The Effects of a Self-Similar Avatar Voice in Educational Games

DOMINIC KAO, Purdue University, USA
RABINDRA RATAN, Michigan State University, USA
CHRISTOS MOUSAS, Purdue University, USA
ALEJANDRA J. MAGANA, Purdue University, USA

Avatar identification is one of the most promising research areas in games user research. Greater identification with one's avatar has been associated with improved outcomes in the domains of health, entertainment, and education. However, existing studies have focused almost exclusively on the *visual appearance* of avatars. Yet audio is known to influence immersion/presence, performance, and physiological responses. We perform one of the first studies to date on avatar *self-similar audio*. We conducted a 2 x 3 (similar/dissimilar x modulation upwards/downwards/none) study in a Java programming game. We find that voice similarity leads to a significant increase in performance, time spent, similarity identification, competence, relatedness, and immersion. Similarity identification acts as a significant mediator variable between voice similarity and all measured outcomes. Our study demonstrates the importance of avatar audio and has implications for avatar design more generally across digital applications.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Games; Avatar; Audio; Voice; Identification; Player Experience

ACM Reference Format:
Dominic Kao, Rabindra Ratan, Christos Mousas, and Alejandra J. Magana. 2021. The Effects of a Self-Similar Avatar Voice in Educational Games. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 238 (September 2021), 28 pages. https://doi.org/10.1145/3474665

1 INTRODUCTION

Virtual identities exist everywhere. From social network profiles, to video games, to virtual reality, there is almost always a representation of the *self*. Because these virtual representations serve as extensions of ourselves, we can identify with them, meaning we temporarily merge their identities with our own self-perception [36]. This identification can be so strong that studies have shown that we conform to the virtual representation's expected behaviors [158, 200]. This influences outcomes including negotiation aggressiveness [200, 201], food choices [66, 169], physical exercise [116, 150, 151], racial bias [149], math performance [113, 159], and creative thinking [26, 45, 76]. Greater identification with a virtual representation—which are often referred to as avatars—is associated with increased motivation [16, 17, 182], performance [100], enjoyment [18, 115, 139, 180], flow [175], and trust in others [101]. Yet, despite the extensive literature attesting to avatars' influence, research has focused almost exclusively on *visual* aspects of the avatar rather than *audial* aspects, potentially because the latter tend to be perceived as a non-critical element of avatar

Authors' addresses: Dominic Kao, kaod@purdue.edu, Purdue University, USA; Rabindra Ratan, rar@msu.edu, Michigan State University, USA; Christos Mousas, cmousas@purdue.edu, Purdue University, USA; Alejandra J. Magana, admagana@purdue.edu, Purdue University, USA.

2573-0142/2021/9-ART238. https://doi.org/10.1145/3474665

use, silent avatars are often perceived as more identifiable, and developing a variety in character voices is resource intensive (e.g., hiring multiple voice actors, programming branching dialog trees) [199]. However, technological solutions to these challenges (e.g., high-quality text-to-speech engines, voice cloning software) can now greatly reduce the resources needed for developing avatar voices, signaling a new opportunity for research on avatar voice effects. For example, consider an avatar in an exercise application for running that speaks using voice characteristics with which the user identifies; greater identification with the avatar's voice could translate into increased exercise performance. Or consider a digital self-help application for smoking cessation; greater identification with the avatar's voice could decrease user attrition, increasing the odds of successfully overcoming addiction. Similarly, when using immersive technologies for learning, such as training how to perform surgery in virtual reality [136], greater identification with the avatar's voice could result in increased presence and motivation, increasing training effectiveness.

There is good reason to believe that an avatar's voice could influence outcomes. A meta-analysis of 83 studies in virtual environments found that the presence of audio contributes a small- to medium-sized effect on presence [43]. Furthermore, audio in games has been linked to greater immersion [55, 109, 135], physiological responses [79], performance [90], and emotional realism [13, 54]. Prior studies give us reason to believe that *avatar audio* in particular could influence avatar identification. Functional neuroimaging shows that perceived similarity is critical to simulating another person's internal state [131]. When a study participant watched a game show contestant with high perceived similarity, the participant experienced a significant increase in vicarious reward [132]. Researchers suggest that similar others trigger likeability, familiarity, and kin-motivated responses [59, 132, 146]. This is often referred to as similarity-attraction [27] and is highly relevant in the extensive literature on pedagogical agents and avatars [100]. Therefore, an avatar with a voice that is more similar to the user's own voice could increase engagement. Nevertheless, one can imagine that a self-similar avatar voice might instead *break* immersion because the avatar is speaking when the user is not. Additionally, Wauck et al. have shown that self-similarity in the context of visual appearance did not make a difference to game performance and experience [196]. As such, self-similar avatar audio might also produce negligible differences.

Our project treats voice similarity as a holistic quality of sound, encompassing characteristics of voice that people use to discriminate between speakers, such as tone, stress, intonation, rhythm, and tempo. Together [107], they comprise a voice identity that individuals might associate with other characteristics they identify with, such as masculinity and femininity [179]. Voice similarity is not a solved technical problem, and the most reliable measure of similarity is subjective ratings—e.g., [89], section 3.2. In this paper, our goal is to study how voice similarity (versus voice dissimilarity) influences users in an educational programming game. We chose to study avatar voice in an educational game as opposed to a game primarily for entertainment because we are also interested in whether STEM gender stereotypes influence the effects of avatar voice. For example, stereotyped avatars' identity characteristics have been found to influence performance in STEM learning contexts. Studies suggest that people perform better on STEM-related tasks when they are represented by a male avatar compared to a female avatar [112, 113, 159]. We expected a similar effect might occur due to avatar voice characteristics associated with masculinity and femininity.

We conducted an online study on Amazon's Mechanical Turk (MTurk) in which half of the participants were given an avatar voice that matched their own voice, while the other half of the participants were given an avatar voice that was randomly chosen from a pool for prior participants. Additionally, we varied voice modulation across all participants; this consisted of pitch shifting the voice upwards, downwards, or not at all. Their avatar's voice was then used inside of the Java programming game as they played. Participants could spend as much time and complete as many

puzzles as they liked, reflecting motivation to engage in and learn from the game. Afterward, we collected measures of need satisfaction, intrinsic motivation, and avatar identification.

Our results show that voice similarity increases performance, time spent, similarity identification, competence, relatedness, and immersion. Similarity identification acts as a significant mediating variable between voice similarity and all measured outcomes. However, there was no evidence that voice modulation significantly influenced outcomes. Our study suggests that games can be made more engaging through self-similar avatar voice audio. Moreover, our study provides motivation for applying similar methods to virtual reality (e.g., effect on presence), voice assistants (e.g., Siri [6]), digital learning (e.g., second-language learning through hearing a self-similar voice), avatar customization (e.g., customizing avatar audio to be similar), and the Proteus effect (the phenomenon that avatar users tend to conform behaviorally to the identity characteristics that they associate with their avatars [200]) in the context of *audio*.

2 RELATED WORK

We describe research in three domains of interest: identification with avatars, audio in games, and player experience in games.

2.1 Avatar Identification

2.1.1 Identification. Identification is a temporary change in a user's self-concept by adoption of a media persona's perceived characteristics [36]. Identification is one of the core components of why media experiences are enjoyable [37, 38]. For example, in literary fiction, the reader is said to adopt the protagonist's emotions, experiences, and objectives such that they feel as if they *are* the protagonist [143]. Or in television, the audience member is said to not only feel sympathetic towards a character, but to feel *with* the character [37]. However, one key difference in video games from other genres of media, such as television, is that players have direct control over the behavior and actions of their characters. Through this active participation, video games can override the distance between players and their avatars [36]. Avatar identification can positively influence enjoyment [18, 115, 139, 180], health outcomes [104], and learning interest [8]. Moreover, it can positively influence intrinsic motivation [16, 182], flow [175], motivation to exercise [115, 190], trust in others [101], self-esteem [195], loyalty to a game [178], and appreciation of the game [23]. Avatar identification has also been associated with aggression [105], addiction [174], and depression [14, 126].

2.1.2 Avatar Identification. Avatar identification is typically operationalized as a multi-faceted construct [50, 185]. *Similarity identification* can be understood as the extent to which we feel similar to the avatar. People expect to be able to build more rewarding interpersonal interactions [80], more easily like, and identify with media characters perceived as being similar [80, 206]. Therefore, avatars that are similar facilitate feelings of closeness and stronger vicarious experiences [185]. This phenomenon (sometimes called similarity-attraction [27, 85]) has been studied for decades in education, wherein pedagogical agents that are similar to users (e.g., gender [75] and race [153, 165]) are more influential. Likewise, greater physical similarity with an on-screen avatar has been shown to significantly increase exercise effort [67]. Nevertheless, avatar dissimilarities can be valuable. For example, users are known to sometimes create avatars that represent idealized versions of themselves [14, 51]. This is known to foster *wishful identification*, wherein the avatar represents an improved version of the real-life self (e.g., leaner, more attractive, and fashionable [51]), represents an ideal, and is someone the user would like to be [185]. *Embodied identification* represents the concept of presence in a virtual environment through a "body container" [15, 121]. This concept refers to *being* the avatar, or feeling as if one is inside the avatar with the body of the avatar as

being one's own [185]. Studies have found that perceived embodiment—induced through increased control of an avatar—heightens the outcomes associated with the avatar's identity [202], supporting the notion that embodied identification should be considered in avatar-effects research. In this paper, we measure the influence of similarity, wishful, and embodied identification as mediators between voice similarity and other outcome variables.

2.1.3 Facilitating Avatar Identification. Currently, the prevailing method of increasing avatar identification is through avatar customization. Customization of one's avatar has been shown to positively increase avatar identification in a variety of contexts [16, 106, 181, 204]. Other factors that can increase identification include the presence of narrative [171] and the character's name [42]. However, no study to date has manipulated the avatar's *voice audio*.

2.2 Audio in Games

2.2.1 Audio Types. Audio can significantly influence player's experiences. A meta-analysis found that the existence of audio, compared to its absence, has a significant effect on presence [43]. Researchers have classified audio into speech and dialog, sound effects, and music [117]. Sound effects are further classified into avatar sounds, object sounds, character sounds, and ornamental sounds [117]. All categories of audio appear to have effects. Game music has been found to influence immersion [135, 170, 197], tension/anxiety [31], risk-taking behavior [163], and concentration [94]. Game sound effects, often an important source of feedback [53, 91, 96, 147, 161], affect immersion [73] and performance [32]. Additionally, the effects of audio are often contextually dependent on game genre [95], device type [164], and preferences [170]. Other studies, though, have found that audio has little effect [164]. Our goal in this paper is to study self-similar avatar voices.

2.2.2 Avatars and Audio. Researchers have suggested that avatar-based sounds, such as breathing (proprioceptive) and footsteps (exteroceptive), can facilitate imaginative immersion and help the player identify with their avatar [74]. It has also been suggested that audio creates a sense of self-representation, which can intensify self-awareness, body ownership, and place illusion [142]. A few early studies have explored how the addition of footstep sounds can influence presence [140] and movement behavior [141]. Researchers have also suggested that using one's own voice to interact with a game (e.g., voice commands) can positively affect avatar identification, despite the dissonance produced in speaking to the game [30].

Several studies show that voice affects users. Voiceovers for non-player characters have been shown to increase engagement in a role-playing game [28]. Virtual customer service representatives that include a text-to-speech voice increase flow [156] and trust [157]. However, not only the presence vs. absence of voice, but voice *similarity* also affects users. In a public-speaking experiment in front of a virtual classroom, participants either gave their own speech out loud, or had another participant's speech audio played back. Participants using their own voice experienced significantly higher presence [7]. However, this may have been a result of the voice similarity group actually having to give the speech while the dissimilarity group only had to act it out. In a study on synthesized voices, participants evaluated voices that were designed to have different personalities (extroverted vs. introverted). The authors found consistent support for similarity-attraction—i.e., participants evaluated higher the voice more similar to their own personality [137]. Studies suggest that the perceived gender of the voice can also influence users.

During a lecture in which the same person spoke as both a male and female (both voice morphed), students evaluated the female as more likeable and the male as more intelligent [49]. In a decision-making study, a male-voiced computer influenced the user's decision significantly more often than the female counterpart [111]. In a study with computerized voice output, three gender stereotypes

were found: male evaluation as more valid than female evaluation; dominance in females as unbecoming; and women knowing more about feminine topics (facts relating to love-and-relationships), with men knowing more about masculine topics (facts relating to computers) [138][1]. Consistent with such stereotyping, one study found that informative male and sociable female voice agents led to more positive assessments of an autonomous vehicle compared to stereotype-inconsistent gender matching (i.e., informative female, sociable male) [114]. Therefore, the gender of the avatar voice may be crucial to its influence on player experience.

Studies show, however, that pitch can also affect voice evaluation. Studies that manipulate voice pitch across multiple languages and cultures have found that men's and women's voices with lowered pitch are perceived as more dominant and masculine than those with raised pitch [5, 22, 93, 110, 145, 152]. People often assess voice pitch as being associated with a certain body mass and height [39, 40, 63, 69, 184], attractiveness [39, 63, 144, 154, 207], and age [63]. Studies show that a pitch manipulation of 20 Hz is sufficient to alter attractiveness ratings of voices [60–63, 92, 189]. In our study, we examine the effects of voice modulation (i.e., pitch manipulation). Specifically, we are interested in the interaction between gender and modulation direction, hypothesizing that a lower modulation will result in more positive outcomes in our programming game because of STEM gender stereotypes. Here, we conduct the one of the first studies to look at either voice similarity or modulation and their effects on player experience (PX) in games.

