post-positivist and critical paradigms in validity procedures (e.g. as a form of triangulation, member checking or collaboration (Creswell & Miller, 2000)). They are likely less suitable in a constructivist paradigm, though some forms of qualitative 'manipulation checks' could still help to clarify whether participants understood and/or interpreted information provided by researchers as intended, to explore whether there is a shared conceptual understanding. Further, to credibly show prolonged engagement in the field, indications of qualitative data characteristics (e.g. duration of interviews, duration of ethnographic presence at the research site) could also be considered a form of 'manipulation check.'

7.4 From the 'Present' to the Future: Gifting Manipulation Checks to Voice UX and Beyond

We now relate our findings within recent VUI literature and illustrate their applicability to audio interfaces more generally. Zargham *et al.* (2021), for instance, investigated the UX of single-agent versus multi-agent voice assistants, using different voices to associate different agents with different tasks. While the multi-agent approach was rated higher on UX factors, there was no manipulation check involved, meaning that we cannot determine if the effects were caused by the number of different agents *or* agent voice characteristics, especially without knowing the perceptions, attitudes and assumptions of participants about the voices. Asking participants to draw how they imagined the different agents might have reduced this ambiguity in the results and uncovered hidden biases.

Similarly, existing commercially available voices were used in a study by Lee *et al.* (2019b) to investigate the effect of VA genderedness and personality style on technology adoption and acceptance. While style (informative and sociable) was validated in a manipulation check, the gendering of the respective voices was not. Considering our findings from the second case study, the findings concerning gender could have been due to the vocal characteristics of the voices used rather than the gender participants might or might not have assigned to the agent.

Audio manipulation checks beyond VUIs and voice agents are also needed. For instance, Altmeyer *et al.* (2022) investigated the impact of sound effects accompanying points in a gamified image classification task. They created various sound effects to evoke certain emotions and tested them in a pre-study to make sure that they actually led to the intended affective experiences. Still, the results were not retained in the main study. A potential reason was the task: Similar to Case Study 1, participants were only presented with the sound effects and asked to rate them in the pre-study, while in the main study, participants performed the image classification task with the sound effects as an accompaniment. This points to the risk of evaluating any sound stimuli *in isolation*, the complexity of stimuli within the *actual context of use* and the benefit of *thoroughfare* manipulation checks—i.e. employing them at multiple points across studies—to (dis)confirm stimuli validation throughout a research project.

A meta-level question going forward will be the degree of investment in sample sizes for manipulation checks. HCI has a long history of balancing sample size by methodological tradition and pragmatics (Caine, 2016). For example, we might find saturation at twenty interviews or drawings (Francis *et al.*, 2010), yet also saturation is not a goal for all types of qualitative research (Braun & Clarke, 2021). Our position is that manipulation checks should aim for the same confidence metrics as main studies. Otherwise, we cannot have confidence in the manipulation itself.

8 Conclusion

Voice is the heart of voice UX. Yet, choice of voice is often taken for granted or under-reported. Our research community has access to a range of free or relatively inexpensive commercial products, TTSs and data sets filled with voice clips. We may rely on the labels already ascribed to these materials or perhaps make an expert call ourselves. Here, we have shown the danger in this, as well as the merits of taking a step back and assessing voice material with manipulation checks. Voice is a pluralistic, perceptual phenomenon that deserves rigorous treatment bereft of expert assumptions, from the start. We end with our key takeaways as a general set of good practices when conducting voice UX manipulation checks:

- **Identify and include all voice characteristics that are theoretically or practically relevant, despite commonsense assumptions and prescribed characteristics.** Make informed choices when selecting voice, i.e. do not base these choices on expert assumptions or commercial prescriptions. Validate expectations for the selected voice (i) in terms of social and other key characteristics and (ii) their relation to the manipulation.
- **Identify and include all measures and levels of measurement that may relate to the manipulation.** Be expansive rather than prescriptive, wherever possible. Make sure to aim for holistic/complete operationalizations of quantitative measures for manipulation checks. Avoid incomplete manipulations lacking important response alternatives (e.g. gender).
- **Identify and include all user groups who may perceive the voice differently**. Be aware than individual perceptions may override general trends. In selecting user groups who may relate differently to the voice, consider social theories that may be less known or unknown in HCI (e.g. SIT).
- **Consider complementary manipulation checks that could be quantitative, qualitative or mixed in nature.** Be aware of characteristics that cannot be easily measured quantitatively. These latent variables may be assessed (additionally or instead) through qualitative methodologies, such as allowing participants to describe their assumptions regarding voice in an unconstrained manner or by observing behavioural cues to detect emotional responses. As demonstrated, a drawing study can work well, but alternatives (such as design fiction, e.g. Dunne & Raby (2013), Troiano *et al.* (2020), storytelling through crafting, e.g. Shaw *et al.* (2023) or embodied approaches like movement analysis, e.g. Newlove & Dalby (2019) in behavioural responses) may also be worthwhile.
- **Consider multiple checks to ensure stability of perceptions over time and across samples**. These may be conducted as pre-studies, in pilot tests, in main studies or as follow-ups.
- **Report the manipulation check in the main paper.** Placement and level of detail may depend on the nature of the manipulation check (e.g. stand-alone, incorporated within a

