Proceedings of the LREC 2020 Workshop Games and Natural Language Processing, pages 68–78
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Demonstration of a Serious Game for Spoken Language Experiments — GDX

Daniel Duran[1] **& Natalie Lewandowski**[2]
[1] Albert-Ludwigs-Universität Freiburg, Germany
[2] High Performance Computing Center Stuttgart (HLRS), Germany
daniel.duran@germanistik.uni-freiburg.de, natalie.lewandowski@hlrs.de

Abstract

Increasing efforts are put into gamification of experimentation software in psychology and educational applications and the development of serious games. Computer-based experiments with game-like features have been developed previously for research on cognitive skills, cognitive processing speed, working memory, attention, learning, problem solving, group behavior and other phenomena. It has been argued that computer game experiments are superior to traditional computerized experimentation methods in laboratory tasks in that they represent holistic, meaningful, and natural human activity. We present a novel experimental framework for forced choice categorization tasks or speech perception studies in the form of a computer game, based on the Unity Engine – the Gamified Discrimination Experiments engine (GDX). The setting is that of a first person shooter game with the narrative background of an alien invasion on earth. We demonstrate the utility of our game as a research tool with an application focusing on attention to fine phonetic detail in natural speech perception. The game-based framework is additionally compared against a traditional experimental setup in an auditory discrimination task. Applications of this novel game-based framework are multifarious within studies on all aspects of spoken language perception.
Keywords: spoken language, gamification, categorization tasks, speech perception

1. Introduction

We present an experimental framework designed as a *computer game*[1] for auditory categorization and perception studies. We demonstrate its utility as a research tool with an application focusing on attention to fine phonetic detail in natural speech perception.

Increasing efforts are put into *gamification* of experimentation software in psychology and educational applications and the development of *serious games* or *games with a purpose* in natural language processing, computational linguistics and other related research disciplines. Computer game paradigms have been applied in studies with adult subjects, children and even monkeys (Berger et al., 2000; Keil et al., 2016; Washburn and Gulledge, 1995). Regarding the tested skills, computer games have been developed for research on cognitive skills (Donchin, 1995; Lindstedt and Gray, 2015), cognitive processing speed (McPherson and Burns, 2007; McPherson and Burns, 2008), working memory (Washburn and Gulledge, 1995), attention (Berger et al., 2000), learning (Nelson et al., 2014), problem solving (Quinn, 1991), or group behavior (Hawkins, 2015; Keil et al., 2016), etc. They have also been developed for computer-assisted language learning (Peterson, 2010). The body of work with applications of computer games as research tools to study some aspects of human language processing, however, is still comparably small.

In this paper we present a novel experimental framework for forced choice categorization tasks or speech perception studies, designed in the form of a computer game – the *Gamified Discrimination Experiments* engine (GDX). The remainder of this paper is structured as follows. First we give a brief overview of related work which comes primarily from research fields other than natural language processing. We also briefly discuss classic experimental approaches which are employed in the study of the mechanisms of human language understanding in psycholinguistics and cognitive sciences. GDX, our novel experimental framework, is described in detail in section 3. along with a first use case in a study on speech perception. In section 4., we compare the application of GDX with a classical test scenario. Finally, we discuss our findings in the context of gamified spoken language experiments.

2. Related work on serious games

Gamification of experimentation software and *serious games* or *games with a purpose* have been employed in various human behavior and language related research disciplines like psychology, cognitive sciences, computational linguistics or natural language processing. Usually, such computer games are custom made for the purpose of a specific study or data acquisition task. Using existing off-the-shelf computer games for research may be possible for some research questions. For example, *Tetris* has been used to study cognitive skills (Kirsh and Maglio, 1994; Maglio and Kirsh, 1996). The commercial games *The Sims* or *World of Warcraft* have been employed for computer-assisted language learning (Peterson, 2010). However, this is in general not possible with all tasks or experimental designs. Donchin (1995), for example, points out: "A game is useful as a research tool if, and only if, the investigator can exercise systematic control over the game's parameters." The researcher needs to know the internal workings of a game in order to develop appropriate empirical procedures and gather the required data from the participants and their interaction with the game (Porter, 1995). One very important aspect is detailed logging of user ac-

[1] A note on terminology: We use the term *computer game* throughout this paper to refer to interactive software programs which represent some sort of game. Most of the general discussion is applicable irrespective of the fact whether it is a competitive or cooperative game, whether it is a single-player or multiplayer game or whether it is made for PC, smartphones or dedicated gaming hardware (i.e. a video game console). The term *video game*, thus, is treated as synonymous with *computer game*. Furthermore, we do not discuss the differences between *serious games* and *games with a purpose*.

tions and game events, which is usually not possible with proprietary computer games (Järvelä et al., 2014; Lindstedt and Gray, 2015).

Apart from experimental research, gamification is also often employed in educational applications (Gruenstein et al., 2009; Habernal et al., 2018; Mayer et al., 2014; McGraw et al., 2009). Picca et al. (2015) review various serious games which employ some NLP techniques with applications in: tutoring systems, computer-assisted foreign language learning, risk management training, communication skills training, conflict resolution training, cognitive-behavioral therapy or scientific and academic education.