2.3 Player Experience in Games

The past two decades have seen the development of a number of instruments to measure PX. These include the Game Immersion Questionnaire (GIQ) [35], the Immersive Experience Questionnaire (IEQ) [88], the Game Engagement Questionnaire (GEQ) [24], the Game Experience Questionnaire (GEQIJ) [84], the Digital Games Motives Scale (DGMS) [44], the Player Experience of Need Satisfaction (PENS) [168], and the Player Experience Inventory (PXI) [1]. In this study, we leverage the PENS because it is based on a well-grounded theoretical framework [168] and allows us to better contextualize our results in the existing literature, which uses the PENS as a theoretical framework to explicate avatar identification (e.g., [16]). More specifically, the PENS is based on self-determination theory (SDT). SDT, as originally conceptualized, consists of three core building blocks to explain human motivation, which in turn lead to greater performance, persistence, and creativity [46, 47, 167]. These building blocks are *competence*, the need for being effective at achieving desired objectives; *autonomy*, the need for having the ability to make decisions; and *relatedness*, the need for social closeness with others. This original model has been extended to games by including *presence/immersion*, the sense of actually being transported into the game world and *intuitive controls*, the intuitiveness of controls. In addition to the PENS, we also leverage the Intrinsic Motivation Inventory (IMI) [125], through which we measure *interest/enjoyment*, *effort/importance*, *pressure/tension*, and *value/usefulness*. Through the *interest/enjoyment* subscale, the IMI measures intrinsic motivation—the desire to complete an activity because of the satisfaction of doing so in and of itself.

Need satisfaction is essential for intrinsic motivation to exist [125]. A study in an endless runner game found that avatar identification increases autonomy, immersion, interest/enjoyment, effort/importance, positive affect, and time spent [16]. A study that involved playing an educational programming game, then making a custom game level for that same game, found that avatar identification increases need satisfaction, intrinsic motivation, self-efficacy, time spent, and quality

[1]In discussing gender stereotypes, we acknowledge that although researchers have found these stereotypes to often be consistent across culture [48] and time [77], other studies have found some variability [52, 64]. Therefore, it is important to note that studies cited in this section were (1) based on stereotypes validated in the social scientific literature during a recent time period prior to each study and (2) for the culture from which the studies' participants are drawn.

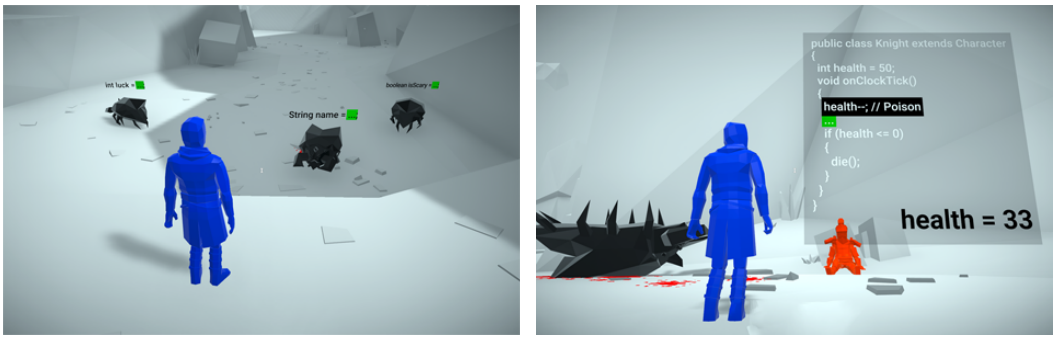

Fig. 1. Data type puzzle (L). Curing a wounded knight (R). Placeholders . . . indicate where code snippets can be thrown.

of game levels [100]. A study in a jumping game found that avatar identification increases need satisfaction and time spent playing [98]. Here, we are interested in determining whether voice similarity influences need satisfaction and intrinsic motivation, as well as the potential mediating effect of avatar identification.

2.4 Hypotheses

Building on the literature and arguments presented thus far, we pose the following hypotheses.
H1: Higher voice similarity will lead to more positive avatar identification, need satisfaction, intrinsic motivation, and performance.
H2: Avatar identification will mediate more positive need satisfaction, intrinsic motivation, and performance—i.e., voice similarity will lead to a higher level of avatar identification, which will in turn increase these outcomes.
H3: Consistent with gender stereotypes in STEM, voice modulation upwards/downwards will have a negative/positive effect, respectively, on avatar identification, need satisfaction, intrinsic motivation, and performance.

3 EXPERIMENTAL TESTBED

3.1 The Game

Our experimental testbed is CodeBreakers[2] [97], which was created for conducting avatar-based studies. CodeBreakers is a Java programming game in which players solve increasingly difficult problems by throwing snippets of code. See Figure 1. CodeBreakers was iteratively created with feedback from professional game developers, game designers, and Java developers, and included informal play testing over an eighteen-month span with playtesters. There were 14 total puzzles, spanning 6 levels. CodeBreakers was designed to incorporate best practices on effective learning curves [119]. Programming topics include data types, conditionals and control flow, classes and objects, inheritance and interfaces, loops and recursion, and data structures. Each puzzle had up to 3 hints, which are increasingly detailed. Players controlled their character using the keyboard and mouse. We measured performance through the number of puzzles completed. Players could exit at any time once they began playing. CodeBreakers was made available for machines running either Microsoft Windows or macOS.

[2]Gameplay video: https://youtu.be/x5U-Jd6tKXA

Fig. 2. Male (L) and female (R) avatars.

3.2 Validating Visual Avatar Design

For this experiment, the player avatar was purposefully designed to avoid known color effects (e.g., the color red is known to reduce mood, affect, and performance in cognitive-oriented tasks [71, 83, 99, 108, 127, 128]), to have ambiguous identity characteristics besides its gender, and to fit the game. We chose blue for the avatar color because it is not associated with negative cognitive effects and has comparable effects to other more neutral colors, such as gray, on test performance and heart rate variability (HF-HRV) [57]. Blue was also chosen to match the aesthetic of the game. The avatar models themselves were designed and created from scratch by a professional 3D game artist and were made intentionally abstract for ambiguity in identity. This was to reduce variance in how much players identified with the avatar's visual appearance. For example, an avatar with unambiguous identity characteristics would have a high similarity with only the subset of players who match those identity characteristics. See Figure 2.

To validate that these goals were met, we ran a validation study with 140 participants on Amazon Mechanical Turk (MTurk). We used a screening survey to retrieve 70 participants who self-identified as male and 70 participants who self-identified as female. After an audio check to ensure participants had their audio turned on, each participant played the base version of CodeBreakers (i.e., without any voice-related aspects) for a minimum of five minutes. All participants played with a gender-matched avatar. After five minutes, participants were allowed to quit at any time. After quitting, we asked participants the following questions: "How appropriate was the avatar color for the game?", "How appropriate was the avatar color for the avatar?", "How appropriate was the avatar clothing for the game?", "How appropriate was the avatar clothing for the avatar?", and "How appropriate was the avatar design overall?" on a scale from 1:*Very Inappropriate* to 5:*Very Appropriate*. Participants then answered two additional questions: "Besides its gender, the identity of my avatar was ambiguous (e.g., ethnicity/race)" and "My avatar resembled me" on a scale from 1:*Strongly Disagree* to 5:*Strongly Agree*. See Table 1. Participants were compensated $5.00 (USD) for taking part in this validation study.

We then performed independent samples t-tests between male and female participants. None of the tests were significant. ColorG: $t(138)=0.54$, $p=0.59$, $d=0.09$; ColorA: $t(138)=0.87$, $p=0.39$, $d=0.15$; ClothesG: $t(138)=0.98$, $p=0.33$, $d=0.17$; ClothesA: $t(138)=0.87$, $p=0.39$, $d=0.15$; DesignO: $t(138)=0.31$, $p=0.76$, $d=0.05$; Ambiguous: $t(138)=0.64$, $p=0.52$, $d=0.11$; Resemble: $t(138)=0.36$, $p=0.72$,

Gender	ColorG		ColorA		ClothesG		ClothesA		DesignO		Ambiguous		Resemble	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Male	4.23	0.80	4.24	0.79	4.24	0.79	4.23	0.80	4.27	0.80	4.34	0.98	2.64	1.14
Female	4.16	0.77	4.11	0.96	4.37	0.77	4.34	0.76	4.23	0.85	4.24	0.86	2.71	1.19

Table 1. Descriptive results from validation study validating that the avatar's visual characteristics were appropriate for the game (ColorG, ClothesG), for the avatar (ColorA, ClothesA), overall (DesignO), that the avatar's identity was viewed as ambiguous (Ambiguous), and that resemblance to the avatar across gender was similar (Resemble). All measures are on a 5-pt Likert scale.

d=0.06. Moreover, the average of all responses scored higher than *Agree* (4)—except for resemblance, which scored only slightly higher than neutral (~2.6–2.7). Therefore, these results validate our goals of avatars that fit the game and have ambiguous identity characteristics, without significant deviations across gender.

3.3 Voice Manipulation Platform

For this study, we developed an online platform that takes as input a single audio voice clip and is able to generate an arbitrary number of similar voice clips. We did this by leveraging recent advances in neural network-based speech synthesis [89, 194]. We started with an open-source implementation of *real-time voice cloning* [87] (see Section 3.3.1) and made several significant additions. In order to run a large-scale study using this, it was necessary to create a version that could be deployed remotely in the cloud and could be accessed on-demand. We did this by first deploying the software to an Amazon EC2 P2 server (type *p2.xlarge*), a GPU-based computing instance that has 1 GPU, 4 vCPUs, and 61 GiB of RAM located in the region us-east-1b [4]. This server leverages NVIDIA's Compute Unified Device Architecture (CUDA), which allows for GPU support in running the software. We then created a server which uses HTTP POST requests to communicate with clients with the following message types: *Upload* (uploads the sample voice clip while specifying any modulation parameters, returns a unique key for subsequent messages), *Status Check* (checks if job completed), and *Download* (get a single voice file, or get all voice files). When a client makes an upload request, this is placed in a queue and then served in order. We use 30 worker threads on our server so that multiple jobs can be processed concurrently. During the study, we kept watch over server performance (i.e., memory and GPU usage), which is important for consistent participant experiences. For example, high GPU usage would delay new requests from being processed in a timely manner. This was managed by limiting the number of concurrent study participants. To reduce wait time, clients only need to download the voice files for the next level to begin playing (while the remaining voice files are downloaded in the background asynchronously). It takes 15±5s (variation dependent on internet speed) for a U.S. based client to request and to download voice files for the first level. All voice files on the server are deleted on download. To perform voice modulation (i.e., pitch shifting), we first measure average fundamental frequency using a pitch floor of 75 Hz (Male) & 100 Hz (Female) and a pitch ceiling of 300 Hz (Male) & 500 Hz (Female) [102]. We performed pitch shifting using pitch synchronous overlap and add (PSOLA) [34]. PSOLA is frequently used in studies that manipulate voice [60–62, 70, 92, 93, 160, 189], and it affects only pitch, leaving other properties of voice perceptually unchanged [60, 61, 63]. After modulation, the sample is then processed through the voice cloning software. To mitigate differences

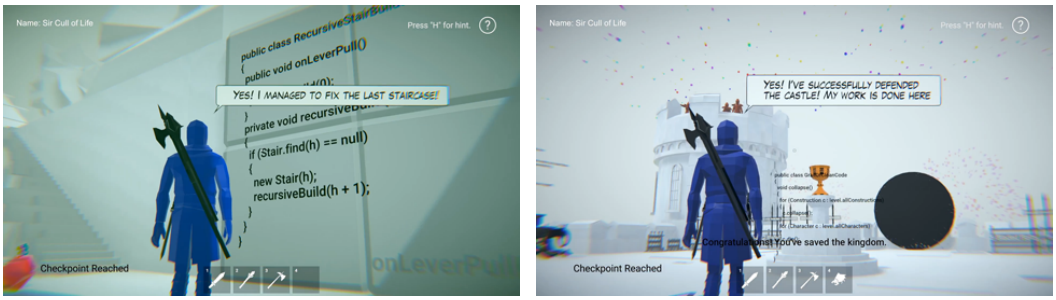

Fig. 3. Voice audio occurs in conjunction with speech bubbles that appear on top of the avatar.

in processing time, unmodulated samples are still dummy pitch-shifted to simulate the same time delay as modulated samples.

3.3.1 Voice Cloning Software Architecture. The open-source implementation of real-time voice cloning is described at a more granular level in [86], and the framework as a whole borrows heavily from [89]. There are three main components to the software architecture which are each trained separately: 1) a *speaker encoder* which creates embedding vectors representing the voice of the speaker [194]; 2) a *synthesizer* which takes input text and, conditioning the text on the speaker embedding vectors, generates a mel spectrogram [172]; and 3) a *vocoder* that converts the spectogram into an audio waveform [183]. Implementation and training details can be found in [89] and [86].

4 METHODS

4.1 Initial Voice Processing

To collect a sample of the participant's voice, during the beginning of the game, the player was asked to speak to an animated robot (Harley) which introduced the game to the participant. The participant was requested to speak the audio line: *Hello, my name is [the character name chosen by the participant], I am about to play CodeBreakers, and it is very nice to meet you, Harley.* We then checked the recorded audio for any long pauses without voice (>1 second) and that the entire audio length fit in a roughly acceptable interval for the voice cloning software (3—7 seconds). If these checks were violated, the participant was asked to re-record the audio until the audio was acceptable. The participant was then asked to listen to their recorded audio to ensure they could hear themselves speaking the sentence, and they were given the option of re-recording. After the voice sample was collected, it was sent to the server and processed while the participant completed the rest of the game introduction. This entire process was identical across conditions. During analysis, we manually checked every sample to ensure that the participant was clearly audible and had followed instructions; ~4.3% of participants were excluded based on this check. For sample rate, we use the default sampling rate from the participant's microphone. All samples are normalized to the same perceived volume using RMS (root mean square) normalization.