study) and when it was conducted (e.g. as a pre-study, in the methods section linked to the material or experimental manipulation or opening the results section).
- **Report whether and how the manipulation check results affected the design or selection of the voice stimuli or main findings.** This could help others make informed choices about certain voices, especially ones commonly used by the community. Manipulation check data may also be used to develop meta-analyses of voice perceptions, which could inform meta-analyses of their use in experimental manipulations.

In short, voice UX research needs better manipulation checks, and accompanying that, a critical and ongoing discussion of how and when to conduct them. We have showcased this in three case studies, argued for why this is needed, and provided a set of implications for practice, within and beyond the scope of voice UX. A fruitful next step could be a systematic review of the literature to draw out more examples and possibilities tied to specific voice stimuli, variables, contexts and user groups. Join us on centring the voice of the machine before jumping headfirst into exploring the intended interaction with it.

Acknowledgments

This work was funded by the Japan Society for the Promotion of Science (JSPS) through a Grants-in-Aid for Early Career Scientists (KAKENHI WAKATE #21K18005). We thank the Fonobono Research Institute for cooperation in producing the older adult voices. We thank Julia Keckeis for assistance with data analysis. We thank Suzuka Yoshida and the members of the Aspire Lab for research support and pilot testing.

Data Availability

The data underlying this article are available in Google Sheets, and can be accessed at https://bit.ly/unboxvoiceux

References

Abbey, J. D. and Meloy, M. G. (2017) Attention by design: using attention checks to detect inattentive respondents and improve data quality. *J. Oper. Manag.*, **53–56**, 63–70. https://doi.org/10.1016/j.jom.2017.06.001.

Allen, M. L. and Butler, H. (2020) Can drawings facilitate symbolic understanding of figurative language in children? *Br. J. Dev. Psychol.*, **38**, 345–362. https://doi.org/10.1111/bjdp.12330.

Altmeyer, M., Hnatovskiy, V., Rogers, K., Lessel, P. and Nacke, L. E. (2022) Here comes no boom! The lack of sound feedback effects on performance and user experience in a gamified image classification task. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. ACM, New York, NY, USA.

Avey, J. B., Avolio, B. J. and Luthans, F. (2011) Experimentally analyzing the impact of leader positivity on follower positivity and performance. *Leadersh. Q.*, **22**, 282–294. https://doi.org/10.1016/j.leaqua.2011.02.004.

Baird, A., Jørgensen, S. H., Parada-Cabaleiro, E., Hantke, S., Cummins, N. and Schuller, B. (2017) Perception of paralinguistic traits in synthesized voices. In *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences, AM '17*. USA. Association for Computing Machinery, New York, NY.

Baird, A., Jørgensen, S., Parada-Cabaleiro, E., Cummings, N., Hantke, S. and Schuller, B. (2018) The perception of vocal traits in synthesized voices: age, gender, and human likeness. *J. Audio Eng. Soc.*, **66**, 277–285. https://doi.org/10.17743/jaes.2018.0023.

Bajaj, N. and Reed, S. M. (2022) Thematic analysis comparing stressors for pediatric residents and subspecialty fellows at a large children's hospital. *Ann. Med.*, **54**, 3332–3340. https://doi.org/10.1080/07853890.2022.2148731.

Bartneck, C., Kulić, D., Croft, E. and Zoghbi, S. (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.*, **1**, 71–81. https://doi.org/10.1007/s12369-008-0001-3.

Behrens, S. I., Egsvang, A. K. K., Hansen, M. and Møllegård-Schroll, A. M. (2018) Gendered robot voices and their influence on trust. In *Companion of the 2018 ACM/IEEE International Conference on Human–Robot Interaction*, pp. 63–64. ACM/IEEE, New York, NY, USA.