Serious games are not only an effective alternative to classic experimentation frameworks – e.g. with respect to participant motivation and naturalness of the gathered data. They are also valuable tools in crowdsourcing and labeling scenarios – e.g. for language data annotations and manual classifications (von Ahn, 2006; Kicikoglu et al., 2019; Madge et al., 2019). Levitan et al. (2018), for example, present a gamification approach for annotation of deceptive speech.

2.1. The computer game paradigm in psychology and cognitive research

Most applications of computer game experiments can probably be found in experimental research in psychology and cognitive sciences. Järvelä et al. (2014) review the use of computer games as "experiment stimulus" and provide a practical guide for game selection and experimental set-up.

Space Fortress is an example of a game developed in the early 1980s for research on skill acquisition (Donchin, 1995). Using the game *Tetris*, it was found that skilled players use more *epistemic actions*, e.g. rotating a piece physically instead of rotating it mentally in order to see if it fits (Kirsh and Maglio, 1994; Maglio and Kirsh, 1996). Later, Lindstedt and Gray (2015) presented *Meta-T*, a *Tetris*-like computer game for cognitive research. They discuss the use of computer games as a means to investigate complex, cognitive behavior of highly skilled experts (gamers) and novices. Other games are employed to study acquisition, categorization or learnability in psycholinguistics experiments (Wade and Holt, 2005; Lim and Holt, 2011; Kimball et al., 2013; Rácz et al., 2017)

2.2. Computer games in linguistics

Games are a well-established paradigm in speech production studies as an elicitation tool. One example is the well-known *Map Task* (Anderson et al., 1991). It provides a pen-and-paper framework to elicit quasi-spontaneous dialogs. In this task, two participants have to find a path on a printed map. Both participants receive a map of their own and they are not able to see the map of their dialog partners. However, the two maps contain different information and the only way to navigate through it is to exchange information verbally. The experimenter can influence the content of the dialog, to a certain extend, by the specific landmarks shown on the maps. Another example is the *Diapix* task (Baker and Hazan, 2011; Van Engen et al., 2010). It is similar to a map task but involves two pictures of various scenes with the task being to spot the differences.

In analogy to these pen-and-paper tasks, cooperative computer games are often used as an elicitation tool for research on human verbal interaction (Garrod and Anderson, 1987; Levitan et al., 2012; Ward and Abu, 2016).

2.3. Classic computerized experimentation methods

Commonly used experimental frameworks in language research are *DMDX* (Forster and Forster, 2003), *PsychoPy* (Peirce, 2007) or *Praat* (Boersma, 2001; Boersma and Weenink, 2020). Classic computerized experimentation methods like these involve explicit instructions for the participants. Their attention is drawn directly to the phenomenon under investigation such that each decision is made consciously. However, human language and speech processing is affected (among many others also) by cognitive factors like attention, distraction and memory (Duran and Lewandowski, 2018) cognitive resources which are likely to be employed in different ways in experimental settings or everyday situations. In addition to the inherently unnatural scenarios created by such experiments, they are most often carried out within an artificial laboratory setting. This raises questions about the validity and naturalness of the obtained data. Consequently, it has been argued that game experiments are superior to traditional experimentation methods. Porter (1995), for example, states: "To a much greater extent than most traditional laboratory tasks, computer games represent holistic, meaningful, and natural human activity."

Lumsden et al. (2016) carried out a simple Go/No-Go experiment where participants have to respond as quickly as possible to some stimuli (Go) but withhold their response to other stimuli (No-Go). They compared this task in different presentation forms: a traditional non-game version, a traditional version with an added scoring mechanism to reward participants for correct actions and a game version (a "cowboy shootout"). They found longer reaction times with the game version as well as lower accuracies. Note, however, that a ceiling effect of accuracy was observed on the non-game variants. It thus can be argued in favor of the game variant, if avoiding ceiling effects is considered desirable. It has to be mentioned, though, that higher visual complexity in the game variant may have contributed to the increased difficulty and the resulting lower performance. A questionnaire about enjoyment and engagement showed that the "non-game control was clearly rated as the least enjoyable and stimulating, the most boring and the most frustrating", and that "participants also reported putting less effort into this variant than others."

3. The GDX framework

In order to alleviate the known issues with classic computerized experimentation methods (mentioned above), we developed a computer-game based experimental framework for forced choice categorization tasks and speech perception studies: GDX – the *Gamified Discrimination Experiments* engine[2]. It was originally motivated by a study

[2]GDX is available for research purposes from the first author upon request. A freely-available version is being prepared for public release.

Figure 1: GDX screenshot: Beginning of a trial during the training phase with visual category information “human”.