4.2 Conditions

The study uses a 2 x 3 factorial design. We manipulate avatar voice similarity (similar vs. dissimilar) and voice modulation (upwards vs. downwards vs. none). The manipulations are as follows:
Similar Voice: Avatar voice is generated using the participant's voice.

Dissimilar Voice: Avatar voice is generated using a gender-matched prior participant's voice. The voice was selected at random from a corpus of 10 (5 male and 5 female) samples collected during pilot testing.
Modulation Upwards: Original voice sample is pitch-shifted upwards by 20 Hz. We choose 20 Hz as it is a manipulation used in prior voice studies [60–63, 92, 189].
Modulation Downwards: Original voice sample is pitch-shifted downwards by 20 Hz.
No Modulation: Original voice sample is used. Dummy pitch-shifted to simulate time delay and ensure consistent user experiences across conditions.

Other than the above, all other aspects of the experiment were identical across conditions. In total, there were 30 possible voice lines that could have been triggered. Other than the first voice line (*What am I doing here? Did my ship crash? How long have I been lying here for? I guess I should get up and look around.*), audio lines typically come before and after each puzzle. For example, prior to puzzle #7: *The castle is under siege!.* And after completing puzzle #7: *It worked! I neutralized all of the bugs by using the staff.* These voice lines were accompanied by speech bubbles (see Figure 3). All audio aside from the avatar voice was identical across conditions, i.e., music and sound effects.

4.3 Measures

We use three validated PX questionnaires and gameplay metrics.

4.3.1 Avatar Identification. For measuring avatar identification, we use the player identification scale (PIS) [185]. The PIS measures three dimensions of avatar identification on a 5-pt Likert scale: similarity identification (e.g., "My character is similar to me"), embodied identification (e.g., "In the game, it is as if I become one with my character"), and wishful identification (e.g., "I would like to be more like my character").

4.3.2 Player Experience of Need Satisfaction. To measure need satisfaction, we use the PENS scale [168]. PENS measures the following dimensions on a 7-pt Likert scale: competence (e.g., "I feel competent at the game"), autonomy (e.g., "The game provides me with interesting options and choices"), relatedness (e.g., "I find the relationships I form in this game fulfilling"), presence/immersion (e.g., "When playing the game, I feel transported to another time and place"), and intuitive controls (e.g., "Learning the game controls was easy").

4.3.3 Intrinsic Motivation Inventory. To measure intrinsic motivation, we use the IMI [125]. We leverage the following IMI dimensions which use a 7-pt Likert scale: interest/enjoyment (e.g., "I enjoyed doing this activity very much"), effort/importance (e.g., "I put a lot of effort into this"), pressure/tension (e.g., "I felt very tense while doing this activity"), and value/usefulness (e.g., "I believe this activity could be of some value to me").

4.3.4 Game Performance. We automatically recorded metrics for game performance, including puzzles completed (max 14) and number of hints accessed (max 42). Re-played puzzles or re-accessed hints are not counted. These metrics were considered a reflection of motivation to engage in and learn from the game, which are clear intended outcomes of educational games.

4.3.5 Time Played. We operationalize motivated behavior as the time spent playing the game.

4.4 Participants

In total, 698 participants (30% female)[3] with an average age of M = 33.53 (SD = 9.55) were recruited through Amazon Mechanical Turk (MTurk). MTurk is a platform in which workers complete Human

[3]The smaller proportion of female participants was unexpected given the female skew overall on MTurk (over 60% in the U.S.) [166], and it was a possible byproduct of male participants being more attracted to playing a programming game.

Intelligence Tasks (HITs), including tasks for research studies. Studies show that MTurk provides data of similar quality [25], diversity [12, 33, 82], and reliability [25, 123] as typical samples (e.g., college students). Participants were each paid $7.50. Participants who answered multiple surveys with zero variance, or multiple surveys with ±3SD, were excluded. Participant voice recordings were manually checked by an author blind to condition to ensure they were audible and had followed instructions. After exclusion based on these criteria, we were left with 657 participants (29% female) for analysis, with an average age of M = 33.45 (SD = 9.66). The HIT was available to workers in the U.S. over the age of 18 who had a computer with a working microphone. For quality control, workers were required to have a HIT approval rate >95%. The Purdue University Institutional Review Board (IRB) approved the study. All participants were asked to provide informed consent.

4.4.1 Experience With Video Games and Programming. Participants reported playing an average of M=11.6 (SD=10.5) hours of video games per week, above the global average of M=8.45 [118]. On a scale from 1:*Minimal* to 7:*Extensive*, participants rated their prior experience playing video games ("How would you rate your prior experience playing video games?") as M=5.38 (SD=1.66) and their prior programming experience ("How would you rate your prior programming experience?") as M=2.64 (SD=1.77). Next, we adapted several questions on programming experience from [173]. On a scale from 1:*Very Inexperienced* to 5:*Very Experienced*, participants rated their programming experience compared to experts ("How do you estimate your programming experience compared to experts with 20 years of practical experience?") as M=1.43 (SD=0.93), their programming experience compared to beginners ("How do you estimate your programming experience compared to beginner programmers?") as M=2.33, (SD=1.29), their programming experience in Java specifically ("How experienced are you with the Java programming language?") as M=1.73 (SD=1.05), and their experience with an object-oriented paradigm ("How experienced are you with the object-oriented programming paradigm?") as M=1.91 (SD=1.21). Therefore, our sample contains participants who are regularly exposed to video games and have low prior programming experience. ANOVAs found that there were no significant differences between conditions on prior gaming experience ($F[5, 651]$=0.053, p=0.998, η_p^2=.000), programming experience ($F[5, 651]$=0.345, p=0.886, η_p^2=.003), and Java programming experience ($F[5, 651]$ =0.459, p=0.807, η_p^2=.004).

4.5 Design

A between-subjects factorial design was used. Each participant was randomly assigned to one of six possible conditions. Participant counts in each condition were approximately equal (M=109.5, SD=4.2), with a similar number of male (M=77.5, SD=3.6) and female (M=32.0, SD=5.3) participants across each condition.

4.6 Procedure

Participants first filled out an IRB-approved consent form. Participants were informed that they could exit the game at any time. Participants then began playing CodeBreakers. At the beginning of the game, participants underwent an audio check during which they were required to type a spoken English word. Next, the participant was asked to speak into their microphone to confirm that we could detect their audio input. Participants then selected a name and gender for their character. For the purposes of the experiment, participants were asked to choose the same gender as their real-life gender. We manually double-checked that their selected gender matched the gender reported post-experiment. A robotic agent (see Figure 4) then engaged in a short conversation with the player. The robot was animated with audio dialog generated through an automatic voice generator [120].

Nevertheless, we proceeded with our analyses as planned because the total number of female participants was still high (>200) and our statistical testing is robust to unequal group sizes.

Fig. 4. The robotic agent during the introduction asks the player to speak (L), and introduces the game (R).

After a brief introduction, the robot asked the participant to introduce himself/herself through their microphone. When the participant was ready to speak, they clicked on the *Record* button, then clicked *Stop Recording* when finished. In case the recording was too short (<3 seconds), too long (>7 seconds), or contained long pauses (>1 second), the player was asked to retry and to keep their dialog a continuous ~5 seconds in length. Once completed, participants were asked to confirm they could hear their recorded audio and to re-record otherwise. Next, the participant's audio was sent to the server for processing as they completed the remainder of the introduction. During the rest of the introduction, the participant was briefed on how to play the game. Participants were told they could exit the game at any time by pressing ESC on their keyboard, then clicking quit game. The participant then began playing the game. In the event that the participant's avatar voice files for the first level had not yet been completely downloaded, we showed a loading screen. Participants played the game with a blue gender-matched avatar. All participant game data was automatically logged. Once participants quit the game (or completed all 6 levels), they filled out a set of manipulation check questions, the PIS, PENS, and IMI. Participants were then asked to describe in their own words any problems encountered and what they thought the purpose of the experiment was. None of the participants correctly guessed the purpose of the experiment. Participants then filled out a set of questions about prior video game experience, programming experience, and demographics.

4.7 Analysis

Data was analyzed using SPSS 23 and the PROCESS macro for SPSS [78]. Independent t-tests were used to compare voice similarity versus voice dissimilarity on the outcomes of performance, time spent, PIS, PENS, and IMI. We then performed a mediation analysis using avatar identification as the mediator. Voice similarity coded as a dichotomous variable is X, avatar identification is the mediator M, and performance, time spent, PENS, and IMI are Y. We use mediation with each dimension of identification modeled individually (similarity, embodied, wishful) rather than parallel mediation. We chose to do this because of multicollinearity between identification dimensions (correlations>0.7), which can affect the estimation of mediation relationships in parallel mediation [78]. To investigate voice modulation, we use a 3x2 (modulation x gender) ANOVA[4] because we expected voice modulation effects to be moderated by gender. We use an α of 0.05.

Voice Condition	Pitch	Similar		Like Me		Gender		Age		Speaking		Friends		Likeable		Friendly		Fit	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Similar	*None*	4.33	1.54	4.31	1.61	6.28	1.08	4.79	1.52	4.22	1.67	4.20	1.70	4.71	1.67	4.79	1.64	4.78	1.50
	Up	4.17	1.51	4.17	1.45	6.13	1.28	4.91	1.34	4.27	1.53	4.25	1.50	4.68	1.42	4.81	1.46	4.75	1.16
	Down	4.19	1.68	4.17	1.79	6.17	1.32	4.81	1.46	4.18	1.81	4.26	1.66	4.40	1.61	4.52	1.66	4.90	1.52
Dissimilar	*None*	3.07	1.78	3.08	1.73	6.12	1.52	4.61	1.71	3.28	1.85	4.10	1.76	4.62	1.76	4.66	1.76	4.76	1.73
	Up	3.06	1.72	2.96	1.67	6.23	1.32	4.45	1.72	3.16	1.88	4.12	1.80	4.66	1.74	4.75	1.63	4.78	1.51
	Down	2.92	1.76	2.75	1.64	6.18	1.48	4.54	1.66	3.01	1.76	3.90	1.73	4.50	1.64	4.58	1.62	4.92	1.59

Table 2. Descriptive results of manipulation check.

Voice Condition	Pitch	Similar		Like Me		Gender		Age		Speaking		Friends		Likeable		Friendly		Fit	
		M_M	M_F	M_M	M_F	M_M	M_F	M_M	M_F	M_M	M_F	M_M	M_F	M_M	M_F	M_M	M_F	M_M	M_F
Similar	*None*	4.32	4.35	4.36	4.22	6.33	6.19	4.86	4.65	4.15	4.35	4.15	4.30	4.74	4.65	4.78	4.81	4.76	4.81
	Up	4.13	4.29	4.11	4.36	6.13	6.11	4.97	4.75	4.21	4.43	4.20	4.39	4.65	4.79	4.79	4.86	4.71	4.86
	Down	4.20	4.17	4.14	4.22	6.22	6.06	4.94	4.53	4.18	4.19	4.32	4.14	4.35	4.50	4.49	4.58	4.84	5.06
Dissimilar	*None*	3.02	3.20	3.05	3.20	6.13	6.08	4.65	4.48	3.29	3.24	4.10	4.12	4.63	4.60	4.71	4.48	4.74	4.84
	Up	3.17	2.79	2.99	2.90	6.26	6.17	4.45	4.45	3.21	3.04	4.17	4.00	4.77	4.38	4.81	4.59	4.77	4.79
	Down	2.97	2.81	2.79	2.68	6.26	6.03	4.61	4.41	3.05	2.92	3.92	3.84	4.55	4.41	4.64	4.46	4.90	4.97

Table 3. Manipulation check mean scores for participants self-identifying as male (M_M) and female (M_F).

5 RESULTS

5.1 Manipulation Check

The manipulation check consisted of 9 questions. The first 8 questions all began with "My avatar's voice sounded..." and ended with "similar to me"; "like me when I talk"; "the same gender as me"; "about my age"; "as if I was the one speaking"; "like someone I would be friends with"; "likeable"; and "friendly" on a 7-pt Likert scale (1:*Strongly Disagree* to 7:*Strongly Agree*). These questions assessed the voice similarity manipulation. The last question, "How well do you feel your avatar's voice fit with your avatar?" was on a 7-pt Likert scale (1:*Very Poorly* to 7:*Very Well*). This question assessed the perceived fit between the voice and the game avatar.

Between-subjects testing found that participants in the voice similarity condition scored significantly higher on "similar to me," $t(655)=9.36$, $p<0.001$, $d=0.73$, "like me when I talk," $t(655)=9.97$, $p<0.001$, $d=0.78$, "about my age," $t(655)=2.45$, $p<0.05$, $d=0.19$, and "as if I was the one speaking," $t(655)=7.87$, $p<0.001$, $d=0.61$, compared to participants in the voice dissimilarity condition. There were no significant differences across the remaining questions. From these results, the voice similarity manipulation was successful at inducing higher perceived avatar voice similarity. All conditions had a slightly above neutral fit between voice and avatar. See Table 2 and Table 3 for a more detailed breakdown of the manipulation check measures by voice similarity conditions as well as the voice modulation conditions and participant gender. The remaining results are organized by our hypotheses.

5.2 Effects of Voice Similarity

H1: *Higher voice similarity will lead to more positive avatar identification, need satisfaction, intrinsic motivation, and performance.*

[4]Both t-tests and ANOVAs are considered robust to non-normality, especially at larger sample sizes [20, 122].