Bellini, R., Strohmayer, A., Alabdulqader, E., Ahmed, A. A., Spiel, K., Bardzell, S. and Balaam, M. (2018) Feminist HCI: taking stock, moving forward, and engaging community. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–4. ACM, New York, NY, USA.

Bergen, H. (2016) 'I'd blush if I could': digital assistants, disembodied cyborgs and the problem of gender. *Word Text J. Lit. Stud. Linguist.*, **6**, 95–113.

Blankenship, A. (1942) Psychological difficulties in measuring consumer preference. *J. Market.*, **6**, 66–75.

Braun, V. and Clarke, V. (2021) To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qual. Res. Sport Exerc. Health*, **13**, 201–216. https://doi.org/10.1080/2159676X.2019.1704846.

Buolamwini, J. and Gebru, T. (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency (FAccT)*, pp. 77–91. Proceedings of Machine Learning Research (PMLR), Cambridge, MA, USA.

Byrne, D. (1969) Attitudes and attraction. In *Advances in Experimental Social Psychology*, vol. **4**, pp. 35–89. Elsevier, Amsterdam, The Netherlands.

Caine, K. (2016) Local standards for sample size at chi. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI'16*. ACM.

Cambre, J. and Kulkarni, C. (2019) One voice fits all? Social implications and research challenges of designing voices for smart devices. *Proc. ACM Hum.-Comput. Interact.*, **3(CSCW)**, 1–19.

Carducci, B. J., Nave, C. S., Di Fabio, A., Mio, J. S., Riggio, R. E., Saklofske, D. H. and Stough, C. (2020) *The Wiley Encyclopedia of Personality and Individual Differences, Set*. John Wiley & Sons, Toronto, ON, Canada.

Carroll, J. M. (1997) Human–computer interaction: psychology as a science of design. *Ann. Rev. Psychol.*, **48**, 61–83. https://doi.org/10.1146/annurev.psych.48.1.61.

Cary, L. A., Chasteen, A. L. and Remedios, J. (2017) The ambivalent ageism scale: developing and validating a scale to measure benevolent and hostile ageism. *Gerontologist*, **57**, e27–e36. https://doi.org/10.1093/geront/gnw118.

Chan, K. (2006) Exploring children's perceptions of material possessions: a drawing study. *Qual. Mark. Res. Int. J.*, **9**, 352–366. https://doi.org/10.1108/13522750610689087.

Chang, R. C.-S., Lu, H.-P. and Yang, P. (2018) Stereotypes or golden rules? Exploring likable voice traits of social robots as active aging companions for tech-savvy baby boomers in taiwan. *Comput. Hum. Behav.*, **84**, 194–210. https://doi.org/10.1016/j.chb.2018.02.025.

Chonody, J. M. (2016) Positive and negative ageism: the role of benevolent and hostile sexism. *Affilia*, **31**, 207–218. https://doi.org/10.1177/0886109915595839.

Clark, L.*et al.* (2019) The state of speech in hci: trends, themes and challenges. *Interact. Comput.*, **31**, 349–371. https://doi.org/10.1093/iwc/iwz016.

Clarke, V. and Braun, V. (2021) *Thematic Analysis: A Practical Guide*. Sage Publications, Thousand Oaks, CA, USA.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver–Fay, A., Baker, D., Popović, Z. and Players, F. (2010) Predicting protein structures with a multiplayer online game. *Nature*, **466**, 756–760. https://doi.org/10.1038/nature09304.

Crenshaw, K. W. (2017) *On Intersectionality: Essential Writings*. The New Press, New York, NY, USA.

Creswell, J. W. and Miller, D. L. (2000) Determining validity in qualitative inquiry. *Theor. Pract.*, **39**, 124–130. https://doi.org/10.1207/s15430421tip3903_2.

Desai, S., Dubiel, M. and Leiva, L. A. (2024) Examining humanness as a metaphor to design voice user interfaces. In *In Proceedings of the 4th Conference on Conversational User Interfaces (CUI), page accepted*. USA. Association for Computing Machinery, New York, NY.

Deterding, N. M. and Waters, M. C. (2018) Flexible coding of in-depth interviews: a twenty-first-century approach. *Sociol. Methods Res.*, **50**, 708–739.

Dillon, A. and Watson, C. (1996) User analysis in hci—the historical lessons from individual differences research. *Int. J. Hum.-Comput. Stud.*, **45**, 619–637. https://doi.org/10.1006/ijhc.1996.0071.

Donoghue, S. (2010) Projective techniques in consumer research. *J. Fam. Ecol. Consum. Sci.*, **28**. https://doi.org/10.4314/jfecs.v28i1.52784.