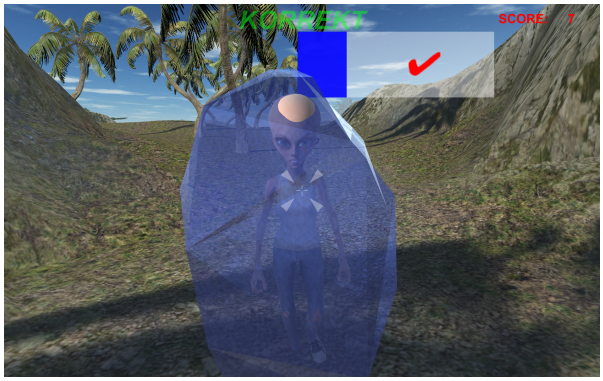

Figure 2: GDX screenshot: Feedback after the end of a trial with a correct classification as “alien”.

on phonetic convergence (see section 4.2. below) to assess individual differences in attention to fine phonetic detail during speech perception in verbal interactions. However, within this study it was not *explicit* attention to fine phonetic detail which we wanted to assess. Explicit, i.e. consciously directed attention probably involves processes which are different from the processes leading to phonetic convergence in natural conversations. We therefore developed an experimental framework based on a gamification approach, where attention could be gauged in an implicit manner. GDX was first employed during the creation of the GECO2 database, which contains spontaneous dialog recordings (Schweitzer et al., 2015), The game was one task among many psychological, social and cognitive tests the subjects had to complete aside the main dialog recordings.

The setting of GDX is that of a first person shooter game with the narrative background of an alien invasion on earth (inspired by an earlier version implemented by Lange et al. (2015)). The remainder of this section describes technical details of the game.

3.1. Design and Implementation

GDX is implemented using the Unity game engine (Unity Technologies, 2016). This provides a state-of-the-art game engine for a high-quality 3D game. Subjects experienced with modern computer games may find this appealing. The game is designed such that experimental parameters are not hard-coded into the game but can be set through a simple configuration file which is loaded by the game at runtime. The game takes place in a virtual 3D environment through which the player has to navigate. Navigation in GDX is controlled via the WASD keys on the keyboard in combination with the computer mouse. This scheme is common in first-person action games of this type. The player encounters agents (“enemies”) to which she/he has to react. In order to minimize interaction between experimenter and participants, all instructions are incorporated into the game and presented subsequently on screen within the game. This approach additionally facilitates the immersion of the participant with the virtual game environment and the background story. Screenshots are shown in Figure 1 and Figure 2.

The narrative of GDX is that of an alien invasion on earth from outer space. The player encounters agents of two categories – “humans” and “aliens”. Within the story of the game, the aliens are disguised as humans. Only during an initial training phase are visual cues shown to indicate the category of an agent (Figure 1). After a few trials, visual cues disappear and the agents are visually indistinguishable. Once the player approaches a given agent, the agent becomes active and starts chasing the player. This starts an experimental trial. A sound stimulus is played once and an optional visual display next to the agent shows a color along with a descriptive text label. The player is equipped with two tools (“weapons”): one that freezes a hit agent in a block of ice and one which beams a hit agent away within a bundle of green light rays. The tools are associated with the left and right mouse buttons and correspond to the two response categories. After the end of a trial, feedback is provided to the player about the true category of an agent (also showing an *alien* figure instead of the default *human* figure in case the agent belonged to the *alien* category, cf. Figure 2). All player actions are logged and stored in a text file for post-processing and evaluation of reaction times and response accuracy.

3.2. Experimental control, logging and reaction time measurements

All game logic (like input handling, agent behavior, experiment control, logging, etc.) is implemented in C#. Experimental parameters are not hard-coded into the game but can be set through a plain text file which is loaded by the game at runtime. The structure of this configuration file corresponds to the familiar Java *.properties* format with key–value pairs. The configurable parameters include, a.o., time limits, trial specifications and also the texts displayed on screen. The actual sound files (using uncompressed wav format) are not compiled into the game, as well, but loaded at runtime from hard disk. This makes GDX very flexible, providing a language-independent framework for various experimental scenarios.

The player’s location and rotation in world-space are logged at key events during the game, e.g. on all mouse clicks (firing one of the two weapons), the beginning of experimental trials, or upon reaching specific landmarks.

Accuracy of time measurements is an important issue in behavioral experiments, which has been discussed for several decades now (Babjack et al., 2015; Segalowitz and Graves,

1990). The DMDX software presented by Forster and Forster (2003), for example, allows running experiments on machines with the Windows operating system. It specifically aims at the minimization of both display timing errors (by keeping track of the system's refresh cycle time) as well as response timing errors, by supporting parallel port input. For high-precision time measurements, GDX relies on the C# `Stopwatch` class (in `System.Diagnostics`) and its property `ElapsedTicks` which refers to the smallest possible unit of time that this class can measure. The actual resolution depends on the underlying operating system and hardware, but it remains constant during experimental runs on the same machine. At the beginning of each session with GDX, the `Stopwatch` update frequency and the high-resolution flag are written to the log file. This aids later analysis of timing *precision*.

4. Game vs. classic perception test

In order to evaluate the utility of GDX, we compared it with a classic perception test in a follow-up study (Lewandowski and Duran, 2018).