Variable	Similar Voice		Dissimilar Voice	
	M	*SD*	*M*	*SD*
**Puzzles Completed	6.58	5.64	5.47	5.28
Hints Accessed	13.95	13.15	12.07	12.50
*Time Spent Sec.	899.75	838.26	743.80	739.30
***PIS Similarity	2.99	1.03	2.75	1.02
PIS Embodied	2.99	1.12	2.93	1.10
PIS Wishful	2.53	1.04	2.50	1.09
***PENS Competence	4.57	1.60	4.21	1.62
PENS Autonomy	4.30	1.58	4.17	1.59
**PENS Relatedness	3.44	1.55	3.10	1.58
***PENS Immersion	3.88	1.58	3.52	1.60
PENS Controls	4.73	1.57	4.75	1.58
IMI Enjoyment	4.73	1.48	4.60	1.55
IMI Effort	5.47	1.23	5.37	1.31
IMI Pressure	3.34	1.56	3.29	1.60
IMI Value	4.64	1.59	4.55	1.63

* significant at $p < .05$; ** significant at $p < .01$; *** significant at $p < .005$.

Table 4. Results for effects of voice similarity (H1). PIS was on a 5-pt Likert scale, while IMI and PENS were on a 7-pt Likert scale.

5.2.1 Performance. Participants in the voice similarity condition completed significantly more puzzles than participants in the voice dissimilarity condition, $t(655)=2.61$, $p<0.01$, $d=0.20$. There was no significant difference in hints used between the voice similarity condition and the voice dissimilarity condition, $t(655)=1.89$, $p=0.06$, $d=0.15$.

5.2.2 Time Spent. Participants in the voice similarity condition played for a significantly longer period of time than participants in the voice dissimilarity condition, $t(655)=2.53$, $p<0.05$, $d=0.20$.

5.2.3 PIS. Participants in the voice similarity condition had significantly higher similarity identification than participants in the voice dissimilarity condition, $t(655)=2.94$, $p<0.005$, $d=0.23$. There was no significant difference in embodied identification between the voice similarity condition and the voice dissimilarity condition, $t(655)=0.70$, $p=0.48$, $d=0.06$. There was no significant difference in wishful identification between the voice similarity condition and the voice dissimilarity condition, $t(655)=0.34$, $p=0.74$, $d=0.03$.

5.2.4 PENS. Participants in the voice similarity condition experienced significantly higher competence than participants in the voice dissimilarity condition, $t(655)=2.89$, $p<0.005$, $d=0.23$. There was no significant difference in autonomy between the voice similarity condition and the voice dissimilarity condition, $t(655)=1.02$, $p=0.31$, $d=0.08$. Participants in the voice similarity condition experienced significantly higher relatedness than participants in the voice dissimilarity condition, $t(655)=2.78$, $p<0.01$, $d=0.22$. Participants in the voice similarity condition experienced significantly higher immersion than participants in the voice dissimilarity condition, $t(655)=2.94$, $p<0.005$, $d=0.23$. There was no significant difference in intuitive controls between the voice similarity condition and the voice dissimilarity condition, $t(655)=0.12$, $p=0.90$, $d=0.01$.

	Similarity Identification					Embodied Identification					Wishful Identification				
	a	*b*	*c′*	*c*	*ab*	*a*	*b*	*c′*	*c*	*ab*	*a*	*b*	*c′*	*c*	*ab*
Performance															
Puzzles	**0.24*****	**0.660*****	**0.959***	**1.114****	**0.155; *CI*[0.038, 0.320]**	0.061	**0.791*****	**1.066***	**1.114****	0.048; *CI*[-0.088, 0.196]	0.028	-0.231	**1.121****	**1.114****	-0.006; *CI*[-0.066, 0.045]
Hints	—	**1.411*****	1.555	1.887	**0.332; *CI*[0.068, 0.704]**	—	**1.975*****	1.767	1.887	0.120; *CI*[-0.213, 0.494]	—	0.137	1.884	1.887	0.004; *CI*[-0.080, 0.107]
Time Spent															
Time Spent	—	**88.74*****	**135.1***	**156.0***	**20.90; *CI*[4.366, 44.60]**	—	**98.66*****	**150.0***	**156.0***	5.99; *CI*[-11.46, 25.23]	—	-21.27	**156.5***	**156.0***	-0.590; *CI*[-8.263, 5.479]
Player Experience of Need Satisfaction (PENS)															
Competence	—	**0.579*****	0.226	**0.362*****	**0.136; *CI*[0.046, 0.231]**	—	**0.596*****	**0.326*****	**0.362*****	0.036; *CI*[-0.065, 0.139]	—	**0.524*****	**0.348*****	**0.362*****	0.015; *CI*[-0.070, 0.101]
Autonomy	—	**0.640*****	-0.025	0.126	**0.151; *CI*[0.050, 0.257]**	—	**0.682*****	0.084	0.126	0.041; *CI*[-0.074, 0.156]	—	**0.592*****	0.109	0.126	0.016; *CI*[-0.081, 0.115]
Relatedness	—	**0.694*****	0.176	**0.340****	**0.164; *CI*[0.055, 0.276]**	—	**0.618*****	**0.302****	**0.340****	0.038; *CI*[-0.068, 0.141]	—	**0.655*****	**0.322*****	**0.340****	0.018; *CI*[-0.088, 0.126]
Immersion	—	**0.760*****	0.186	**0.365*****	**0.179; *CI*[0.057, 0.302]**	—	**0.769*****	**0.318*****	**0.365*****	0.047; *CI*[-0.084, 0.176]	—	**0.743*****	**0.344*****	**0.365*****	0.021; *CI*[-0.100, 0.142]
Controls	—	**0.460*****	-0.123	-0.015	**0.108; *CI*[0.034, 0.189]**	—	**0.498*****	-0.045	-0.015	0.030; *CI*[-0.056, 0.116]	—	**0.370*****	-0.025	-0.015	0.010; *CI*[-0.051, 0.070]
Intrinsic Motivation Inventory (IMI)															
Enjoyment	—	**0.591*****	-0.004	0.135	**0.139; *CI*[0.047, 0.236]**	—	**0.628*****	0.097	0.135	0.038; *CI*[-0.069, 0.145]	—	**0.479*****	0.122	0.135	0.013; *CI*[-0.065, 0.095]
Effort	—	0.032	0.094	0.102	0.008; *CI*[-0.016, 0.035]	—	**0.118****	0.095	0.102	0.007; *CI*[-0.014, 0.032]	—	0.042	0.101	0.102	0.001; *CI*[-0.009, 0.015]
Tension	—	-0.066	0.063	0.048	-0.016; *CI*[-0.051, 0.015]	—	-0.063	0.052	0.048	-0.004; *CI*[-0.023, 0.011]	—	0.023	0.047	0.048	0.001; *CI*[-0.010, 0.014]
Usefulness	—	**0.594*****	-0.050	0.090	**0.140; *CI*[0.048, 0.239]**	—	**0.584*****	0.055	0.090	0.036; *CI*[-0.065, 0.133]	—	**0.529*****	0.076	0.090	0.015; *CI*[-0.071, 0.102]

* significant at $p < .05$; ** significant at $p < .01$; *** significant at $p < .005$; significant *ab* based on 95% CI.

Table 5. Mediation results with voice similarity (X), avatar identification (M), and outcome (Y). Regression coefficients a ($X \rightarrow M$), b ($M \rightarrow Y$), c' (direct $X \rightarrow Y$), c (total $X \rightarrow Y$), and *ab*. Significant results are bold.

5.2.5 IMI. There was no significant difference in enjoyment between the voice similarity condition and the voice dissimilarity condition, $t(655)=1.14$, $p=0.25$, $d=0.09$. There was no significant difference in effort between the voice similarity condition and the voice dissimilarity condition, $t(655)=1.03$, $p=0.31$, $d=0.08$. There was no significant difference in pressure between the voice similarity condition and the voice dissimilarity condition, $t(655)=0.39$, $p=0.70$, $d=0.03$. There was no significant difference in value between the voice similarity condition and the voice dissimilarity condition, $t(655)=0.72$, $p=0.47$, $d=0.06$.

5.2.6 Summary of Results. Higher voice similarity leads to a significant increase in performance, time spent, similarity identification, competence, relatedness, and immersion. Effect sizes (d) range from 0.2 to 0.23, making these effects small. However, given the complexity of player-game interactions, small effect sizes are not uncommon in games user research [16, 19, 177, 205]. Embodied and wishful identification, autonomy, controls, and intrinsic motivation were unaffected. See Table 4.

5.3 Avatar Identification as a Mediator

H2: *Avatar identification will mediate more positive need satisfaction, intrinsic motivation, and performance.*

From Table 5, we can see that voice similarity leads to higher similarity identification (*a*) and that higher similarity identification was subsequently related to higher performance, time spent, need satisfaction, interest/enjoyment, and value/usefulness (*b*). A 95% bias-corrected confidence interval based on 10,000 bootstrap samples indicates that the indirect effects (*ab*) are also significant. Therefore, we conclude that similarity identification significantly mediates the relationship between voice similarity and performance, time spent, need satisfaction, and intrinsic motivation.

Embodied identification was related to higher performance, time spent, need satisfaction, interest/enjoyment, effort/importance, and value/usefulness. Wishful identification was related to higher need satisfaction, interest/enjoyment, and value/usefulness. Indirect effects for embodied and wishful identification were non-significant.

5.4 Effects of Voice Modulation

H3: *Consistent with gender stereotypes in STEM, voice modulation upwards/downwards will have a negative/positive effect, respectively, on avatar identification, need satisfaction, intrinsic motivation, and performance.*

	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)		(9)		(10)		(11)		(12)		(13)		(14)		(15)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Male																														
Pitch–	6.63	5.63	13.68	13.13	876.4	765.8	3.02	1.05	3.06	1.05	2.58	1.08	4.62	1.53	4.38	1.61	3.42	1.59	3.80	1.60	4.90	1.47	4.68	1.38	5.36	1.23	3.10	1.49	4.60	1.58
Pitch↑	6.53	5.39	13.60	12.53	820.4	766.3	2.99	0.98	3.08	1.09	2.56	1.07	4.69	1.58	4.50	1.54	3.32	1.56	3.86	1.56	5.18	1.46	4.87	1.48	5.55	1.29	2.90	1.39	4.84	1.57
Pitch↓	6.56	5.39	13.37	13.14	809.0	712.5	2.92	1.03	3.01	1.08	2.59	1.08	4.53	1.49	4.20	1.41	3.32	1.55	3.64	1.60	4.82	1.45	4.55	1.54	5.22	1.22	3.04	1.36	4.45	1.58
Female																														
Pitch–	4.45	5.07	11.61	11.84	693.1	644.6	2.59	1.06	2.74	1.18	2.39	1.07	3.89	1.73	3.69	1.80	3.06	1.50	3.63	1.65	4.12	1.72	4.51	1.73	5.55	1.28	4.11	1.76	4.30	1.66
Pitch↑	4.30	5.11	11.21	12.81	820.7	940.3	2.58	0.92	2.73	1.15	2.23	0.93	3.57	1.67	3.84	1.63	2.94	1.68	3.25	1.54	4.16	1.58	4.53	1.59	5.59	1.27	3.71	1.68	4.56	1.60
Pitch↓	5.18	5.74	12.15	13.34	845.1	1039.4	2.64	1.09	2.76	1.16	2.47	1.04	4.00	1.68	4.18	1.61	3.22	1.59	3.67	1.66	4.32	1.79	4.68	1.57	5.49	1.37	4.23	1.73	4.62	1.75
Main Effect Gender																														
F	**17.035**		2.915		0.512		**18.217**		**10.353**		**5.359**		**34.152**		**11.391**		**4.289**		3.345		**33.320**		0.915		2.394		**59.511**		0.949	
p	**<0.001**		0.088		0.474		**<0.001**		**<0.005**		**<0.05**		**<0.001**		**<0.001**		**<0.05**		0.068		**<0.001**		0.339		0.122		**<0.001**		0.330	
η_p^2	**0.026**		0.004		0.001		**0.027**		**0.016**		**0.008**		**0.050**		**0.017**		**0.007**		0.005		**0.049**		0.001		0.004		**0.084**		0.001	
Main Effect Pitch																														
F	0.352		0.035		0.149		0.033		0.021		0.737		0.379		0.497		0.393		0.435		0.460		0.234		1.331		2.501		1.079	
p	0.703		0.966		0.861		0.968		0.980		0.479		0.684		0.609		0.675		0.648		0.632		0.792		0.265		0.083		0.340	
η_p^2	0.001		0.000		0.000		0.000		0.000		0.002		0.001		0.002		0.001		0.001		0.001		0.001		0.004		0.008		0.003	
Interaction Effect																														
F	0.357		0.104		0.996		0.296		0.104		0.466		1.556		2.818		0.451		1.854		1.271		1.115		0.359		0.727		1.308	
p	0.700		0.901		0.370		0.744		0.901		0.628		0.212		0.060		0.637		0.157		0.281		0.328		0.698		0.484		0.271	
η_p^2	0.001		0.000		0.003		0.001		0.000		0.001		0.005		0.009		0.001		0.006		0.004		0.003		0.001		0.002		0.004	

Gender df=1, Pitch df=2, Interaction df=2, Error df=651

(1) Puzzles Comp.
(2) Hints Accessed
(3) Time Spent Sec.
(4) PIS Similarity
(5) PIS Embodied
(6) PIS Wishful
(7) PENS Comp.
(8) PENS Autonomy
(9) PENS Related.
(10) PENS Immersion
(11) PENS Controls
(12) IMI Enjoyment
(13) IMI Effort
(14) IMI Pressure
(15) IMI Value

Table 6. Results for effects of voice modulation (H3). Significant results are bold.

From Table 6, 2x3 ANOVAs (gender x voice modulation) found main effects of gender for puzzles completed, similarity identification, embodied identification, wishful identification, competence, autonomy, relatedness, intuitive controls, and pressure/tension. No main effects of voice modulation were found. No interaction effects between gender and voice modulation were found. Therefore, voice modulation had a negligible impact on outcomes.