Druga, S., Williams, R., Breazeal, C. and Resnick, M. (2017) "Hey google is it ok if i eat you?": initial explorations in child–agent interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children, IDC '17*, pp. 595–600. Association for Computing Machinery, New York, NY, USA.

Dubiel, M., Sergeeva, A. and Leiva, L. A. (2024) Impact of voice fidelity on decision making: a potential dark pattern? In *Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI '24*, pp. 181–194. Association for Computing Machinery, New York, NY, USA.

Dunne, A. and Raby, F. (2013) Speculative everything: design, fiction, and social dreaming. In *BusinessPro Collection*. MIT Press, Cambridge, MA, USA.

Egan, D. E. (1988) Individual differences in human–computer interaction. In *Handbook of Human–Computer Interaction*, pp. 543–568. Elsevier, Amsterdam, The Netherlands.

Ejelöv, E. and Luke, T. J. (2020) " Rarely safe to assume": evaluating the use and interpretation of manipulation checks in experimental social psychology. *J. Exp. Soc. Psychol.*, **87**, 103937. https://doi.org/10.1016/j.jesp.2019.103937.

Fayant, M.-P., Sigall, H., Lemonnier, A., Retsin, E. and Alexopoulos, T. (2017) On the limitations of manipulation checks: an obstacle toward cumulative science. *Int. Rev. Soc. Psychol.*, **30**, 125. https://doi.org/10.5334/irsp.102.

Fiedler, K., McCaughey, L. and Prager, J. (2021) Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspect. Psychol. Sci.*, **16**, 816–826. https://doi.org/10.1177/1745691620970602.

Fleury, A. (2011) Drawing as a user experience research tool. In *Proceedings of the 11th Danish Human–Computer Interaction Research Symposium (DHRS2011)*, 14–17.

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P. and Grimshaw, J. M. (2010) What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychol. Health*, **25**, 1229–1245. https://doi.org/10.1080/08870440903194015.

Franco, M. G., Katz, R. and O'Brien, K. M. (2016) Forbidden identities: a qualitative examination of racial identity invalidation for black/white biracial individuals. *Int. J. Intercult. Relat.*, **50**, 96–109. https://doi.org/10.1016/j.ijintrel.2015.12.004.

Furnham, A. (1986) Response bias, social desirability and dissimulation. *Pers. Individ. Dif.*, **7**, 385–400. https://doi.org/10.1016/0191-8869(86)90014-0.

Gruijters, S. L. K. (2022) Making inferential leaps: manipulation checks and the road towards strong inference. *J. Exp. Soc. Psychol.*, **98**, 104251. https://doi.org/10.1016/j.jesp.2021.104251.

Guest, G., MacQueen, K. M. and Namey, E. E. (2011) *Applied Thematic Analysis*. Sage Publications, Oliver's Yard, London, UK.

Guillemin, M. (2004) Understanding illness: using drawings as a research method. *Qual. Health Res.*, **14**, 272–289. https://doi.org/10.1177/1049732303260445.

Guillemin, M. and Drew, S. (2010) Questions of process in participant-generated visual methodologies. *Vis. Stud.*, **25**, 175–188. https://doi.org/10.1080/1472586X.2010.502676.

Gurtman, M. B. (2009) Exploring personality with the interpersonal circumplex. *Soc. Personal Psychol. Compass*, **3**, 601–619. https://doi.org/10.1111/j.1751-9004.2009.00172.x.

Guzman, A. L. (2019) Voices in and of the machine: source orientation toward mobile virtual assistants. *Comput. Hum. Behav.*, **90**, 343–350. https://doi.org/10.1016/j.chb.2018.08.009.

Hauser, D. J., Ellsworth, P. C. and Gonzalez, R. (2018) Are manipulation checks necessary? *Front. Psychol.*, **9**, 998. https://doi.org/10.3389/fpsyg.2018.00998.

Head, T. C., Griffin, R. W., Bateman, T. S., Lohman, L. and Yates, V. L. (1988) The priming effect in task design research. *J. Manag.*, **14**, 33–39. https://doi.org/10.1177/014920638801400104.

Henrich, J., Heine, S. J. and Norenzayan, A. (2010) The weirdest people in the world? *Behav. Brain Sci.*, **33**, 61–83. https://doi.org/10.1017/S0140525X0999152X.

Hochard, K. D., Ashcroft, S., Carroll, J., Heym, N. and Townsend, E. (2017) Exploring thematic nightmare content and associated self-harm risk. *Suicide Life Threat. Behav.*, **49**, 64–75. https://doi.org/10.1111/sltb.12402.