4.1. Test case: the role of attention in phonetic convergence

To demonstrate the utility of GDX as a testing environment for auditory stimuli, or more broadly, within all kinds of forced choice categorization tasks and speech perception studies, data were collected in conjunction with a phonetic convergence study. Within the GECO2 project (Schweitzer et al., 2015), we gathered data of thirty adult subjects, who performed the GDX game in the scenario described below. The test set-up of GDX was targeted at measuring the attention given to fine phonetic detail in speech, when no explicit instructions are given to the players.

4.2. Background: a socio-cognitive model of phonetic convergence

Phonetic convergence (sometimes also called *accommodation*, *alignment* or *entrainment*[3]) is the phenomenon when two speakers become more alike in their speech productions within the course of a dialog. It occurs (1) in laboratory set-ups, e. g. in shadowing tasks or question-answer sequences (Bailly and Lelong, 2010; Delvaux and Soquet, 2007; Namy et al., 2002; Nielsen, 2011); (2) between native or non-native speakers in (quasi) spontaneous dialogs (De Looze et al., 2011; Kim et al., 2011; Lewandowski, 2012; Lewandowski and Jilka, 2019; Schweitzer and Lewandowski, 2013; Schweitzer et al., 2015); (3) between non-native speakers in a shared L2 (Trofimovich and Kennedy, 2014); and (4) even in human–machines interaction (Beňuš et al., 2018; Gessinger et al., 2019).

Previous attempts to explain convergence (not only at the phonetic level) can be categorized into two branches. Probably the most prominent one is a socio-linguistic model: the *Communication Accommodation Theory* (Cat) (Giles, 2016). It attempts to model the motives and evaluations of switching in terms of a balance of social psychological processes focusing on social integration and differentiation (Sachdev and Giles, 2006). The fundamental assumption is that individuals use communication, in part, to indicate their attitudes toward each other, and, as such, communication is a barometer of the level of social distance between them (Sachdev and Giles, 2006). According to this model, convergence is an expression of attitudes towards the interlocutor, and is affected by intentions, goals and knowledge of the involved speakers. Thus, convergence is essentially a conscious means of expression.

The second model is a mechanistic one, as proposed by Pickering and Garrod (2013), for example. The goal of interaction for speakers is to achieve mutual understanding or "common ground" (Trofimovich and Kennedy, 2014). At least one way of doing so is to align or coordinate language at several linguistic levels (lexical, syntactic, and phonological) (Trofimovich and Kennedy, 2014). According to this model, phonetic convergence is caused by the adoption of perceived phonetic details, based on psychological and cognitive processes which link perception and production – the *perception-production feedback loop*. Thus, convergence is modeled as an automatic process here, and potential (social or other) influencing factors are not discussed by Pickering and Garrod in their original model.

As Babel (2012) correctly points out, a crucial aspect has been left out of the discussion between the above models – the reasons for the lack of convergence, which is fairly often observed. She points to several possible solutions, including the incapacity to resolve perceptual details, production biases, or a lack of sufficient attention.

Research on convergence during the last years shows more and more that it is affected not only by social aspects (Schweitzer et al., 2017; Schweitzer and Lewandowski, 2014), but also by psychological (personality-related) and cognitive (processing skill-related) individual differences (Babel and McGuire, 2015; Lewandowski, 2013; Lewandowski and Jilka, 2019; Vais et al., 2015) as well. Amongst the cognitive factors, one feature seems to be especially involved – namely attention. As defined by Segalowitz (2007), attention control is the ability to focus and refocus attention on different semantic levels. The executive control part of attention might also operate beyond mere semantic levels, for instance, when switching between different levels/dimensions of the speech signal, e.g. between meaning vs. form. Lewandowski and Jilka (2019) also find attention skills (as tested by a mental flexibility task – the Simon Test (Craft and Simon, 1970)) to modulate the amount of convergence in their study, next to personality features such as, for instance, openness. The lower the switch costs in the Simon Test (i.e., the faster the subjects were able to switch between the dimensions in the test), the more phonetic convergence they displayed during the conversations. Another dimension which proved to be related to convergence in the study above was the Behavior Inhibition Scale (BIS). Results indicate that speakers displaying less behavioral inhibition (i.e., they are put off to a smaller degree by negative encounters or the fear of bad outcomes) again show more convergence. The authors conclude that some speakers (those showing more talent) seem

[3]Note on terminology: *Imitation* is not considered to be a synonym for phonetic convergence occurring in conversational speech. Compare, for instance, the discussion in (Lewandowski and Jilka, 2019).

to be more skilled in switching between different signal types (in their case: meaning vs. sound) and potentially giving more weight to their speaking partners' pronunciation, opposed to just focusing on transmitting information in the dialog (Lewandowski and Jilka, 2019). This in turn, is a phenomenon observable in its purest form within actual conversations, where attention towards certain communicative aspects usually arises (or does not) without any explicit instructions, just as it can be tested with the here presented serious game GDX. Therefore, the first described use case is a comparison of a test for attention to phonetic detail using our GDX engine (no explicit instructions necessary) and a classic perception test (inherently containing explicit instructions pointing the subject towards "areas of interest" in the speech signal).