6 DISCUSSION

Existing literature has shown that an avatar's visual appearance affects its user [200]. Such effects are moderated by how much we *identify* with the avatar. This identification can be increased through visual avatar customization. However, it remains unclear whether the *audial* aspects of an avatar influences identification and other outcomes.

Here, we conducted a 2 x 3 (voice similarity x voice modulation) experiment with neural network voice cloning. Higher voice similarity directly increases game performance[5], time spent, similarity identification, competence, relatedness, and immersion. Mediation analysis found that similarity identification (M) mediates between voice similarity (X) and performance, time spent, need satisfaction, and intrinsic motivation (Y). Therefore, avatar voice influences crucial PX outcomes.

Surprisingly, pitch shifting had no significant effect. Although studies have shown that manipulating pitch by 20 Hz alters attractiveness of a voice [60–63, 92, 189], our measurement instruments did not focus on attractiveness and were instead geared towards PX and performance outcomes. Furthermore, our study takes place during gameplay, not an environment where the player's sole focus is on evaluating the audio being presented. The player's focus is instead divided between

[5]For the similar voice condition, we see a significant increase in performance, but also a (non-significant) increase in hints accessed (see Table 4). The mean increase in puzzles completed is ~1, while the mean increase in hints accessed is ~2. Each puzzle contains three hints, with the first two designed to guide the player towards the answer (e.g., "Look around the environment.") and the last hint providing the answer. Therefore, the increase in hints accessed alone does not explain the performance increase. We interpret both the increased performance and the increased time spent as behavioral indicators of increased motivation to engage in and learn from the game.

cognitive processes—e.g., visually interpreting the scene and learning gameplay. Therefore, voice modulation may have been too subtle for observable effects. Larger modulations (e.g., 40 Hz) should be considered in future research on the potential for avatar voices to influence stereotype effects.

6.1 Applications to Games

Although the amount of dialogue in our game can be considered minimal compared to other games, such as *Mass Effect* [130], even this amount of higher player-similar audio significantly promoted gameplay performance and influenced PX. This has significant implications for audio in games. Game companies can create more engaging experiences through similar voice audio, leading to greater commercial success. Similarly, games that promote health (e.g., exercise [10]), learning (e.g., educational games [2]), and discovery (e.g., citizen science [41]) could benefit from increased engagement. Engagement can translate to better habits, greater learning gains, and increased scientific discoveries. Our results show that increasing similarity identification through higher voice similarity results in increased need satisfaction, intrinsic motivation, and motivated behavior. These outcomes are important across virtually all games.

6.2 Broader Voice Applications

Virtual environments more generally that contain voiced characters could also benefit from voice similarity. For example, consider an intelligent agent, designed for math learning, whose voice resembles the user's. An intelligent agent perceived as being similar could improve learning outcomes [9, 11, 75, 103]. Applications for learning a new language could similarly benefit users, as hearing how one's own voice *should* sound could help users more easily imitate speech. Or consider VR oil rig safety training where the narrator's voice resembles the trainee's. A similar narrator could lead to more engaged and immersive training.

Many real-world devices incorporate voice assistants such as Siri [6], Cortana [129], Google Assistant [72], and Alexa [3], and these are increasingly prevalent in homes, cars, and mobile devices. Although the effect of voice similarity with these assistants has not been studied directly, the present findings suggest that increasing voice similarity would lead to more positive interactions with such voice assistants. More research is needed on the extensive number of potential use cases for voice similarity.

6.3 Audio Customization

While we demonstrated these results in a controlled lab experiment, players will likely experience even greater similarity identification and affected outcomes in realistic volitional play contexts where players engage with their virtual representations over a longer period of time. For instance, research suggests that over time we become more congruent with our virtual identities [51, 159, 200, 203]. Of the types of identification measured (similarity, embodied, and wishful), only similarity was affected. While expected due to the manipulation of voice similarity, the avatar customization process on the other hand has been shown to increase similarity, embodied, and wishful identification [16]. For example, the options during avatar customization allow players to create not only themselves but an ideal that they would like to become [14]. This leads us to believe that customization of avatar audio, similar to customization of an avatar's visual appearance, would be beneficial for fostering avatar identification.

Although still not common, some games allow for customization of avatar audio. Games such as *Final Fantasy XIV* [176], *Saints Row IV* [188], and *Monster Hunter: World* [29], allow for selection of different pre-created collections of voice audio. Other games allow the player to directly manipulate the voice itself. *Black Desert Online* [148] and *Red Dead Redemption 2* [162] both allow for customization of pitch, with the latter introducing an additional "clarity" parameter. *The Sims*

4 [56] allows pitch adjustment and choosing between 'sweet,' 'melodic,' and 'lilted' for women, and between 'clear,' 'warm,' and 'brash' for men. However, more extensive audio customization in games does not currently exist. With these limited parameters, a self-similar voice is not possible in most circumstances.

Nevertheless, more complex avatar audio customization could be highly beneficial. Allowing users to create similar (and perhaps embodied and wishful, as is possible with visual avatar customization) audial identities gives rise to new possibilities for identification (possibly leading to stronger emotional attachments [21]), thereby enhancing a wide range of PX outcomes.

6.4 Behavioral Influence

This line of research on audial avatar identities is also relevant to the Proteus effect, the phenomenon that avatar users tend to conform behaviorally to the identity characteristics that they associate with their avatars [200]. This phenomenon has been studied extensively with respect to avatar appearance [158], but not with respect to avatar voice characteristics. Just as taller avatars lead to more aggressive negotiation [201], healthier-weight avatars lead to more physical activity [116], and inventor-looking avatars lead to more creative brainstorming [76], an avatar that sounds more confident, healthy, and creative could also cause enactments of those attributes. Future research on the Proteus effect could use the methods adopted in the present study to confirm these expectations.

7 LIMITATIONS

Controlled experiments with random assignment are considered robust. However, compensating participants to play a game in a controlled lab setting is fundamentally dissimilar than playing of one's own volition. Future studies should seek to understand whether these results extend to voluntary play.

As our study design was relatively complex, there was an inherent degree of randomness in our conditions. For example, the three conditions that that were aggregated into the similar voice condition had slight degrees of dissimilarity due to the pitch modulation. Similarly, the dissimilar voice was cloned from a random corpus of 10 participant voices and was also aggregated with pitch modulated versions. Nevertheless, these comparisons can be performed given the large sample size and the manipulation check. For example, Table 2 validates that pitch modulated voices did not differ greatly in similarity from their unmodulated counterparts. That being said, it is important to note that technically, while the similar voice was viewed as having higher than average similarity with the player (~4.23), this cannot be considered to be truly a *very* similar voice. This is mostly a technological constraint in that the state-of-the-art in voice cloning is currently unable to consistently generate very similar voices across all speakers. Future studies might address this through collecting a larger corpus of participant audio to train deep learning algorithms to create an even better matching voice prior to conducting the experiment. Moreover, the effect sizes of our results fall in the small range. Nonetheless, this study has successfully compared a voice that sounds more like the player to a voice that sounds less like the player, illustrating significant differences in PX. The implications of such results can be of value to the HCI community more broadly, as audio is often understudied in comparison to visual aspects of games and other systems.

This study used a single, education-oriented game that was designed for research purposes. Hence, generalizability was not established for the types of games or media applications that are used more commonly, such as entertainment-oriented action games or mobile phone operating systems. The influence of voice similarity may depend on facets of the media design (e.g., pacing, opportunities for voice-based interaction) as well as user orientations toward the media (e.g., playing for fun or to learn). Future research could examine such factors as moderating effects of voice similarity on PX.

This research was designed to examine voiced avatars that speak for the avatar user, presumably within single-player games or applications. However, many multi-user applications offer voice-based communication [191, 192], which enhances user experiences and social trust [198], although users rarely actually hear their own voices. That said, previous research suggests that when gender is communicated through voices in online games, women are more likely to receive toxic treatment [186, 193], which potentially triggers stereotype threat and causes psychological harm [65]. Although the present research did not find any differences in stereotype-related outcomes due to voice pitch modulation, the findings do suggest that user voices are malleable, just like the visual characteristics of avatars. Technologies are currently available to consumers that facilitate voice modification in multi-user games and other applications (e.g., [124, 187]), offering the potential to switch genders or even species. Future research could use such tools to examine voice avatars and stereotype effects in multi-user voice-communication contexts, e.g., social VR [68].

There were aspects of the experiment that were not entirely under our control. The quality of the microphone and audio, for example, depend on what devices are owned by the participant. However, using participants' own devices increases ecological validity as this is more typical to how a person would play a game compared to a lab. Other aspects, however, could have also played a role in the experiment. For example, we performed an audio check to ensure participants could hear audio at the beginning of the experiment, and we additionally recorded participants' system audio level whenever a voice line was triggered, but we had no control over the specific volume being used or whether they were really listening (e.g., putting their headphones down on the table).

Our research on voice pitch and stereotypes is based on decades of work on evolutionary behavior. There are common associations between voice pitch and masculinity, femininity, and dominance, and these associations exist across animal species and nonhuman primates [81, 134]. Furthermore, the "universality of voice pitch sexual dimorphism" has led researchers to argue that such associations are expected to hold across cultures [155]. Nevertheless, this should not be taken for granted and such studies should be replicated in non-U.S. contexts.

One aspect not directly studied is the degree of similarity. For example, with too little similarity, there may be no effect; too much similarity and it may be strange (e.g., an audial analogue to the uncanny valley, which refers to revulsion for nearly human-looking avatars [133]). Similarly, there are ethical concerns that need to be explored prior to broadly deploying voice manipulation. A recent workshop hosted by the U.S. Federal Trade Commission (FTC) discussed both the risks and benefits of voice cloning [58]. Risks include fraud and harassment, while benefits include synthesizing voices for those suffering from amyotrophic lateral sclerosis (ALS), Huntington's disease, and autism. Nevertheless, the full implications of voice cloning are still unfolding.

8 CONCLUSION

Avatar identification is a topic of extensive research. Despite widespread acknowledgment of how avatar identification benefits users, existing studies have focused on visual appearance of avatars. We presented one of the first studies to date on avatar self-similar audio. Higher voice similarity leads to a significant increase in performance, time spent, similarity identification, competence, relatedness, and immersion. Similarity identification acts as a significant mediator variable between voice similarity and performance, time spent, need satisfaction, and intrinsic motivation. We discussed the wide-ranging implications of these results for games and beyond. This study is an important step towards understanding voice audio effects.

REFERENCES

[1] Vero Vanden Abeele, Katta Spiel, Lennart Nacke, Daniel Johnson, and Kathrin Gerling. 2020. Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and

psychosocial consequences. *International Journal of Human Computer Studies* 135, January 2019 (2020), 102370. https://doi.org/10.1016/j.ijhcs.2019.102370

[2] Ma Victoria Almeda, Erica Kleinman, Chaima Jemmali, Carter Ithier, Elizabeth Rowe, and Magy Seif El-Nasr. 2020. Labeling debugging in may's journey gameplay. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. https://doi.org/10.1145/3328778.3372624

[3] Amazon. 2020. Alexa. https://developer.amazon.com/en-US/alexa

[4] Amazon. 2020. Amazon EC2 P2 Instances. https://aws.amazon.com/ec2/instance-types/p2/

[5] Moya L. Andrews and Charles P. Schmidt. 1997. Gender presentation: Perceptual and acoustical analyses of voice. *Journal of Voice* 11, 3 (1997), 307–313. https://doi.org/10.1016/S0892-1997(97)80009-4

[6] Apple. 2020. Siri. https://www.apple.com/siri/

[7] Laura Aymerich-Franch, Cody Karutz, and Jeremy N Bailenson. 2012. Effects of Facial and Voice Similarity on Presence in a Public Speaking Virtual Environment. *ISPR Presence Live Conference* (2012), 1–7.

[8] Christine M. Bachen, Pedro Hernández-Ramos, Chad Raphael, and Amanda Waldron. 2016. How do presence, flow, and character identification affect players' empathy and interest in learning from a serious computer game? *Computers in Human Behavior* 64 (2016), 77–87. https://doi.org/10.1016/j.chb.2016.06.043

[9] Jeremy N. Bailenson, Jim Blascovich, and Rosanna E. Guadagno. 2008. Self-representations in immersive virtual environments. *Journal of Applied Social Psychology* 38, 11 (2008), 2673–2690.

[10] Anna Barenbrock, Marc Herrlich, Kathrin Maria Gerling, Jan David Smeddinck, and Rainer Malaka. 2018. Varying avatar weight to increase player motivation: Challenges of a gaming setup. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018), 1–6. https://doi.org/10.1145/3170427.3188634

[11] Al Baylor and Yanghee Kim. 2004. Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. *Intelligent Tutoring Systems* 1997 (2004), 592–603. https://doi.org/10.1007/978-3-540-30139-4_56

[12] Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. 2012. Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis* 20, 3 (2012), 351–368.

[13] Axel Berndt and Knut Hartmann. 2008. The functions of music in interactive media. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 5334 LNCS (2008), 126–131. https://doi.org/10.1007/978-3-540-89454-4_19

[14] K Bessière, AF Seay, and S Kiesler. 2007. The ideal elf: Identity exploration in World of Warcraft. *CyberPsychology & Behavior* (2007). http://online.liebertpub.com/doi/abs/10.1089/cpb.2007.9994

[15] Frank Biocca. 1997. Cyborg's dilemma: Embodiment in virtual environments. In *Proceedings of the International Conference on Cognitive Technology*.

