A novel experimental design for the study of listener-to-listener convergence in phoneme categorization

Qingye Shen, Leonardo Lancia, Noël Nguyen

Interspeech 2024
1-5 September 2024, Kos, Greece

A novel experimental design for the study of listener-to-listener convergence in phoneme categorization

Qingye Shen[1,2], *Leonardo Lancia*[1,2], *Noël Nguyen*[1,2]

[1]Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
[2]Institute for Language, Communication and the Brain, Aix Marseille University, France
qingye.shen@univ-amu.fr, leonardo.lancia.1@univ-amu.fr, noel.nguyen-trong@univ-amu.fr

Abstract

We present a novel experimental design that combines highly accurate psychometric methods with an interactive task to characterize how two or more listeners can converge towards each other in the categorization of speech sounds. The design is implemented as a cooperative game, in which listeners are presented with a sequence of sounds that range on a continuum between two endpoints unambiguously associated with two phoneme categories in a joint phoneme identification task. To play the game successfully, listeners must comply with both a distinctivity constraint (identify the endpoints as being different from each other) and an agreement constraint (identify the stimuli in the same way as their partner). Our first results show that our experimental design opens new avenues for research on convergence between listeners in speech perception.
Index Terms: speech perception, phoneme categorization, perceptual convergence, serious game

1. Introduction

In spoken language interactions, and for people to understand each other, speech sounds must be categorized consistently across listeners. Within a linguistic community, a common set of criteria must therefore be agreed upon, as regards how speech sound categories are delineated in the perceptual space. We still need to understand the cognitive and cerebral mechanisms that allow this shared perceptual landscape to emerge. Our long-term goal is to identify the cognitive and cerebral mechanisms that make it possible for listeners to build up a shared perceptual landscape in the processing of speech.

Work on inter-individual convergence in the *production* of speech already has a rich and long history, ever since [1] and [2]'s pioneer studies (see [3] for a review). These studies have endeavored to identify the mechanisms that lead a speaker to imitate another person's way of producing speech sounds when exposed to that person's speech. In the speech perception domain, there is likewise a large body of literature on the brain and cognitive underpinnings of perceptual adaptation in listeners, whether adaptation takes place as a function of the speaker's idiosyncratic characteristics (e.g., [4, 5]), lexical context ([6]) or phonetic context ([7]). However, while phonetic convergence studies have centered on speaker-to-speaker adaptation, and perceptual adaptation studies have been up to now mainly concerned with listener-to-speaker adaptation, our focus of interest is different, and lies in *listener-to-listener* adaptation. The question we ask is whether a listener can infer how *another listener* categorizes the speech sounds that both listeners are exposed to. In other words, to what extent can we ascertain whether other listeners perceive speech in the way we do?

Although this appears to be an essential prerequisite for speech to successfully function as a communication device, and to the best of our knowledge, little attention has been devoted so far to that issue. In our view, this is due to two main reasons. First, at the theoretical level, previous studies have focused on the single listener, i.e., how individuals perceive speech sounds in situations of short, artificial social isolation. Second, at the methodological level, technical obstacles may have prevented researchers from examining the impact of social interactions between listeners on speech perception.

One may argue that studies on phonetic convergence already have shed light on listener-to-listener adaptation, because it may be assumed that convergence in production necessarily entails convergence in perception. However, previous work has shown that this assumption is not necessarily true. In [8] for example, participants were exposed to words spoken by a speaker of another regional accent, whose phonemic inventory differed from that of the participants' native accent, in both a shadowing task and a word-identification task. The participants phonetically converged towards the speaker in shadowing, but showed no perceptual adaptation to the speaker in the word identification task.

In one pilot study [9] and a large-scale one [10], we endeavored to explore potential between-listener convergence effects in the identification of phonemes using a novel experimental design. In this design, participants are presented with a set of stimuli at equidistant locations on an acoustic continuum between two endpoints unambiguously associated with two phoneme categories (such as /s/-/ʃ/ or /b/-p/). These stimuli are presented a number of times in a random order and, for each stimulus, participants must identify the stimulus as either of the two proposed categories. By contrast with a standard phoneme identification task, carried out individually, participants are asked in this novel design to perform the task along with one or several partners.