4.3. Classic categorization experiment

The *classic experiment* is a categorization test with acoustic stimuli, designed in a way to maximally resemble the game scenario (involving the category labels "human" and "alien", just as in the game). All manipulated items belonged to the "alien" category, whereas the original recordings were used as the "human" samples. The nature of the manipulation was not communicated to the participants (neither in the perception test nor in the game). However, since the setting was an auditory categorization test, it was obvious to the participants that they were supposed to focus on cues in the sound of the stimuli. This is in stark contrast to the game scenario, where the target dimension of the signal was never explicitly nor circumstantially revealed to the participants. Similarly to the game, after a short training phase, subjects had to categorize the stimuli in three blocks, with one manipulation at a time (as in the three game levels).

4.4. Participants and method

Our subjects in the comparison study were 24 German native speakers (aged 20–31, 12 female) divided into two groups with 12 subjects each, which differed in testing order (game first vs. perception test first).

The test group – *Group 1 (G1)* – played the game first and then completed the classic perception test, the control group – *Group 2 (G2)* – took part in the classic perception test first and played the game afterwards. The two test sessions followed each other with a 3–7 days' break. Analyzed were accuracy and reaction times, as well as individual post-hoc questionnaires on the evaluation of the two methods. Two participants suffered from a mild case of cybersickness while playing the game (Frey et al., 2007; Rebenitsch and Owen, 2016). After a short break, however, they were able to continue with the experiment. Since the break occurred still within the training phase before any RTs were measured, the data did not have to be discarded but was included in the evaluation.

4.5. Post-hoc questionnaires

The first post-hoc questionnaire for every participant included sociodemographic information and questions on the usage of computers and other electronic devices, and the frequency and type of games played either on a computer, console or smartphone. The data was summarized in the following variables: *isGamer* (yes/no), *GamingFrequency* in days per week, *GamingScore* (i.e. How many types of games and on how many devices are usually played), and *ElectronicsUseScore* reflecting how many devices (smartphone, console, laptop, computer, tablet etc.) are being used on a daily basis. The second questionnaire was filled out directly after the respective experiment (game and perception test) and included a.o. questions on the difficulty and fun of the game/test (on a scale from 1–5), and also questions on the used "strategy" during the experiment in order to distinguish between aliens and humans.

4.6. Hypotheses

G1 (who started with the game) is expected to perform worse in their first task than G2, who began with the perception test. This difference should be the result of the explicitness of the instructions G2 received regarding the task at hand/ target to attend to. Furthermore, the classic design allows to focus solely on the experienced auditory stimuli, without any distractors present – thus differing considerably from the game. In consequence, we should also see G2 outperforming G1 in their second task, the game, since they already know which cue is essential (i.e., the sounds uttered within the game) and would not be held back by a false reliance on semantic or other unrelated cues.

4.7. Results

A full discussion of the results within the context of its original study (Schweitzer et al., 2015), the cognitive aspects in dialog situations, is beyond the scope of this paper. We present the results of the game vs. classic perception experiment and demonstrate the utility of the framework to collect reaction time and behavioral data.

The data sets were transformed and prepared for analysis using R version 3.4.3 (R Core Team, 2017) and the packages *tidyverse* (Wickham, 2017), *dplyr* and *stringr*. The statistical analyses were performed using *afex* (Singmann et al., 2018) and *lmerTest* (Kuznetsova et al., 2017), and visualized with *ggplot2* (Wickham, 2016). Raw reaction times (RT) were first log-transformed before supplying it to the model. Visual inspection of normality plots did not show any obvious deviations. Table 1 shows descriptive statistics. The best fitting linear mixed model (lmer) for predicting the variable *RT(log)* was obtained by maximum likelihood t-tests using Satterthwaite approximations to degrees of freedom (lmerMod) after fitting a large model first and applying an automatic stepwise reduction with the *step* procedure in the *lmerTest* package, which was manually overseen and double-checked with model comparison anovas.

The resulting model with the best fit contains random intercepts for *stimulus* and *subject*, and the fixed factors shown in Table 2 (model parameters: AIC 816.9, BIC 887.1, logLik -394.5, deviance 788.9, df.resid 1102). The number of *correct responses* in the two test scenarios was predicted by fitting a maximal generalized linear model (GLM) of type *binomial* and a subsequent reduction of factors to achieve the best fit (see Table 3).

The lmer shows that correct responses came hand in hand with shorter reaction times, and perceived *fun* in the ex-

		Accuracy		reaction time	
Group	**Test**	**Mean**	**SD**	**Mean**	**SD**
1	game	0.42	0.49	3.09	1.60
1	classic	0.69	0.46	3.96	1.22
2	classic	0.80	0.40	3.30	1.05
2	game	0.76	0.43	1.54	0.68

Table 1: Proportion correct responses (accuracy) and reaction times (sec) in both tests and groups, without the training phase. SD = standard deviation.

periments reduced RTs. Furthermore, there was an effect for the type of the acoustic manipulation of the stimuli and strong interactions between *test*group* and *test* and participants' *gaming score*, with more gaming experience actually prolonging reaction times in the game (see Table 2). Post-hoc pairwise comparisons with Tukey HSD Tests were performed on the factors in the fixed effects of the linear mixed model. For the interaction of *group* and *test* all between-group and within-group comparisons reached significance, indicating that subjects in both groups and tests responded to the stimuli with differing RTs.