[16] Max V Birk, Cheralyn Atkins, Jason T Bowey, and Regan L Mandryk. 2016. Fostering Intrinsic Motivation through Avatar Identification in Digital Games. *CHI* (2016). https://doi.org/10.1145/2858036.2858062

[17] Max V. Birk and Regan L. Mandryk. 2018. Combating Attrition in Digital Self-Improvement Programs using Avatar Customization. *CHI '18: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2018), 1–15. https://doi.org/10.1145/3173574.3174234

[18] Max V Birk, Regan L Mandryk, and Cheralyn Atkins. 2016. The Motivational Push of Games. *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16* April 2017 (2016), 291–303. https://doi.org/10.1145/2967934.2968091

[19] Max V. Birk, Regan L. Mandryk, Matthew K. Miller, and Kathrin M. Gerling. 2015. How self-esteem shapes our interactions with play technologies. In *CHI PLAY 2015 - Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. https://doi.org/10.1145/2793107.2793111

[20] María J. Blanca, Rafael Alarcón, Jaume Arnau, Roser Bono, and Rebecca Bendayan. 2017. Non-normal data: Is ANOVA still a valid option? *Psicothema* (2017). https://doi.org/10.7334/psicothema2016.383

[21] Julia Ayumi Bopp, Livia J. Müller, Lena Fanya Aeschbach, Klaus Opwis, and Elisa D. Mekler. 2019. Exploring emotional attachment to game characters. *CHI PLAY 2019 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (2019), 313–324. https://doi.org/10.1145/3311350.3347169

[22] Barbara Borkowska and Boguslaw Pawlowski. 2011. Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour* 82, 1 (2011), 55–59. https://doi.org/10.1016/j.anbehav.2011.03.024

[23] Nicholas David Bowman, Mary Beth Oliver, Ryan Rogers, Brett Sherrick, Julia Woolley, and Mun-Young Chung. 2016. In control or in their shoes? How character attachment differentially influences video game enjoyment and appreciation. *Journal of Gaming & Virtual Worlds* 8, 1 (2016), 83–99. https://doi.org/10.1386/jgvw.8.1.83_1

[24] Jeanne H Brockmyer, Christine M Fox, Kathleen A Curtiss, Evan McBroom, Kimberly M Burkhart, and Jacquelyn N Pidruzny. 2009. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology* 45, 4 (2009), 624–634.

[25] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1 (2011), 3–5.

[26] Stéphanie Buisine, Jérôme Guegan, Jessy Barré, Frédéric Segonds, and Améziane Aoussat. 2016. Using avatars to tailor ideation process to innovation strategy. *Cognition, Technology and Work* (2016). https://doi.org/10.1007/s10111-016-0378-y

[27] Donn Byrne and Don Nelson. 1965. Attraction as a linear function of proportion of positive reinforcements. *Journal of Personality and Social Psychology* 1, 6 (1965), 659–663. https://doi.org/10.1037/h0022073

[28] Jaehwan Byun and Christian S. Loh. 2015. Audial engagement: Effects of game sound on learner engagement in digital game-based learning environments. *Computers in Human Behavior* 46, May (2015), 129–138. https://doi.org/10.1016/j.chb.2014.12.052

[29] Capcom. 2018. Monster Hunter: World. Game [Multiple Platforms].

[30] Marcus Carter, Fraser Allison, John Downs, and Martin Gibbs. 2015. Player identity dissonance and voice interaction in games. *CHI PLAY 2015 - Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (2015), 265–270. https://doi.org/10.1145/2793107.2793144

[31] Gianna Cassidy and Raymond MacDonald. 2009. The effects of music choice on task performance: A study of the impact of self-selected and experimenter-selected music on driving game performance and experience. *Musicae Scientiae* (2009). https://doi.org/10.1177/102986490901300207

[32] G G Cassidy and Raymond A R MacDonald. 2010. The effects of music on time perception and performance of a driving game. *Scandinavian journal of psychology* 51, 6 (2010), 455–464.

[33] Jesse Chandler and Danielle Shapiro. 2016. Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology* 12 (2016).

[34] Ff Charpentier and M Stella. 1986. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 11. IEEE, 2015–2018.

[35] M-T Cheng, H-C She, and Leonard A Annetta. 2015. Game immersion experience: its hierarchical structure and impact on game-based science learning. *Journal of Computer Assisted Learning* 31, 3 (2015), 232–253.

[36] Klimmt Christoph, Hefner Dorothée, and Vorderer Peter. 2009. The Video Game Experience as "True" Identification: A Theory of Enjoyable Alterations of Players' Self-Perception. *Communication theory* 19, 4 (2009), 351–373.

[37] Jonathan Cohen. 2001. Defining identification: A theoretical look at the identification of audiences with media characters. *Mass communication & society* 4, 3 (2001), 245–264.

[38] Jonathan Cohen. 2006. Audience identification with media characters. *Psychology of entertainment* 13 (2006), 183–197.

[39] Sarah A Collins. 2000. Men's voices and women's choices. *Animal behaviour* 60, 6 (2000), 773–780.

[40] Sarah A Collins and Caroline Missing. 2003. Vocal and visual attractiveness are related in women. *Animal behaviour* 65, 5 (2003), 997–1004.

[41] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit Players. 2010. Predicting protein structures with a multiplayer online game. *Nature* (2010). https://doi.org/10.1038/nature09304

[42] Nicole Crenshaw and Bonnie Nardi. 2014. What's in a Name? Naming Practices in Online Video Games. *CHI PLAY* (2014), 67–76.

[43] James J. Cummings and Jeremy N. Bailenson. 2016. How Immersive Is Enough? A Meta-Analysis of the Effect of Immersive Technology on User Presence. *Media Psychology* 19, 2 (2016), 272–309. https://doi.org/10.1080/15213269.2015.1015740

[44] Frederik De Grove, Verolien Cauberghe, and Jan Van Looy. 2016. Development and validation of an instrument for measuring individual motives for playing digital games. *Media Psychology* 19, 1 (2016), 101–125.

[45] Alwin De Rooij, Sarah Van Der Land, and Shelly Van Erp. 2017. The creative proteus effect: How self-similarity, embodiment, and priming of creative stereotypes with avatars influences creative ideation. In *C and C 2017 - Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition.* https://doi.org/10.1145/3059454.3078856

[46] Edward Deci and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior.* Plenum Press.

[47] Edward L. Deci and Richard M. Ryan. 2000. The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry* (2000). https://doi.org/10.1207/S15327965PLI1104_01

[48] Michel Désert and Jacques-Philippe Leyens. 2006. Social comparisons across cultures I: Gender. *Social comparison and social psychology: Understanding cognition, intergroup relations, and culture* (2006), 303.

[49] Mats Deutschmann, Anders Steinvall, Anna Lagerström, Mats Deutschmann, Anders Steinvall, and Anna Lagerström. 2011. Gender-Bending in Virtual Space - Using Voice-morphing in Second Life to Raise Sociolinguistic Gender Awareness. *V-lang International Conference, Warsaw* November (2011), 54–61.

[50] Edward Downs, Nicholas D. Bowman, and Jaime Banks. 2019. A polythetic model of player-avatar identification: Synthesizing multiple mechanisms. *Psychology of Popular Media Culture* (2019). https://doi.org/10.1037/ppm0000170

[51] Nicolas Ducheneaut, MH Wen, Nicholas Yee, and Greg Wadley. 2009. Body and mind: a study of avatar personalization in three virtual worlds. *CHI 2009* (2009). http://dl.acm.org/citation.cfm?id=1518877

[52] Alice H. Eagly, Christa Nater, David I. Miller, Michèle Kaufmann, and Sabine Sczesny. 2020. Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist* (2020). https://doi.org/10.1037/amp0000494

[53] Inger Ekman. 2005. Meaningful noise: Understanding sound effects in computer games. *Proc. Digital Arts and Cultures* 17 (2005).

[54] Inger Ekman. 2008. Psychologically Motivated Techniques for Emotional Sound in Computer Games. *Proc. AudioMostly 2008* January 2008 (2008), 20–26. https://meaningfulnoise.wordpress.com/psychologically-motivated-techniques-for-emotional-sound-in-computer-games/

[55] Inger Ekman. 2013. On the desire to not kill your players: Rethinking sound in pervasive and mixed reality games. *FDG* (2013), 142–149.

[56] Electronic Arts. 2014. The Sims 4. Game [Multiple Platforms].

[57] Andrew J. Elliot, Vincent Payen, Jeanick Brisswalter, Francois Cury, and Julian F. Thayer. 2011. A subtle threat cue, heart rate variability, and cognitive performance. *Psychophysiology* (2011). https://doi.org/10.1111/j.1469-8986.2011.01216.x

[58] Federal Trade Commission. 2020. You Don't Say: An FTC Workshop on Voice Cloning Technologies. https://www.ftc.gov/news-events/events-calendar/you-dont-say-ftc-workshop-voice-cloning-technologies

[59] Ernst Fehr and Urs Fischbacher. 2003. The nature of human altruism. *Nature* 425, 6960 (2003), 785–791.

[60] David R Feinberg, Lisa M DeBruine, Benedict C Jones, and Anthony C Little. 2008. Correlated preferences for men's facial and vocal masculinity. *Evolution and Human Behavior* 29, 4 (2008), 233–241.

[61] David R. Feinberg, Lisa M. Debruine, Benedict C. Jones, and David I. Perrett. 2008. The role of femininity and averageness of voice pitch in aesthetic judgments of women's voices. *Perception* (2008). https://doi.org/10.1068/p5514

[62] D. R. Feinberg, B. C. Jones, M. J. Law Smith, F. R. Moore, L. M. DeBruine, R. E. Cornwell, S. G. Hillier, and D. I. Perrett. 2006. Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice. *Hormones and Behavior* (2006). https://doi.org/10.1016/j.yhbeh.2005.07.004

[63] David R Feinberg, Benedict C Jones, Anthony C Little, D Michael Burt, and David I Perrett. 2005. Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal behaviour* 69, 3 (2005), 561–568.

[64] Susan T. Fiske. 2017. Prejudices in Cultural Contexts: Shared Stereotypes (Gender, Age) Versus Variable Stereotypes (Race, Ethnicity, Religion). *Perspectives on Psychological Science* (2017). https://doi.org/10.1177/1745691617708204

[65] Joseph Fordham, Rabindra Ratan, Kuo-Ting Huang, and Kyle Silva. 2020. Stereotype Threat in a Video Game Context and Its Influence on Perceptions of Science, Technology, Engineering, and Mathematics (STEM): Avatar-Induced Active Self-Concept as a Possible Mitigator. *American Behavioral Scientist* (2020), 0002764220919148.

[66] Jesse Fox, Jeremy Bailenson, and Joseph Binney. 2009. Virtual experiences, physical behaviors: The effect of presence on imitation of an eating avatar. *Presence: Teleoperators and Virtual Environments* 18, 4 (2009), 294–303.

[67] Jesse Fox and Jeremy N. Bailenson. 2009. Virtual Self-Modeling: The Effects of Vicarious Reinforcement and Identification on Exercise Behaviors. *Media Psychology* 12 (2009), 1–25. https://doi.org/10.1080/15213260802669474

[68] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Alexandra Adkins. 2020. My body, my avatar: How people perceive their avatars in social virtual reality. *Conference on Human Factors in Computing Systems - Proceedings* (2020), 1–8. https://doi.org/10.1145/3334480.3382923

[69] Asif A Ghazanfar and Drew Rendall. 2008. Evolution of human vocal production. *Current Biology* 18, 11 (2008), R457—-R460.

[70] Asif A. Ghazanfar, Hjalmar K. Turesson, Joost X. Maier, Ralph van Dinther, Roy D. Patterson, and Nikos K. Logothetis. 2007. Vocal-Tract Resonances as Indexical Cues in Rhesus Monkeys. *Current Biology* (2007). https://doi.org/10.1016/j.cub.2007.01.029

[71] Timo Gnambs, Markus Appel, and Bernad Batinic. 2010. Color red in web-based knowledge testing. *Computers in Human Behavior* 26, 6 (2010), 1625–1631. https://doi.org/10.1016/j.chb.2010.06.010

[72] Google. 2020. Google Assistant. https://assistant.google.com/

[73] Mark Grimshaw. 2007. Sound and immersion in the first-person shooter. In *Proceedings of CGAMES 2007 - 11th International Conference on Computer Games: AI, Animation, Mobile, Educational and Serious Games*.

[74] Mark Grimshaw. 2007. Sound and immersion in the first-person shooter. *Proceedings of CGAMES 2007 - 11th International Conference on Computer Games: AI, Animation, Mobile, Educational and Serious Games* January 2007 (2007), 119–124.

[75] Rosanna E Guadagno, Jim Blascovich, Jeremy N Bailenson, and Cade McCall. 2007. Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology* 10, 1 (2007), 1–22. https://doi.org/10.1080/15213260701300865

[76] Jérôme Guegan, Stéphanie Buisine, Fabrice Mantelet, Nicolas Maranzana, and Frédéric Segonds. 2016. Avatar-mediated creativity: When embodying inventors makes engineers more creative. *Computers in Human Behavior* 61 (2016), 165–175. https://doi.org/10.1016/j.chb.2016.03.024

[77] Elizabeth L. Haines, Kay Deaux, and Nicole Lofaro. 2016. The Times They Are a-Changing … or Are They Not? A Comparison of Gender Stereotypes, 1983-2014. *Psychology of Women Quarterly* (2016). https://doi.org/10.1177/0361684316634081

[78] Andrew F Hayes. 2017. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

[79] Sylvie Hébert, Renée Béland, Odrée Dionne-Fournelle, Martine Crête, and Sonia J. Lupien. 2005. Physiological stress response to video-game playing: The contribution of built-in music. *Life Sciences* 76, 20 (2005), 2371–2380. https://doi.org/10.1016/j.lfs.2004.11.011

[80] Cynthia Hoffner and Martha Buchanan. 2005. Young adults' wishful identification with television characters: The role of perceived similarity and character attributes. https://doi.org/10.1207/S1532785XMEP0704_2

[81] T. C. Holyoke, Eugene S. Morton, and Jake Page. 1992. Animal Talk: Science and the Voices of Nature. *The Antioch Review* (1992). https://doi.org/10.2307/4612642

[82] John J Horton, David G Rand, and Richard J Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* 14, 3 (2011), 399–425.