Hogg, M. A., Terry, D. J. and White, K. M. (1995) A tale of two theories: a critical comparison of identity theory with social identity theory. *Soc. Psychol. Q.*, **58**, 255–269. https://doi.org/10.2307/2787127.

Hwang, G., Lee, J., Oh, C. Y. and Lee, J. (2019) It sounds like a woman: exploring gender stereotypes in South Korean voice assistants. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA '19*, pp. 1–6. Association for Computing Machinery, New York, NY, USA.

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C. and van Anders, S. M. (2019) The future of sex and gender in psychology: five challenges to the gender binary. *Am. Psychol.*, **74**, 171–193. https://doi.org/10.1037/amp0000307.

Inuhiko, Y. (2006) *"Kawaii" ron [The Theory of Kawaii]*. In *Number 578 in Chikuma Shinsho. Chikuma Shobō*, Tokyo, Japan.

Iseli, M., Shue, Y.-L. and Alwan, A. (2006) Age-and gender-dependent analysis of voice source characteristics. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. **1**, pp. I–I. IEEE, New York, NY, USA.

Jestin, I., Fischer, J., Galvez Trigo, M. J., Large, D. and Burnett, G. (2022) Effects of wording and gendered voices on acceptability of voice assistants in future autonomous vehicles. In *Proceedings of the 4th Conference on Conversational User Interfaces, CUI '22*. USA. Association for Computing Machinery,New York, NY.

Jolley, R. P. and Thomas, G. V. (1995) Children's sensitivity to metaphorical expression of mood in line drawings. *Br. J. Dev. Psychol.*, **13**, 335–346. https://doi.org/10.1111/j.2044-835X.1995.tb00684.x.

Joshi, A.*et al.* (2014) Supporting treatment of people living with hiv / aids in resource limited settings with ivrs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14*, pp. 1595–1604. USA. Association for Computing Machinery, New York, NY.

Jung, H., Kim, H. J., So, S., Kim, J. and Oh, C. (2019) Turtletalk: an educational programming game for children with voice user interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–6. ACM, New York, NY, USA.

Kao, D., Ratan, R., Mousas, C. and Magana, A. J. (2021) The effects of a self-similar avatar voice in educational games. *Proc. ACM Hum.-Comput. Interact.*, **5**, 1–28. https://doi.org/10.1145/3474665.

Kearney, K. S. and Hyle, A. E. (2004) Drawing out emotions: The use of participant-produced drawings in qualitative inquiry. *Qual. Res.*, **4**, 361–382. https://doi.org/10.1177/1468794104047234.

Knowles, E. S. and Nathan, K. T. (1997) Acquiescent responding in self-reports: cognitive style or social concern? *J. Res. Pers.*, **31**, 293–301. https://doi.org/10.1006/jrpe.1997.2180.

Kornadt, A. E. and Rothermund, K. (2012) Internalization of age stereotypes into the self-concept via future self-views: a general model and domain-specific differences. *Psychol. Aging*, **27**, 164–172. https://doi.org/10.1037/a0025110.

Lee, E. J., Nass, C. and Brave, S. (2000) Can computer-generated speech have gender? An experimental test of gender stereotype. In *CHI'00 Extended Abstracts on Human Factors in Computing Systems*, pp. 289–290. ACM, New York, NY, USA.

Lee, S., Kim, S. and Lee, S. (2019a) "What does your agent look like?": A drawing study to understand users' perceived persona of conversational agent. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA '19*, pp. 1–6. Association for Computing Machinery, New York, NY, USA.

Lee, S., Ratan, R. and Park, T. (2019b) The voice makes the car: enhancing autonomous vehicle perceptions and adoption intention through voice agent gender and style. *Multimodal Technol. Interact.*, **3**, 20. https://doi.org/10.3390/mti3010020.

Lieber-Milo, S. (2021) Cute at an older age: a case study of Otona-Kawaii. *Mutual Image. J.*, **10**, 93–108.

Lopatovska, I. and Williams, H. (2018) Personification of the amazon alexa: Bff or a mindless companion. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, CHIIR '18*, pp. 265–268. Association for Computing Machinery, New York, NY, USA.

Machover, K. (1949) Personality projection in the drawing of the human figure (A method of personality investigation). In *Personality Projection in the Drawing of the Human Figure: A Method of Personality Investigation, American Lectures in Psychology*, pp. 3–32. Charles C. Thomas Publisher, Springfield, IL, US.

Malchiodi, C. (2006) *Art Therapy Sourcebook*. McGraw Hill Professional, New York, NY, USA.