The GLM for *accuracy* shows an effect for test type (i.e. a considerable negative effect for the *game*), and a main negative effect of perceived difficulty of the experiment. The subjective evaluation of the game's difficulty level seems to correlate with an actual decrease in accuracy for the classic test, which is, however, reversed for the game. The significant interaction of *group* and *test* confirms that G2 performed better in the game than G1. There also is a small bias for *fun* in favor of the game, mediated by group (post-hoc Tukey: game(g2)-classic(g1), diff 0.246735, p adj. = 0.005566).

A further analysis focused on the performance of both groups on their respective first test – *Time 1* – treating the game for G1 and the categorization test for G2 as two conditions of one variable, since the subjects had no knowledge as to the nature of the target cues prior to Time 1. The difference in accuracy on the first performed test per group was significant (compare Table 1, *Wilcoxon Rank Sum*: $W = 23705$, $p < 0.001$) – G2 was better able to correctly categorize the stimuli in the perception test than G1 was in the game. The same was true for Time 2 – G2 playing the game (76% correct) outperformed G1 completing the perception test with 69% correct ($W = 23705$, $p < 0.001$). For the logged RTs, the differences between both tests at Time 1 (*Tukey multiple comparisons of means*: diff. 0.156863, $p < 0.001$) and at Time 2 were significant (diff. -0.977677, $p < 0.001$). Figures 3 and 4 display the individual differences in performance of our subjects in both tests and groups.

5. Discussion and Conclusion

Our first validation study was designed as a comparison between a classical perception test, as used in speech perception research, and our gamified framework GDX. Several aspects have been found to speak in favor of using gamified testing environments. First, unsurprisingly, we have found a noticeable individual variation between our participants. Also, as expected in a scenario without any explicit instruc-

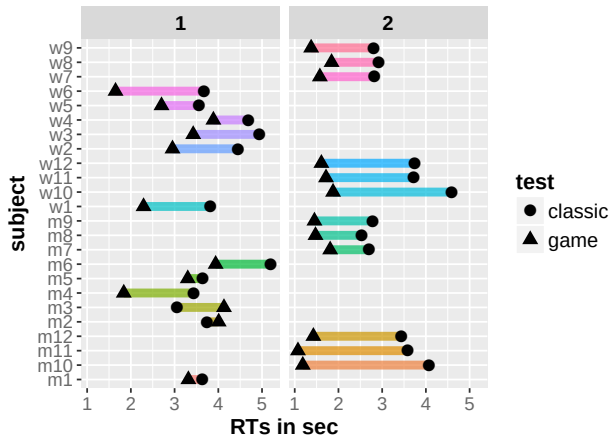

Figure 3: RTs in seconds in both tests per subject and group.

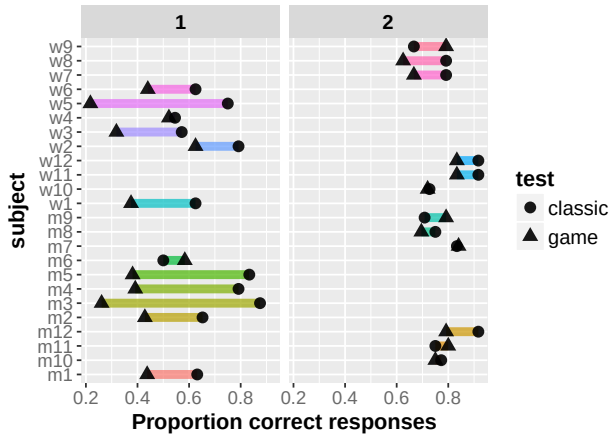

Figure 4: Proportion correct responses in both tests per subject and group.

tions, subjects have focused more on the (more salient) semantic content than on the acoustic information present in the game stimuli. Firstly, this seems to be a more genuine reflection of everyday communication, where meaning is the key, and the phonetic-acoustic part primarily serves as the means of transmission. Secondly, this casts doubt onto the validity of classical perception tests aiming at phonetic dimensions, or at the very minimum, onto the effect sizes observed in such tests. We presume that classic test designs might lead to exaggerated outcomes due to the explicit instructions subjects usually receive in these tasks (or, for example, in games like the one presented by Levitan et al. (2018)). Making participants aware which dimension they need to pay attention to, reduces task complexity considerably, and probably bypasses naturally occurring attention control or attention switching mechanisms, since the person already is focused on the "correct" task. In the reversed situation of the no-instruction gamified design, fewer subjects directed their attention towards the target dimension. Nevertheless, a number of participants in GDX were very successfully able to identify the task at hand (i.e. paying attention to the sounds) and reacted accordingly. These might be precisely those subjects who naturally pay more attention to sound properties in general, or, specifically in