[83] Bart Hulshof. 2013. The influence of colour and scent on people's mood and cognitive performance in meeting rooms. *Master Thesis* May (2013), 1–97.

[84] W IJsselsteijn, Y De Kort, K Poels, A Jurgelionis, and Francesco Bellotti. 2007. Characterising and Measuring User Experiences in Digital Games. *International Conference on Advances in Computer Entertainment Technology* 620 (2007), 1–4. https://doi.org/10.1007/978-1-60761-580-4

[85] Katherine Isbister and Clifford Nass. 2000. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies* 53 (2000), 251–267. https://doi.org/10.1006/ijhc.2000.0368

[86] Corentin Jemine. 2019. Master's thesis: Real-Time Voice Cloning. (2019). https://matheo.uliege.be/handle/2268.2/6801

[87] Corentin Jemine. 2020. Real-Time Voice Cloning. https://github.com/CorentinJ/Real-Time-Voice-Cloning

[88] Charlene Jennett, Anna L. Cox, Paul Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs, and Alison Walton. 2008. Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies* 66, 9 (sep 2008), 641–661. https://doi.org/10.1016/j.ijhcs.2008.04.004

[89] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems*. arXiv:1806.04558

[90] Colby Johanson and Regan L. Mandryk. 2016. Scaffolding Player Location Awareness through Audio Cues in First-Person Shooters. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (2016), 3450–3461. https://doi.org/10.1145/2858036.2858172

[91] Colby Johanson and Regan L. Mandryk. 2016. Scaffolding Player Location Awareness through Audio Cues in First-Person Shooters. (2016), 3450–3461. https://doi.org/10.1145/2858036.2858172

[92] Benedict C Jones, David R Feinberg, Lisa M DeBruine, Anthony C Little, and Jovana Vukovic. 2008. Integrating cues of social interest and voice pitch in men's preferences for women's voices. *Biology Letters* 4, 2 (2008), 192–194.

[93] Benedict C. Jones, David R. Feinberg, Lisa M. DeBruine, Anthony C. Little, and Jovana Vukovic. 2010. A domain-specific opposite-sex bias in human preferences for manipulated voice pitch. *Animal Behaviour* 79, 1 (2010), 57–62. https://doi.org/10.1016/j.anbehav.2009.10.003

[94] Kristine Jørgensen. 2008. *Left in the dark: playing computer games with the sound turned off*. Ashgate.

[95] Kristine Jørgensen. 2008. Left in the dark: playing computer games with the sound turned off. *From Pac-Man to Pop Music: Interactive Audio in Games and New Media* (2008), 163–176. http://hdl.handle.net/1956/7855

[96] Kristine Jørgensen. 2010. Time for new terminology? Diegetic and non-diegetic sounds in computer games revisited. In *Game Sound Technology and Player Interaction: Concepts and Developments*. https://doi.org/10.4018/978-1-61692-828-5.ch005

[97] Dominic Kao. 2019. JavaStrike: A Java Programming Engine Embedded in Virtual Worlds. In *Proceedings of The Fourteenth International Conference on the Foundations of Digital Games*.

[98] Dominic Kao. 2019. The Effects of Anthropomorphic Avatars vs. Non-Anthropomorphic Avatars in a Jumping Game. In *The Fourteenth International Conference on the Foundations of Digital Games*.

[99] Dominic Kao and D. Fox Harrell. 2016. Exploring the Impact of Avatar Color on Game Experience in Educational Games. *Proceedings of the 34th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI 2016)* (2016).

[100] Dominic Kao and D. Fox Harrell. 2018. The Effects of Badges and Avatar Identification on Play and Making in Educational Games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI'18*.

[101] Changsoo Kim, Sang Gun Lee, and Minchoel Kang. 2012. I became an attractive person in the virtual world: Users' identification with virtual communities and avatars. *Computers in Human Behavior* 28, 5 (2012), 1663–1669. https://doi.org/10.1016/j.chb.2012.04.004

[102] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. Crepe: A Convolutional Representation for Pitch Estimation. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. https://doi.org/10.1109/ICASSP.2018.8461329 arXiv:1802.06182

[103] Yanghee Kim and Amy L. Baylor. 2006. Pedagogical agents as learning companions: The role of agent competency and type of interaction. *Educational Technology Research and Development* 54, 3 (2006), 223–243.

[104] Youjeong Kim and S Shyam Sundar. 2012. Visualizing ideal self vs. actual self through avatars: Impact on preventive health outcomes. *Computers in Human Behavior* 28, 4 (2012), 1356–1364.

[105] Elly A Konijn, Marije Nije Bijvank, and Brad J Bushman. 2007. I wish I were a warrior: the role of wishful identification in the effects of violent video games on aggression in adolescent boys. *Developmental psychology* 43, 4 (2007), 1038.

[106] Jordan Koulouris, Zoe Jeffery, James Best, Eamonn O'Neill, and Christof Lutteroth. 2020. Me vs. Super(wo)man: Effects of Customization and Identification in a VR Exergame. (2020), 1–17. https://doi.org/10.1145/3313831.3376661

[107] Jody Kreiman, Diana Vanlancker-Sidtis, and Bruce R Gerratt. 2003. Defining and Measuring Voice Quality. *VOQUAL'03, Geneva, August 27-29, 2003* (2003).

[108] Christof Kuhbandner and Reinhard Pekrun. 2013. Joint effects of emotion and color on memory. *Emotion (Washington, D.C.)* 13, 3 (2013), 375–9. https://doi.org/10.1037/a0031821

[109] Pontus Larsson, Aleksander Väljamäe, Daniel Västfjäll, Ana Tajadura-Jiménez, and Mendel Kleiner. 2010. Auditory-Induced Presence in Mixed Reality Environments and Related Technology. (2010), 143–163. https://doi.org/10.1007/978-1-84882-733-2_8 arXiv:arXiv:1011.1669v3

[110] Marianne Latinus and Margot J Taylor. 2012. Discriminating male and female voices: differentiating pitch and gender. *Brain topography* 25, 2 (2012), 194–204.

[111] Eun Ju Lee, Clifford Nass, and Scott Brave. 2000. Can computer-generated speech have gender? An experimental test of gender stereotype. *Conference on Human Factors in Computing Systems - Proceedings* (2000), 289–290. https://doi.org/10.1145/633292.633461

[112] Jong Eun Roselyn Lee and Clifford Nass. 2012. Distinctiveness-based stereotype threat and the moderating role of coaction contexts. *Journal of Experimental Social Psychology* (2012). https://doi.org/10.1016/j.jesp.2011.06.018

[113] Jong-Eun Roselyn Lee, Clifford I Nass, and Jeremy N Bailenson. 2014. Does the mask govern the mind?: Effects of arbitrary gender representation on quantitative task performance in avatar-represented virtual groups. *Cyberpsychology, Behavior, and Social Networking* 17, 4 (2014), 248–254.

[114] Sanguk Lee, Rabindra Ratan, and Taiwoo Park. 2019. The voice makes the car: Enhancing autonomous vehicle perceptions and adoption intention through voice agent gender and style. *Multimodal Technologies and Interaction* (2019). https://doi.org/10.3390/mti3010020

[115] Benjamin J. Li and May O. Lwin. 2016. Player see, player do: Testing an exergame motivation model based on the influence of the self avatar. *Computers in Human Behavior* 59 (2016), 350–357. https://doi.org/10.1016/j.chb.2016.02.034

[116] Benjamin J. Li, May O. Lwin, and Younbo Jung. 2014. Wii, Myself, and Size: The Influence of Proteus Effect and Stereotype Threat on Overweight Children's Exercise Motivation and Behavior in Exergames. *Games for Health Journal* (2014). https://doi.org/10.1089/g4h.2013.0081

[117] Mats Liljedahl. 2011. Sound for Fantasy and Freedom. *Game Sound Technology and Player Interaction* (2011), 264–285. https://doi.org/10.4018/978-1-61692-828-5.ch017

[118] Limelight Networks. 2021. State of Online Gaming 2021. (2021). https://www.limelight.com/lp/state-of-online-gaming-2021/

[119] Conor Linehan, George Bellord, Ben Kirman, Zachary H. Morford, and Bryan Roche. 2014. Learning curves: Analysing pace and challenge in four successful puzzle games. *CHI PLAY 2014 - Proceedings of the 2014 Annual Symposium on Computer-Human Interaction in Play* (2014), 181–190. https://doi.org/10.1145/2658537.2658695

[120] LingoJam. 2020. Robot Voice Generator. https://lingojam.com/RobotVoiceGenerator

[121] Van Looy and De Grove. 2013. Avatar identification in serious games - The role of avatar identification in the learning experience of a serious game. *Proceeding of: The Power of Play : Motivational Uses and Applications. Pre-Conference to the 63rd International Communication Association (ICA) Annual Conference, Abstracts* (2013).

[122] Thomas Lumley, Paula Diehr, Scott Emerson, and Lu Chen. 2002. The importance of the normality assumption in large public health data sets. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

[123] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (2012), 1–23. https://doi.org/10.3758/s13428-011-0124-6 arXiv:/ssrn.com/abstract=1691163 [http:]

[124] Oscar Mayor, Jordi Bonada, and Jordi Janer. 2009. Kaleivoicecope: Voice transformation from interactive installations to video-games. In *Proceedings of the AES International Conference*.
[125] Edward McAuley, Terry Duncan, and Vance V Tammen. 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport* 60, 1 (1989), 48–58.
[126] Daniel G McDonald and Hyeok Kim. 2001. When I die, I feel small: Electronic game characters and the social self. *Journal of Broadcasting & Electronic Media* 45, 2 (2001), 241–258.
[127] Ravi Mehta and Rui(Juliet) Zhu. 2008. Blue or Red? Exploring the Effect of Color on Cognitive Task Performances. *Science* 323, February (2008), 1226–1229. https://doi.org/10.1126/science.1169144
[128] M.A. Meier, Russell A. Hill, Andrew J. Elliot, and R.A. Barton. 2015. Color in Achievement Contexts in Humans. *Handbook of Color Psychology* 44, February (2015), 0–103. https://doi.org/10.1063/1.2756072 arXiv:arXiv:0811.2183v2
[129] Microsoft. 2020. Cortana. https://www.microsoft.com/en-us/cortana
[130] Microsoft Game Studios and Electronic Arts. 2007. Mass Effect. Game [Multiple Platforms].
[131] Jason P. Mitchell, C. Neil Macrae, and Mahzarin R. Banaji. 2006. Dissociable Medial Prefrontal Contributions to Judgments of Similar and Dissimilar Others. *Neuron* 50, 4 (2006), 655–663. https://doi.org/10.1016/j.neuron.2006.03.040
[132] Dean Mobbs, Rongjun Yu, Marcel Meyer, Luca Passamonti, Ben Seymour, Andrew J Calder, Susanne Schweizer, Chris D Frith, and Tim Dalgleish. 2009. A key role for similarity in vicarious reward. *Science* 324, 5929 (2009), 900. https://doi.org/10.1126/science.1170539
[133] Masahiro Mori. 1970. The uncanny valley. *Energy* 7, 4 (1970), 33–35. https://doi.org/10.1109/MRA.2012.2192811
[134] Eugene S. Morton. 1977. On the Occurrence and Significance of Motivation-Structural Rules in Some Bird and Mammal Sounds. *The American Naturalist* (1977). https://doi.org/10.1086/283219
[135] Lennart E. Nacke and Mark Grimshaw. 2011. Player-Game Interaction Through Affective Sound. *Game Sound Technology and Player Interaction* (2011), 264–285. https://doi.org/10.4018/978-1-61692-828-5.ch013
[136] Myura Nagendran, Kurinchi Selvan Gurusamy, Rajesh Aggarwal, Marilena Loizidou, and Brian R. Davidson. 2013. Virtual reality training for surgical trainees in laparoscopic surgery. https://doi.org/10.1002/14651858.CD006575.pub3
[137] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied* 7, 3 (2001), 171–181. https://doi.org/10.1037/1076-898X.7.3.171
[138] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology* 27, 10 (1997), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x
[139] Raymond Ng and Robb Lindgren. 2013. Examining the effects of avatar customization and narrative on engagement and learning in video games. *Proceedings of CGAMES 2013 USA - 18th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational and Serious Games* (2013), 87–90. https://doi.org/10.1109/CGames.2013.6632611
[140] Rolf Nordahl. 2005. Self-induced Footsteps Sounds in Virtual Reality: Latency, Recognition, Quality and Presence. (2005).
[141] Rolf Nordahl. 2006. Increasing the Motion of Users in Photo-realistic Virtual Environments by Utilising Auditory Rendering of the Environment and Ego-motion. *Presence 2006* (2006), 57–62.
[142] Rolf Nordahl and Niels C Nilsson. 2014. The sound of being there. In *The Oxford handbook of interactive audio*.
[143] Keith Oatley. 1995. A taxonomy of the emotions of literary response and a theory of identification in fictional narrative. *Poetics* (1995). https://doi.org/10.1016/0304-422X(94)P4296-S
[144] Takashi Oguchi and Hiroto Kikuchi. 1997. Voice and interpersonal attraction. *Japanese Psychological Research* 39, 1 (1997), 56–61.
[145] Yumiko Ohara. 1999. Performing gender through voice pitch: A cross-cultural analysis of Japanese and American English. In *Wahrnehmung und Herstellung von Geschlecht*. Springer, 105–116.
[146] Justin H Park and Mark Schaller. 2005. Does attitude similarity serve as a heuristic cue for kinship? Evidence of an implicit cognitive association. *Evolution and Human Behavior* 26, 2 (2005), 158–170.
[147] Jim R Parker and John Heerema. 2008. Audio interaction in computer mediated games. *International Journal of Computer Games Technology* 2008 (2008).
[148] Pearl Abyss. 2015. Black Desert Online. Game [Multiple Platforms].
[149] Tabitha C. Peck, Sofia Seinfeld, Salvatore M. Aglioti, and Mel Slater. 2013. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and Cognition* (2013). https://doi.org/10.1016/j.concog.2013.04.016
[150] Jorge Peña, Subuhi Khan, and Cassandra Alexopoulos. 2016. I Am What I See: How Avatar and Opponent Agent Body Size Affects Physical Activity Among Men Playing Exergames. *Journal of Computer-Mediated Communication* (2016). https://doi.org/10.1111/jcc4.12151