McCrae, R. R., Kurtz, J. E., Yamagata, S. and Terracciano, A. (2011) Internal consistency, retest reliability, and their implications for personality scale validity. *Pers. Soc. Psychol. Rev.*, **15**, 28–50. https://doi.org/10.1177/1088868310366253.

McEnaney, T. (2019) This american voice: the odd timbre of a new standard in public radio. In Eidsheim, N., Meizel, K. (eds), *The Oxford Handbook of Voice Studies, chapter 6*, pp. 97–123. Oxford University Press, Oxford, United Kingdom.

McNeal, J. U. and Ji, M. F. (2003) Children's visual memory of packaging. *J. Consum. Mark.*, **20**, 400–427. https://doi.org/10.1108/07363760310489652.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021) A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, **54**, 1–35.

Mendelson, J. and Aylett, M. P. (2017) Beyond the listening test: an interactive approach to tts evaluation. In *Interspeech 2017*, pp. 249–253. ISCA.

Morris, M. R., Wobbrock, J. O. and Wilson, A. D. (2010) Understanding users' preferences for surface gestures. In *Proceedings of Graphics Interface 2010, GI '10*, pp. 261–268. Canadian Information Processing Society, CAN.

Mutz, D. C. and Pemantle, R. (2015) Standards for experimental research: encouraging a better understanding of experimental methods. *J. Exp. Political Sci.*, **2**, 192–215. https://doi.org/10.1017/XPS.2015.4.

Nass, C. and Lee, K. M. (2000) Does computer-generated speech manifest personality? An experimental test of similarity-attraction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '00*, pp. 329–336. USA. Association for Computing Machinery, New York, NY.

Nass, C., Moon, Y. and Green, N. (1997) Are machines gender neutral? Gender-stereotypic responses to computers with voices. *J. Appl. Soc. Psychol.*, **27**, 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x.

Naumburg, M. (1966) *Dynamically Oriented Art Therapy: Its Principles and Practices, Illustrated With Three Case Studies*. Grune & Stratton, New York.

Newlove, J. and Dalby, J. (2019) *Laban for All*. Routledge, Milton Park, Abington, UK.

Nittono, H. (2010) A behavioral science framework for understanding kawaii. In *Proceedings of The Third International Workshop on Kansei*, pp. 80–83. Editorial Committee of the Third International Workshop on Kansei, Fukuoka, Japan.

Nittono, H. (2016) The two-layer model of 'kawaii': a behavioural science framework for understanding kawaii and cuteness. *East Asia. J. Pop. Culture*, **2**, 79–95. https://doi.org/10.1386/eapc.2.1.79_1.

Nittono, H., Fukushima, M., Yano, A. and Moriya, H. (2012) The power of kawaii: viewing cute images promotes a careful behavior and narrows attentional focus. *PLOS ONE*, **7**, e46362. https://doi.org/10.1371/journal.pone.0046362.

Nittono, H. and Ihara, N. (2017) Psychophysiological responses to kawaii pictures with or without baby schema. *SAGE Open*, **7**, 2158244017709321. https://doi.org/10.1177/2158244017709321.

Overend, A. (2022) Situatedness. In *Showing Theory to Know Theory*.

Palmore, E. (1999) *Ageism: Negative and Positive*. Springer Publishing Company, New York, NY, USA.

Patton, M. Q. (1990) *Qualitative Evaluation and Research Methods*. Sage Publications, Oliver's Yard, London, UK.

Piaget, J. and Inhelder, B. (1969) *The Psychology of The Child*. Routledge and Keegan Paul, London, England.

Poeller, S., Dechant, M. J., Klarkowski, M. and Mandryk, R. L. (2023) Suspecting sarcasm: how league of legends players dismiss positive communication in toxic environments. *Proc. ACM Hum.-Comput. Interact.*, **7**, 1–26. https://doi.org/10.1145/3611020.

Poyatos, F. (1993) *Paralanguage: A Linguistic and Interdisciplinary Approach to Interactive Speech and Sounds*, vol. **92**. John Benjamins Publishing, Amsterdam, The Netherlands.

Pradhan, A., Findlater, L. and Lazar, A. (2019) "Phantom friend" or "just a box with information": personification and ontological

categorization of smart speaker-based voice assistants by older adults. *Proc. ACM Hum.-Comput. Interact.*, **3**, 1–21. https://doi.org/10.1145/3359316.

Rode, J. A. (2011) Reflexivity in digital anthropology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11*, pp. 123–132. USA. Association for Computing Machinery, New York, NY.