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	1.82	0.17	44.12	10.90	0.00
testgame	-0.05	0.07	1081.66	-0.68	0.49
group	-0.14	0.08	24.47	-1.80	0.08
GamingScore	-0.05	0.03	24.63	-2.02	0.05
fun	-0.07	0.02	427.61	-3.29	0.00
manipulationF2	0.21	0.05	26.21	4.07	0.00
manipulationFRIC	0.11	0.05	31.66	2.26	0.03
manipulationOriginal	0.14	0.04	28.18	3.35	0.00
correct	-0.07	0.02	1091.13	-2.90	0.00
testgame:group	-0.47	0.04	1067.83	-11.69	0.00
testgame:GamingScore	0.07	0.01	1063.90	5.27	0.00

Table 2: Fixed factors in lmer: *RTlog ~test * (group + GamingScore) + fun + manipulation + correct + (1|stimulus) + (1|subject)*. Random effects: *stimulus (Intercept)*, var. 0.0032, SD 0.0564; *subject (Intercept)*: var 0.0324, SD 0.1801; *resid.*: var. 0.1100, SD 0.3317.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	1.3275	0.4327	3.07	0.0022
test(game)	-3.2427	0.5900	-5.50	0.0000
group	0.3055	0.2127	1.44	0.1509
difficulty	-0.3399	0.0902	-3.77	0.0002
test(game):group	1.2073	0.2852	4.23	0.0000
test(game):difficulty	0.3622	0.1229	2.95	0.0032

Table 3: GLM output for the proportion of correct responses (accuracy) in both tests and groups. Model formula: *correct ~test * (group + difficulty)*; null deviance: 1416.5 on 1115 df, residual deviance: 1295.1 on 1110 df, AIC: 1307.1

situations where this becomes a relevant factor in communication (e.g., in L1–L2 encounters, in the presence of dialect, in order to allow situation-adequate style choices, or for the purpose of convergence in dialogs). We are at the same time aware of certain limitations of this first validation study. Most importantly, a third testing condition with participants knowing before starting the game that sound is important (however without knowing the exact acoustic-phonetic feature targeted) would bridge the current gap between the two experimental conditions and allow an even more refined conclusion on the ability to pay attention to fine phonetic detail.

There are two aspects of computer game-like experiments which are frequently discussed: (1) the appeal of the task or motivation of participants, and (2) the quality of the collected data.

Motivation of participants through game-like features has been mentioned repeatedly in the literature as desirable (Lindstedt and Gray, 2015; Nelson et al., 2014), although it has also been argued that this may not necessarily improve data (Hawkins et al., 2012). Howes (2017) points out that "games are so motivating that [...] people actively *choose* to engage with them and, today, action games are a significant and growing part of the fabric of everyday human experience" (emphasis in original). He compares game paradigms with "extremely simple paradigms" (Gray, 2017) where studies focus on isolated cognitive processes in order to build a big picture. Referring to Newell (1973), he emphasizes that "the pieces never seem to get put back together". Lindstedt and Gray (2015) point out the aspect of participant *motivation* as an advantage of using a Tetris-like game for psychological studies stating that it "is not a boring experimental paradigm, but a fascinating game that has a life outside of academia".

Motivation is not only relevant in terms of engaging participants with the task during the experiment. It is also an important aspect in recruitment of participants for experiments, in the first place. Järvelä et al. (2014), for example, note that "the high penetration in the population serves to make games more approachable than abstract psychological tasks, which helps in recruiting participants." With computer games, social groups could be reached and recruited as subjects who usually do not find their ways into the labs of speech and language scientists. The kind of setup presented in this paper might not be suitable for all experiments or groups of subjects (e.g. taking into account the issue of cybersickness or different levels of experience with action games). Increased reaction times, as we find them (section 4.), for example, might indicate that the performance of (highly) experienced gamers is negatively affected by deviations from common game conventions. Further research is needed in order to asses the suitability of serious games in favour of classic experimental designs with participants beyond the usual subject group of undergraduate students.

Note also, that intrinsic motivation and *fun* may affect labeling and data annotation tasks. As a possible annotation tool, GDX exploits natural implicit judgments and does not require specially trained or skilled expert annotators. In comparison to common crowdsourcing methods (e.g. Amazon Mechanical Turk[4]), the gamification in GDX exploits intrinsic motivation of the participants in a more "natural environment". In comparison to explicit categorization tasks, the gamification in GDX thus allows for the elicitation of spontaneous behavioral (linguistic) data.

[4] https://www.mturk.com/

We would like to address the question, why we opt for conducting computer experiments in the lab rather than online via the internet. Online experiments face several issues which are easier to address in scenarios with local computer experiments, like control over test subjects (e.g. personal features like age, gender, language skills etc.) and how often they participate, or protection against malicious attacks on the system. Additionally, local experiments in the lab allow for control over the test situation and the used hardware and software equipment making results more consistent and comparable (Babjack et al., 2015). Other problems with web-based online experiments are: premature drop-out or loss of attention resulting in the participant's switching to other activities in the middle of an experiment as discussed, for example, by Hawkins (2015).