[151] Jorge Peña and Eunice Kim. 2014. Increasing exergame physical activity through self and opponent avatar appearance. *Computers in Human Behavior* (2014). https://doi.org/10.1016/j.chb.2014.09.038
[152] Cyril R Pernet and Pascal Belin. 2012. The role of pitch and timbre in voice gender categorization. *Frontiers in psychology* 3 (2012), 23.
[153] Jean A. Pratt, Karina Hauser, Zsolt Ugray, and Olga Patterson. 2007. Looking at human-computer interface design: Effects of ethnicity in computer agents. *Interacting with Computers* 19, 4 (2007), 512–523.
[154] David Andrew Puts. 2005. Mating context and menstrual phase affect women's preferences for male voice pitch. *Evolution and Human Behavior* 26, 5 (2005), 388–397.
[155] David Andrew Puts, Steven J.C. Gaulin, and Katherine Verdolini. 2006. Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior* (2006). https://doi.org/10.1016/j.evolhumbehav.2005.11.003
[156] Lingyun Qiu and Izak Benbasat. 2005. An investigation into the effects of text-to-speech voice and 3D avatars on the perception of presence and flow of Live Help in electronic commerce. *ACM Transactions on Computer-Human Interaction* 12, 4 (2005), 329–355. https://doi.org/10.1145/1121112.1121113
[157] Lingyun Qiu and Izak Benbasat. 2005. Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars. *International Journal of Human-Computer Interaction* 19, 1 (2005), 75–94. https://doi.org/10.1207/s15327590ijhc1901_6
[158] Rabindra Ratan, David Beyea, Benjamin J. Li, and Luis Graciano. 2019. Avatar characteristics induce users' behavioral conformity with small-to-medium effect sizes: a meta-analysis of the proteus effect. *Media Psychology* 0, 0 (2019), 1–25. https://doi.org/10.1080/15213269.2019.1623698
[159] Rabindra Ratan and Young June Sah. 2015. Leveling up on stereotype threat: The role of avatar customization and avatar embodiment. *Computers in Human Behavior* 50 (2015), 367–374. https://doi.org/10.1016/j.chb.2015.04.010
[160] David Reby, Karen McComb, Bruno Cargnelutti, Chris Darwin, W. Tecumseh Fitch, and Tim Clutton-Brock. 2005. Red deer stags use formants as assessment cues during intrasexual agonistic interactions. *Proceedings of the Royal Society B: Biological Sciences* (2005). https://doi.org/10.1098/rspb.2004.2954
[161] James Robb, Tom Garner, Karen Collins, and Lennart E. Nacke. 2017. The Impact of Health-Related User Interface Sounds on Player Experience. *Simulation and Gaming* (2017). https://doi.org/10.1177/1046878116688236
[162] Rockstar Games. 2018. Red Dead Redemption 2. Game [Multiple Platforms].
[163] Katja Rogers, Matthias Jörg, and Michael Weber. 2019. Effects of background music on risk-taking and general player experience. *CHI PLAY 2019 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (2019), 213–224. https://doi.org/10.1145/3311350.3347158
[164] Katja Rogers, Giovanni Ribeiro, Rina R. Wehbe, Michael Weber, and Lennart E. Nacke. 2018. Vanishing Importance. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (2018), 1–13. https://doi.org/10.1145/3173574.3173902
[165] Rinat B. Rosenberg-Kima, E. Ashby Plant, Celestee E. Doerr, and Amy Baylor. 2010. The influence of computer-based model's race and gender on female students' attitudes and beliefs towards engineering. *Journal of Engineering Education* (2010), 35–44. https://doi.org/10.1002/j.2168-9830.2010.tb01040.x
[166] Joel Ross, Lilly Irani, M Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *CHI'10 extended abstracts on Human factors in computing systems*. 2863–2872.
[167] R M Ryan and E L Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American psychologist* 55, 1 (2000), 68–78. https://doi.org/10.1037/0003-066X.55.1.68 arXiv:0208024 [gr-qc]
[168] Richard M. Ryan, C. Scott Rigby, and Andrew Przybylski. 2006. The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion* 30, 4 (2006), 344–360. https://doi.org/10.1007/s11031-006-9051-8
[169] Young June Sah, Rabindra Ratan, Hsin-Yi Sandy Tsai, Wei Peng, and Issidoros Sarinopoulos. 2017. Are you what your avatar eats? Health-behavior effects of avatar-manifested self-concept. *Media Psychology* 20, 4 (2017), 632–657.
[170] Timothy Sanders and Paul Cairns. 2010. Time perception, immersion and music in videogames. In *Proceedings ofthe 24th BCS interaction specialist group conference*.
[171] Edward F. Schneider, Annie Lang, Mija Shin, and Samuel D. Bradley. 2004. Death with a story: How story impacts emotional, motivational, and physiological responses to first-person shooter video games. *Human Communication Research* (2004). https://doi.org/10.1093/hcr/30.3.361
[172] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2017. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *arXiv* (2017), 4779–4783.

[173] Janet Siegmund, Christian Kästner, Jörg Liebig, Sven Apel, and Stefan Hanenberg. 2014. Measuring and modeling programming experience. *Empirical Software Engineering* 19, 5 (2014), 1299–1334. https://doi.org/10.1007/s10664-013-9286-4
[174] David Smahel, Lukas Blinka, and Ondrej Ledabyl. 2008. Playing MMORPGs: connections between addiction and identifying with a character. *Cyberpsychology & Behavior* 11, 6 (2008), 715–718. https://doi.org/10.1089/cpb.2007.0210
[175] Alistair Raymond Bryce Soutter and Michael Hitchens. 2016. The relationship between character identification and flow state within video games. *Computers in Human Behavior* 55, December 2015 (2016), 1030–1038. https://doi.org/10.1016/j.chb.2015.11.012
[176] Square Enix. 2013. Final Fantasy XIV. Game [Multiple Platforms].
[177] Sharon T. Steinemann, Elisa D. Mekler, and Klaus Opwis. 2015. Increasing Donating Behavior Through a Game for Change. https://doi.org/10.1145/2793107.2793125
[178] Ching-I Teng. 2017. Impact of avatar identification on online gamer loyalty: Perspectives of social identity and social capital theories. *International Journal of Information Management* 37, 6 (2017), 601–610. https://doi.org/10.1016/j.ijinfomgt.2017.06.006
[179] Larry Terango. 1966. Pitch and Duration Characteristics of the Oral Reading of Males on a Masculinity-Femininity Dimension. *Journal of Speech and Hearing Research* (1966). https://doi.org/10.1044/jshr.0904.590
[180] Sabine Trepte and Leonard Reinecke. 2010. Avatar creation and video game enjoyment. *Journal of Media Psychology* (2010).
[181] Sabine Trepte, Leonard Reinecke, and Katharina-Maria Behr. 2010. Avatar Creation and Video Game Enjoyment: Effects of Life-Satisfaction, Game Competitiveness and Identification with the Avatar. In *60th Annual Conference of the International Communication Association (ICA)*. https://doi.org/10.1027/1864-1105/a000022
[182] Selen Turkay and Charles K. Kinzer. 2017. *The Relationship between Avatar-Based Customization, Player Identification, and Motivation*. 48–79 pages. https://doi.org/10.4018/978-1-5225-1817-4.ch003
[183] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. (2016), 1–15. arXiv:1609.03499 http://arxiv.org/abs/1609.03499
[184] Wim A Van Dommelen and Bente H Moxness. 1995. Acoustic parameters in speaker height and weight identification: sex-specific behaviour. *Language and speech* 38, 3 (1995), 267–287.
[185] Jan Van Looy, Cédric Courtois, Melanie De Vocht, and Lieven De Marez. 2012. Player Identification in Online Games: Validation of a Scale for Measuring Identification in MMOGs. *Media Psychology* 15, 2 (2012), 197–221. https://doi.org/10.1080/15213269.2012.674917
[186] Kellie Vella, Madison Klarkowski, Selen Turkay, and Daniel Johnson. 2020. Making friends in online games: gender differences and designing for greater social connectedness. *Behaviour and Information Technology* (2020). https://doi.org/10.1080/0144929X.2019.1625442
[187] Voicemod. 2020. Voicemod. https://www.voicemod.net/
[188] Volition and Deep Silver. 2013. Saints Row IV. Game [Multiple Platforms].
[189] Jovana Vukovic, David R Feinberg, Benedict C Jones, Lisa M DeBruine, Lisa L M Welling, Anthony C Little, and Finlay G Smith. 2008. Self-rated attractiveness predicts individual differences in women's preferences for masculine men's voices. *Personality and Individual Differences* 45, 6 (2008), 451–456.
[190] T Franklin Waddell, S Shyam Sundar, and Joshua Auriemma. 2015. Can customizing an avatar motivate exercise intentions and health behaviors among those with low health ideals? *Cyberpsychology, Behavior, and Social Networking* 18, 11 (2015), 687–690.
[191] Greg Wadley, Marcus Carter, and Martin Gibbs. 2015. Voice in virtual worlds: The design, use, and influence of voice chat in online play. *Human–Computer Interaction* 30, 3-4 (2015), 336–365.
[192] Greg Wadley, Martin Gibbs, and Peter Benda. 2007. Speaking in character: using voice-over-IP to communicate within MMORPGs. In *Proceedings of the 4th Australasian conference on Interactive entertainment*. 1–8.
[193] Greg Wadley, Martin R. Gibbs, and Nicolas Ducheneaut. 2009. You can be too rich: Mediated communication in a virtual world. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group - Design: Open 24/7, OZCHI '09*. https://doi.org/10.1145/1738826.1738835
[194] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4879–4883.
[195] Melissa Watts. 2016. *Avatar Self-Identification, Self-Esteem, and Perceived Social Capital in the Real World: A Study of World of Warcraft Players and their Avatars*. University of South Florida.
[196] Helen Wauck, Gale Lucas, Ari Shapiro, Andrew Feng, Jill Boberg, and Jonathan Gratch. 2018. Analyzing the effect of avatar self-similarity on men and women in a search and rescue game. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018). https://doi.org/10.1145/3173574.3174059

[197] Alexander Wharton and Karen Collins. 2011. Subjective measures of the influence of music customization on the video game play experience: A pilot study. *Game Studies* (2011).
[198] Dmitri Williams, Scott Caplan, and Li Xiong. 2007. Can you hear me now? The impact of voice in an online gaming community. *Human Communication Research* (2007). https://doi.org/10.1111/j.1468-2958.2007.00306.x
[199] Hanna Elina Wirman and Rhys Jones. 2017. *Voice and Sound: Player Contributions to Speech.* Peter Lang, Digital Formations Series.
[200] Nick Yee and J Bailenson. 2007. The Proteus Effect: The Effect of Transformed Self-Representation on Behavior. *Human communication research* (2007), 1–38. http://onlinelibrary.wiley.com/doi/10.1111/j.1468-2958.2007.00299.x/full
[201] Nick Yee, Jeremy N Bailenson, and Nicolas Ducheneaut. 2009. *The Proteus Effect Implications of Transformed Digital Self-Representation on Online and Offline Behavior*. Vol. 36. 285–312 pages. https://doi.org/10.1177/0093650208330254
[202] Nick Yee, Jeremy N. Bailenson, Mark Urbanek, Francis Chang, and Dan Merget. 2007. The unbearable likeness of being digital: The persistence of nonverbal social norms in online virtual environments. *Cyberpsychology and Behavior* 10, 1 (2007), 115–121. https://doi.org/10.1089/cpb.2006.9984
[203] Nick Yee, Nicolas Ducheneaut, Mike Yao, and Les Nelson. 2011. Do Men Heal More When in Drag? *CHI 2011* (2011), 1–4.
[204] Sukkyung You, Euikyung Kim, and Donguk Lee. 2017. Virtually Real: Exploring Avatar Identification in Game Addiction among Massively Multiplayer Online Role-Playing Games (MMORPG) Players Sukkyung. *Games and Culture* (2017). https://doi.org/10.1177/1555412015581087
[205] David Zendle, Paul Cairns, and Daniel Kudenko. 2015. Higher graphical fidelity decreases players' access to aggressive concepts in violent video games. In *CHI PLAY 2015 - Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. https://doi.org/10.1145/2793107.2793113
[206] Dolf Zillmann. 1995. Mechanisms of emotional involvement with drama. *Poetics* (1995). https://doi.org/10.1016/0304-422X(94)00020-7
[207] Miron Zuckerman and Kunitate Miyake. 1993. The attractive voice: What makes it so? *Journal of nonverbal behavior* 17, 2 (1993), 119–135.

Received February 2021; revised June 2021; accepted July 2021