Rogers, K. and Weber, M. (2019) Audio habits and motivations in video game players. In *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound, AM '19*, pp. 45–52. USA. Association for Computing Machinery, New York, NY.

Rood, L. (2011) The role of rumination in the development of depressive symptoms in youth. PhD thesis, Maastricht University. Chapter 7: Causal-analysis of stressful events, not focus on negative inferences, increases anxious affect compared to focus on contextual facts.

Sawa, Y., Keckeis, J. and Seaborn, K. (2023) Right for the job or opposites attract? Exploring cross-generational user experiences with "younger"' and "older"' voice assistants. In *Designing Interactive Systems Conference, DIS '23*, pp. 160–163. ACM.

Sawa, Y. and Seaborn, K. (2022) Localizing the ambivalent ageism scale for Japan. In *The 8th Asian Conference on Aging & Gerontology 2022: Official Conference Proceedings*, pp. 33–36. IAFOR, Tokyo, Japan.

Schleinschok, K., Eitel, A. and Scheiter, K. (2017) Do drawing tasks improve monitoring and control during learning from text? *Learn. Instruct.*, **51**, 10–25. Bridging Cognitive Load and Self-Regulated Learning Research.

Schuman, H. and Presser, S. (1996a) *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Sage Publications, Oliver's Yard, London, UK.

Schuman, H. and Presser, S. (1996b) *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. SAGE, Thousand Oaks, CA, USA.

Seaborn, K. (2022) From identified to self-identifying: social identity theory for socially embodied artificial agents. In *Robo-Identity: Exploring Artificial Identity and Emotion via Speech Interactions (HRI 2022 Workshop on Robo-Identity 2), co-Located with the 2022 International Conference on Human-Robot Interaction (HR I 2022)*, Sapporo, Japan.

Seaborn, K. and Frank, A. (2022) What pronouns for pepper? a critical review of gender/ing in research. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22*. Association for Computing Machinery, New York, NY, USA.

Seaborn, K., Miyake, N. P., Pennefather, P. and Otake-Matsuura, M. (2021) Voice in human–agent interaction: a survey. *ACM Comput. Surv.*, **54**.

Seaborn, K., Nam, S., Keckeis, J. and Itagaki, T. (2023) Can voice assistants sound cute? Towards a model of kawaii vocalics. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, CHI EA '23*, pp. 1–7. Association for Computing Machinery, New York, NY, USA.

Seaborn, K., Pennefather, P. and Kotani, H. (2022) Exploring gender-expansive categorization options for robots. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, CHI EA '22*. Association for Computing Machinery, New York, NY, USA.

Seaborn, K. and Urakami, J. (2021) Measuring voice ux quantitatively: a rapid review. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, CHI EA '21*. Association for Computing Machinery, New York, NY, USA.

Seaborn, K., Urakami, J., Pennefather, P. and Miyake, N. P. (2024) Qualitative approaches to voice ux. *ACM Comput. Surv.*, **56**, 1–34. https://doi.org/10.1145/3658666.

Seymour, W. and Van Kleek, M. (2021) Exploring interactions between trust, anthropomorphism, and relationship development in voice assistants. *Proc. ACM Hum.-Comput. Interact.*, **5**, 1–16. https://doi.org/10.1145/3479515.

Shaw, M. S., Coleman, J. J., Thomas, E. E. and Kafai, Y. B. (2023) Restorying a black girl's future: using womanist storytelling methodologies to reimagine dominant narratives in computing education. *J. Learn. Sci.*, **32**, 52–75. https://doi.org/10.1080/10508406.2023.2179847.

Shiokawa, K. (1999) Cute but deadly: women and violence in Japanese comics. In Lent, J. A. (ed.), *Themes and Issues in Asian Cartooning: Cute, Cheap, Mad, and Sexy*, pp. 93–126. Bowling Green State University Popular Press, Bowling Green, OH, USA.

Sigall, H. and Mills, J. (1998) Measures of independent variables and mediators are useful in social psychology experiments: but are they necessary? *Pers. Soc. Psychol. Rev.*, **2**, 218–226. https://doi.org/10.1207/s15327957pspr0203_5.

Spiel, K., Haimson, O. L. and Lottridge, D. (2019) How to do better with gender on surveys: a guide for hci researchers. *Interactions*, **26**, 62–65. https://doi.org/10.1145/3338283.

Stigall, B., Waycott, J., Baker, S. and Caine, K. (2019) Older adults' perception and use of voice user interfaces: a preliminary review of the computing literature. In *Proceedings of the 31st Australian Conference on Human–Computer-Interaction*, pp. 423–427.