On hearing a stimulus, participants first have to each provide a response to that stimulus. Once all of the responses are collected, the response(s) given by her/his partner(s) is shown on each participant's screen. Participants are instructed to seek to respond to each stimulus in the same way as their partner. The goal is to induce participants to predict, on being presented with a stimulus, how their partner will categorize that stimulus and to respond identically. As the experiment unfolds, the design makes it possible to determine to what extent participants come to align with each other in how they identify the speech sounds. In the abovementioned studies, participants did the task in the lab and in the presence of each other [9] or online [10], and with either a human partner [9] or, unbeknownst to the participants, an artificial one [10].

10.21437/Interspeech.2024-1598

2. New experimental design

The experimental design described above has a number of limitations. First, the goal of the game (respond in the same way as your partner) may be seen by participants as being of limited meaningfulness, and it is only remotely connected to the objectives that two people may seek in perceptually converging towards each other, such as the improvement of the coordination between them in the accomplishment of a joint action. Second, the feedback that participants receive about their partners' response forms the final element of each trial, and no further event occurs that would depend on whether participants agree or not with each other in the categorization of the stimulus. It is therefore uncertain to what extent participants really pay attention to that feedback. Third, the design opens the possibility for participants to adopt a strategy that consists in always both opting for the same response (/sa/ for example) irrespective of the stimulus presented, since this allows them to comply with the rule of the game at a minimal cognitive cost. That strategy appears to have been adopted by a significant number of dyads in [11].

We therefore have developed a new variant of that experimental design whose aim was to overcome those limitations and, in addition, to enhance the participants' engagement and interest in the task. This variant takes the form of a video game entitled "Let's save the alien together!" and played by two participants. In that game, an alien has landed on Earth, and is chased by a pursuer that throws scoops of icecream at the alien. Both the alien and its pursuer can only move horizontally on the screen, to the left or right. Verbal instructions for movement are sent to the alien by its mother ship, which is said to be located in the air above the scene. However, these instructions cannot be conveyed directly to the alien, and have to be transmitted through the two participants. Furthermore, both participants have to interpret each instruction in the same way for it to get to the alien. The verbal instructions are formed by CV syllables whose initial consonants belong to either of two phonemic categories and differ by one feature, such as /sa/ and /ʃa/.

At the beginning of the game, participants are told that one of the two syllables (e.g., /sa/) means "go to the left" and the other (/ʃa/) "go to the right". The participants are both equipped with headphones and a button box that has two buttons on the left and right. In a first, training phrase, the participants are presented with unambiguous auditory tokens of the two syllables and for each token, are instructed to press the button that is said to be associated with it. In the following, test phase, a sequence of auditory stimuli is played to the participants, who both have to identify them using their respective button boxes. The stimuli are equally spaced on an acoustic continuum between and including the two, unambiguous ones (the endpoints, hereafter) presented in the training phase. On each trial, the alien can move in the indicated direction if either of the two following conditions is fulfilled: 1) the stimulus was one of the two endpoints and both participants identify it correctly, i.e., in accordance with the instructions given to them in the training phase, 2) the stimulus is neither of the two endpoints and both participants identify it in the same way. This allows us to ensure that participants comply with both a distinctivity constraint (Condition 1, differentiate the two endpoints from each other) and an agreement constraint (Condition 2, agree on how stimuli that are ambiguous to various extents should be categorized). As a result, the same-response strategy should be avoided by participants since it is at variance with the distinctivity constraint. On each trial, the alien escapes its pursuer when moving to either the left or right (Condition 1 or 2 fulfilled), and is hit by the scoops of icecream if it stays still (none of the conditions fulfilled).

This amounts to making the two participants a transmission relay in a chain of communication whose the alien is the final receptor. The participants are led to understand that the transmission channel is noisy and that some of the verbal instructions are to some degree unclear. Their goal is to determine how to identify these instructions in the same way. To do this, they must come to an agreement on where the categorical boundary between the initial consonants should be located in the acoustic space.

The stimuli are presented to the participants a number of times in a fully randomized order. This is consistent with a strategy that one may assume to be the best one for the alien in order to escape its pursuer, and which consists in moving in the most unpredictable way. Randomization is required to minimize potential order effects in the participants' responses to the sequence of stimuli, but it therefore also makes full sense within the game's scenario.

3. Implementation of the game

We have implemented the game in Python, using the pygame (pygame.org) library, a free set of Python modules designed for developing real-time video games. Our pygame script for running the game is freely available at `osf.io/8mukx`.

The experimental setup is shown in Figure 1. The two participants are seated side by side in front of a large screen on which the video game is displayed. They are separated by a wood board that prevents them from seeing each other's hands. The game is run from a computer operated by the experimenter and connected to both the screen and the participants' button boxes. Because a participant cannot see their partner's hands on the button box, the only information available to them about their partner's response to a stimulus is that conveyed through the alien and its movements on the screen.

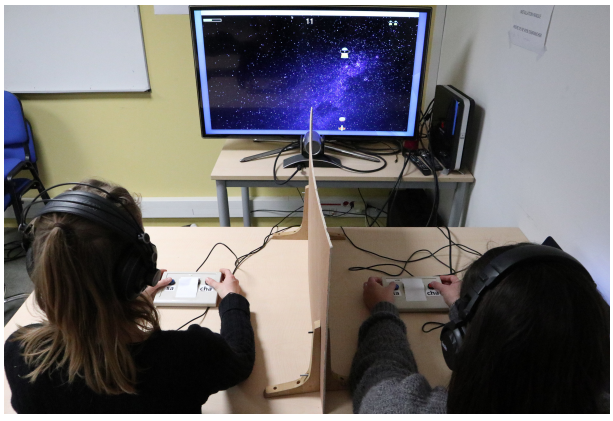

Figure 1: *New experimental setup*

4. A first test

In a first experiment using the game, we recruited 40 pairs of same-sex participants (33 female-female and 7 male-male dyads, aged from 18-36 y.o., mean age 21 y.o.), all native speakers of French with no known hearing defects. Nine acoustic stimuli were used which ranged at equal distances between and

including two endpoints associated with /sa/ and /ʃa/ respectively. These stimuli were generated from two natural recordings of /sa/ and /ʃa/ by a male speaker of French using a procedure adapted from [12]. The participants first passed an individual phoneme identification test in which they were presented with 10 repetitions of each of the 9 stimuli in a fully randomized order and had to categorize each stimulus as /sa/ or /ʃa/. We analyzed the participants' responses by means of a Bayesian logistic regression using the `brms` R package [13], whose main results are shown in Figure 2. The dark orange curve represents the probability of /sa/ responses over the /sa/-/ʃa/ continuum across all participants, as averaged from 4000 samples of the posterior distributions for the parameters of the logistic regression model, whereas the light orange ribbon represents the highest density interval on either side of the average curve, with a credible mass of 95%. The vertical dashed line corresponds to the location of the /sa/-/ʃa/ categorical boundary across participants.

As expected, Stimulus 1 was perceived as an unambiguous token of /sa/ and Stimulus 9 as a close to unambiguous token of /ʃa/. Figure 2 also indicates that the participants' responses shifted from /sa/ to /ʃa/ in a gradual way across the 9 stimuli, and that there was a substantial amount of inter-individual variability in how participants categorized Stimuli 4 to 8. This corresponds to the response pattern we sought to get in designing our acoustic material, as initial inter-individual differences were required for us to be able to subsequently capture potential convergence effects between partners in the game.

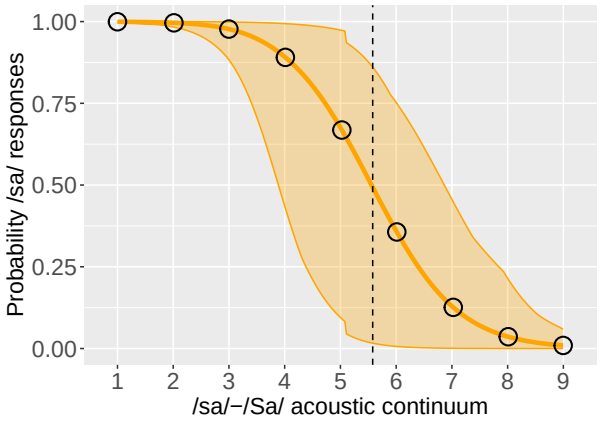

Figure 2: *Probability of* /sa/ *responses as a function of the location of the stimulus on the* /sa/-/ʃa/ *continuum as estimated using a Bayesian logistic regression model.*

The game itself was divided in 6 rounds. In each round, Stimuli 1-2 and 8-9 were presented 4 times each while Stimuli 3 to 7 were presented 7 times each, in a fully randomized order. We chose to have a higher number of presentations for the stimuli closer to the continuum midpoint because these are more ambiguous, and are therefore likely to lend themselves to a higher degree than the peripheral stimuli to mutual perceptual adaptation between partners.