Strait, M. K., Aguillon, C., Contreras, V. and Garcia, N. (2017) The public's perception of humanlike robots: online social commentary reflects an appearance-based uncanny valley, a general fear of a "technology takeover", and the unabashed sexualization of female-gendered robots. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE.

Straub, D. and Gefen, D. (2004) Validation guidelines for IS positivist research. *Commun. Assoc. Inf. Syst.*, **13**. https://doi.org/10.17705/1CAIS.01324.

Sundar, S. S. and Nass, C. (2000) Source orientation in human-computer interaction: programmer, networker, or independent social actor. *Commun. Res.*, **27**, 683–703. https://doi.org/10.1177/009365000027006001.

Sutton, S. J., Foulkes, P., Kirk, D. and Lawson, S. (2019) Voice as a design material: sociophonetic inspired design strategies in human–computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, pp. 1–14. Association for Computing Machinery, New York, NY, USA.

Tajfel, H. and Turner, J. C. (2004) The social identity theory of intergroup behavior. In *Political Psychology*, pp. 276–293. Psychology Press, Hove, East Sussex, UK.

Tajfel, H., Turner, J. C., Austin, W. G. and Worchel, S. (1979) An integrative theory of intergroup conflict. *Organ. Identity Reader*, **56**, 9780203505984–9780203505916.

Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J. and Schiebinger, L. (2019) Sex and gender analysis improves science and engineering. *Nature*, **575**, 137–146. https://doi.org/10.1038/s41586-019-1657-6.

Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J. S., Leimeister, J. M. M. and Bernstein, A. (2021) Female by default? Exploring the effect of voice assistant gender and pitch on trait and trust attribution. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–7. ACM, New York, NY, USA.

Torre, I., Lagerstedt, E., Dennler, N., Seaborn, K., Leite, I. and Székely, E. (2023) Can a gender-ambiguous voice reduce gender stereotypes

in human-robot interactions? In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE.

Torre, I., Latupeirissa, A. B. and McGinn, C. (2020) How context shapes the appropriateness of a robot's voice. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, vol. **33**, pp. 215–222. IEEE.

Troiano, G. M., Wood, M. and Harteveld, C. (2020) "And this, kids, is how I met your mother"': consumerist, mundane, and uncanny futures with sex robots. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, pp. 1–17. Association for Computing Machinery, New York, NY, USA.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D. and Wetherell, M. S. (1987) *Rediscovering the Social Group: A Self-Categorization Theory*. Basil Blackwell, Hoboken, NJ, USA.

Urakami, J., Qie, N., Kang, X. and Patrick Rau, P.-L. (2021) Cultural adaptation of "kawaii" in short mobile video applications: how the perception of "kawaii"' is shaped by the cultural background of the viewer and the gender of the performer. *Comput. Hum. Behav. Rep.*, **4**, 100109. https://doi.org/10.1016/j.chbr.2021.100109.

van Berkel, N. and Hornbæk, K. (2023) Implications of human–computer interaction research. *Interactions*, **30**, 50–55. https://doi.org/10.1145/3600103.

Waltz, J., Addis, M. E., Koerner, K. and Jacobson, N. S. (1993) Testing the integrity of a psychotherapy protocol: assessment of adherence and competence. *J. Consult. Clin. Psychol.*, **61**, 620–630. https://doi.org/10.1037/0022-006X.61.4.620.

Westen, D. (1988) Transference and information processing. *Clin. Psychol. Rev.*, **8**, 161–179. https://doi.org/10.1016/0272-7358(88)90057-8.

Wobbrock, J. O. and Kientz, J. A. (2016) Research contributions in human–computer interaction. *Interactions*, **23**, 38–44. https://doi.org/10.1145/2907069.

Yurrita, M., Draws, T., Balayn, A., Murray-Rust, D., Tintarev, N. and Bozzon, A. (2023) Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*. Association for Computing Machinery, New York, NY, USA.

Zajonc, R. B. (1968) Attitudinal effects of mere exposure. *J. Pers. Soc. Psychol.*, **9**, 1–27.

Zargham, N., Bonfert, M., Porzel, R., Doring, T. and Malaka, R. (2021) Multi-agent voice assistants: an investigation of user experience. In *Proceedings of the 20th International Conference on Mobile and Ubiquitous Multimedia*, pp. 98–107. USA. ACM, New York, NY.

Interacting with Computers, 2026, 127–148
https://doi.org/10.1093/iwc/iwae062
Article