To characterize convergence, we computed a) the location of the /sa/-/ʃa/ categorical boundary by means of a logistic regression for each participant across all of the trials, and b) the distance in that location (referred to as delta hereafter) between the two participants in each of the 40 dyads. We then built up 1000 sets of 40 pseudo dyads, by randomly assigning each participant to another participant than her/his true partner and computed the value of delta for each of the pseudo dyads. To capture potential convergence effects between participants in how the sounds were categorized, we compared the mean value of delta in the real vs. pseudo dyads.[1] Our reasoning was that convergence should be associated with a lower mean delta in the real relative to the pseudo dyads (see [14] for a similar approach in a different domain). The results are displayed in Figure 3.

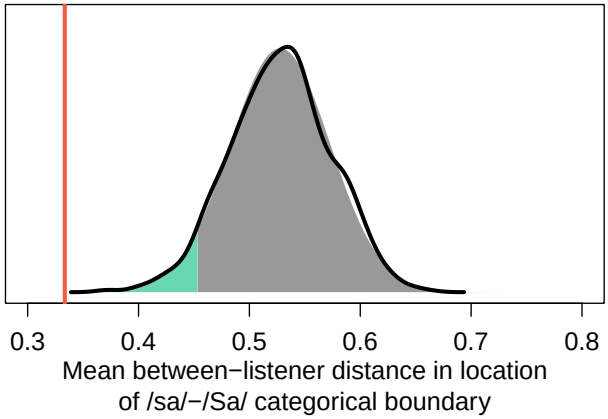

Figure 3: *Normal distribution fitted to the kernel density of the mean delta value for each of 1000 sets of 40 pseudo dyads. Red vertical line: mean delta value for the 40 real dyads. Green area: 5th percentile of the normal distribution.*

The mean delta value for the real dyads was 0.333, and is clearly below the range of values covered by the mean deltas for the pseudo dyads. The probability of the real-dyad delta relative to the normal distribution of delta for the pseudo dyads is lower than 0.001. Note that the delta measure is relative to the size of the interval between two adjacent stimuli on the /sa/-/ʃa/ acoustic continuum. Thus, a delta of 1 means that the categorical boundary locations for the two participants within each dyad are on average one step apart. On average, the between-partner distance in the categorical boundary location in real dyads was about a third of a step, i.e., very small. This indicates that in real dyads, the two partners succeeded in adapting to each other's response pattern very closely.

5. Discussion

Our first results show that our new experimental design is well suited to making perceptual convergence effects arise between listeners in a joint phoneme categorization task. We found that perceptual convergence occurs within a short time frame, as the interactive task lasted less than 10 minutes, and in dyads of listeners exposed to a limited number of stimuli (< 310). In the interactive task, listeners were found to set the /sa/-/ʃa/ categorical boundary at a location close to that of their partner, and this resulted in their categorizing the stimuli in a way which mirrored their partner's own response pattern. Thus, our experimental design allows us to observe in the laboratory how listeners build up a shared perceptual space.

[1] Note that what we compare here are mean values associated with each set of dyads, which is why real dyads are represented by a single value.

5.1. Distinguishing convergence from common fate

In studies on inter-individual convergence, whether in the production or perception of speech, one major issue is to disentangle genuine convergence from common fate. Common fate refers to the fact that two or more people tend to behave in an increasingly similar way, not because they seek to be more like each other, but because they are all subject to the influence of the same factors [15]. In a phoneme identification task, it may be the case that, as they hear the stimuli repeatedly, participants come to have an internal measure of the interval over which the stimuli take place in the acoustic space, and shift the location of the categorical boundary between the two phoneme categories towards the center of that interval as the experiment unfolds. In that scenario, participants' responses would be more similar to those of their partner at the end of the experiment compared with the beginning, not because they sought to perceptually adapt to their partner, but because both of them independently moved their categorical boundary in the same direction. To address this issue, we used a method that draws on [16]'s work on vocal accommodation between speakers. This method entails assessing to what extent between-participant distances in the location of the categorical boundary are shorter in real dyads compared with pseudo dyads. We found that participants are indeed closer on average to their real partner in the game, than to another participant with whom they did not actually interact.

5.2. Perspectives for future studies

Our experimental design opens new avenues for research on convergence between listeners in speech perception. One of these avenues relates to the development of adaptive artificial agents that may be assigned as partners to participants in the game. Using artificial partners allows us to accurately control the feedback provided to the participant and characterize how the participant responds as a function of the feedback's timing and content. Endowing artificial agents with the capacity to adapt to the participants' response pattern may contribute to making the interaction with participants more human-like and to reinforcing the tendency that participants may themselves show to converge towards the artificial agent. In [10]'s recent study on perceptual convergence, artificial agents were already used but did not have an adaptive mechanism and this, according to the authors, may have caused the participants to converge towards their artificial partner to a lesser extent than they would have done had adaptation been reciprocal.

As a first step in that direction, we undertook to model the responses provided by each participant in our interactive task by means of an iterative Bayesian logistic regression classifier. Each model included prior, normal distributions for both the intercept and slope parameters of the logistic regression. We presented the model with the participant's responses to the sequence of stimuli in the interactive task, and updated the prior distributions for the model's parameter after each stimulus+response pair in an iterative manner. An example of the models' output is given in Figure 4. The figure shows the location of each participant's /sa/-/ʃa/ categorical boundary as estimated from the prior distributions of the model's parameters at each trial. We take here the logistic regression classifiers as computational devices to model both the participants' prior stimulus-to-phoneme mapping, and the learning mechanism that allows participants to update this mapping as a function of their partner's own responses. As can be seen, the two trajectories clearly converge towards each other as the game unfolds. A variant of our experimental design will therefore involve participants doing the game with a human confederate, whose responses will actually be generated by a Bayesian artificial agent, with a view to determining to what extent the participant's perceptual convergence towards the agent is sensitive to the agent's own degree of adaptation to the participant.

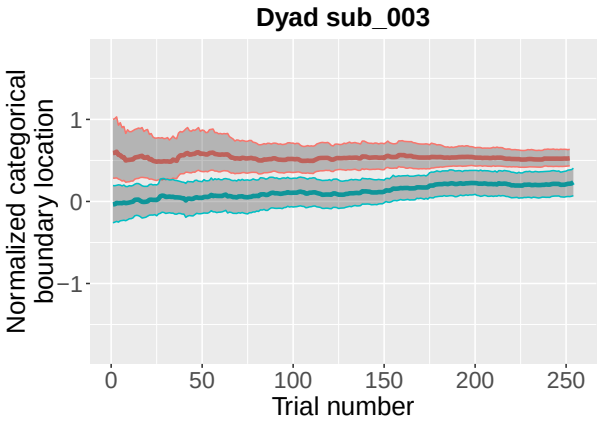

Figure 4: *Normalized location of* /sa/-/ʃa/ *boundary as estimated from the prior distributions of the Bayesian classifier's parameters at each trial in Dyad 3. Blue curve: Participant A, red curve: Participant B. The 95% highest density interval for each curve is also shown.*

Another issue that our experimental design can contribute to addressing, is whether listeners are better at categorizing speech sounds when they do this collectively rather than individually. There is a large body of evidence showing that groups of people are better than individuals in a variety of perceptual tasks [17, 18, 19, 20]. To our knowledge, however, this remains to be demonstrated for speech. Another version of our game will consist in asking participants to learn to categorize sounds in a non-native language, in either an individual or collective manner. Based on the so-called group benefit found in other joint perceptual tasks [20], we may predict that the listeners' performance will be better in the collective vs. individual condition. If this proves to be the case, this will indicate that, in the speech perception domain too, "two heads are better than one" [17] with highly interesting implications for theories of learning and adaptation in speech communication.

6. Conclusion

Gamification is a strongly growing trend in behavioral sciences [21], including psycholinguistics [22] and speech perception studies [23]. It contributes to enhancing the attention, engagement and motivation of the participants. Experimental games that are played by two or more participants make it possible to study the impact of social interactions on a wide variety of sensori-motor, perceptual and cognitive mechanisms. In this paper, we have presented a novel experimental design that aims to combine highly accurate psychometric methods with an interactive task to characterize how two or more listeners can converge towards each other in the categorization of speech sounds. Our first results indicate that our proposed experimental design fulfills these objectives and opens new perspectives for future research on listener-to-listener adaptation in speech perception.

Work carried out with the prior approval of the Ethics Committee of Aix-Marseille University (approval #2024-01-11-01).

7. References

[1] S. D. Goldinger, "Echoes of echoes? an episodic theory of lexical access," *Psychological Review*, vol. 105, pp. 251–279, 1998.

[2] J. S. Pardo, "On phonetic convergence during conversational interaction," *Journal of the Acoustical Society of America*, vol. 119, pp. 2382–2393, 2006.

[3] J. S. Pardo, A. Urmanche, S. Wilman, and J. Wiener, "Phonetic convergence across multiple measures and model talkers," *Attention, Perception, & Psychophysics*, vol. 79, pp. 637–659, 2017.

[4] T. Kraljic and A. G. Samuel, "Perceptual learning for speech: Is there a return to normal?" *Cognitive Psychology*, vol. 51, pp. 141–178, 2005.

[5] D. Norris, J. McQueen, and A. Cutler, "Perceptual learning in speech," *Cognitive Psychology*, vol. 47, pp. 204—238, 2003.

[6] W. F. Ganong, "Phonetic categorization in auditory word perception," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 6, pp. 110–125, 1980.

[7] V. A. Mann and B. H. Repp, "Influence of vocalic context on perception of the [ʃ]-[s] distinction," *Perception & Psychophysics*, vol. 28, pp. 213–228, 1980.

[8] N. Nguyen, S. Dufour, and A. Brunellière, "Does imitation facilitate word recognition in a non-native regional accent?" *Frontiers in Psychology*, vol. 3, 2012, Article 480.

[9] L. Lancia and N. Nguyen, "The joint perception and categorization of speech sounds: A pilot study," in *Proc. 7th Joint Action Meeting*, Genova, Italy, July 2019.

[10] N. Nguyen, L. Lancia, L. Huttner, J.-L. Schwartz, and J. Diard, "Listeners' convergence towards an artificial agent in a joint phoneme categorization task," *Glossa Psycholinguistics*, vol. 2, no. 1, pp. 1–48, 2023.

[11] L. Huttner and N. Nguyen, "Alignment in speech sound perception," in *Proc. AMLaP 2022 – 28th Conference on Architectures and Mechanisms for Language Processing*, York, GB, Sep. 2022.

[12] D. C. Stevenson, "Categorical perception and selective adaptation phenomena in speech," Ph.D. dissertation, University of Alberta, 1979.

[13] P. C. Bürkner, "brms: An R package for Bayesian multilevel models using Stan," *Journal of Statistical Software*, vol. 80, pp. 1–28, 2017.

[14] V. Gelardi, J. Godard, D. Paleressompoulle, N. Claidière, and A. Barrat, "Measuring social networks in primates: wearable sensors versus direct observations," *Proceedings of the Royal Society A*, vol. 476, 2020, Article 20190737.

[15] D. A. Kenny, D. A. Kashy, and W. L. Cook, *Dyadic data analysis*. New York: Guilford Publications, 2020.

[16] S. W. Gregory and S. Webster, "A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions," *Journal of Personality and Social Psychology*, vol. 70, pp. 1231–1240, 1996.

[17] B. Bahrami, K. Olsen, P. E. Latham, A. Roepstorff, G. Rees, and C. D. Frith, "Optimally interacting minds," *Science*, vol. 329, pp. 1081–1085, 2010.

[18] A. Koriat, "When are two heads better than one and why?" *Science*, vol. 336, pp. 360–362, 2012.

[19] R. D. Sorkin, C. J. Hays, and R. West, "Signal-detection analysis of group decision making," *Psychological Review*, vol. 108, pp. 183–203, 2001.

[20] B. Wahn, A. Kingstone, and P. König, "Group benefits in joint perceptual tasks—a review," *Annals of the New York Academy of Sciences*, vol. 1426, pp. 166–178, 2018.

[21] B. Long, J. Simson, A. Buxó-Lugo, D. G. Watson, and S. A. Mehr, "How games can make behavioural science better," *Nature*, vol. 613, pp. 433–436, 2023.

[22] K. Christianson, J. Dempsey, A. Tsiola, and M. Goldshtein, "What if they're just not that into you (or your experiment)? on motivation and psycholinguistics," *Psychology of Learning and Motivation*, vol. 76, pp. 51–88, 2022.

[23] D. Duran and N. Lewandowski, "Demonstration of a serious game for spoken language experiments — GDX," in *Proceedings of the LREC 2020 Workshop Games and Natural Language Processing*, Marseille, France, 2020, pp. 68–78.