Another important issue in behavioral experiments which requires continued attention is the accuracy of time measurements. This has been discussed for several decades now (Babjack et al., 2015; Segalowitz and Graves, 1990). Experimentation software like DMDX specifically addresses timing issues by specific optimizations for operating systems and support of specific hardware Forster and Forster (2003). Babjack et al. (2015) observe that different configurations (operating system, sound card, API, etc.) introduce significant timing variability. They found mean sound onset latencies of approximately 25–35 ms on PC and 6–25 ms on laptops (running Windows 7 and 8). Often, such timing issues are tackled by the use of dedicated hardware for stimulus presentation and response detection. Within a game-like environment this is not feasible and also counters the goal of providing a low-cost, easy to use framework. Segalowitz and Graves (1990) strongly recommend external measurements of the timing accuracy of the employed computer systems and that "corrections of any systematic errors be made, and that such accuracy measurements and corrections be reported in published research articles". Unfortunately, timing *accuracy* is in general not easy to assess. It depends on various factors and may even change over time during running experiments. The timing mechanism implemented in GDX offers high precision time measurements which allow for analyses of reaction times. However, experimenters need to be aware of potential problems introduced by various combinations of hardware, operating system and other aspects of the experimental environment. In use cases where GDX is employed as a data annotation tool rather than a behavioral experiment framework, timing, of course, might not be of relevance.

In conclusion, we have demonstrated the utility of GDX for categorization tasks (within the scope of the described use case in the study of speech perception). We are positive that GDX offers a useful tool to researchers for experiments on human spoken language processing as well as categorization tasks such as data annotation.

6. Acknowledgements

The initial development of the game has been carried out at the Institute for Natural Language Processing at the University of Stuttgart within Project A4 of the Collaborative Research Center SFB 732 funded by the German Research Council (DFG), PI: Grzegorz Dogil and Antje Schweitzer. This work also benefitted from many discussions within the scientific network *Simphon.Net* (funded by DFG). We thank Jenny Krüwald who carried out some of the experiments within her Bachelor's thesis (Krüwald et al., 2018). We also thank the participants in our experiments and the anonymous reviewers for very helpful feedback.

7. Bibliographical References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

Babel, M. and McGuire, G. (2015). The effects of talker variability on phonetic accommodation. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, pages 1–5, Glasgow, UK. Paper number 661.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, 40(1):177–189.

Babjack, D. L., Cernicky, B., Sobotka, A. J., Basler, L., Struthers, D., Kisic, R., Barone, K., and Zuccolotto, A. P. (2015). Reducing audio stimulus presentation latencies across studies, laboratories, and hardware and operating system configurations. *Behavior Research Methods*, 47(3):649–665.

Bailly, G. and Lelong, A. (2010). Speech dominoes and phonetic convergence. In *Proceedings of Interspeech*, pages 1153–1156, Tokio (Japan).

Baker, R. and Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, 43(3):761–770.

Beňuš, Š., Trnka, M., Kuric, E., Marták, L., Gravano, A., Hirschberg, J., and Levitan, R. (2018). Prosodic entrainment and trust in human-computer interaction. In *Proceedings of the 9th International Conference on Speech Prosody 2018*, pages 220–224.

Berger, A., Jones, L., Rothbart, M. K., and Posner, M. I. (2000). Computerized games to study the development of attention in childhood. *Behavior Research Methods, Instruments, & Computers*, 32(2):297–303.

Boersma, P. and Weenink, D. (2020). Praat: doing phonetics by computer. [Computer program]. Version 6.1.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Craft, J. L. and Simon, J. R. (1970). Processing symbolic information from a visual display: Interference from an irrelevant directional cue. *Journal of Experimental Psychology*, 83(3, Pt.1):415–420.

De Looze, C., Oertel, C., Rauzy, S., and Campbell, N. (2011). Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In *International Conference on Phonetic Sciences (ICPhS). Hong Kong*, page 1294–1297.

Delvaux, V. and Soquet, A. (2007). Inducing imitative phonetic variation in the laboratory. In *Proceedings of the 16th ICPhS*, pages 369–372, Saarbrücken.

Donchin, E. (1995). Video games as research tools: The

Space Fortress game. *Behavior Research Methods, Instruments, & Computers*, 27(2):217–223.

Duran, D. and Lewandowski, N. (2018). Untersuchung der kognitiven Beanspruchung durch Sprachassistenzsysteme. In André Berton, et al., editors, *Elektronische Sprachsignalverarbeitung 2018: Tagungsband der 29. Konferenz*, pages 159–166. TUDpress, Dresden.

Forster, K. I. and Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35(1):116–124.

Frey, A., Hartig, J., Ketzel, A., Zinkernagel, A., and Moosbrugger, H. (2007). The use of virtual environments based on a modification of the computer game Quake III Arena® in psychological experimenting. *Computers in Human Behavior*, 23(4):2026–2039, July.

Garrod, S. and Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Gessinger, I., Möbius, B., Fakhar, N., Raveh, E., and Steiner, I. (2019). A Wizard-of-Oz Experiment to Study Phonetic Accommodation in Human-Computer Interaction. In *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 1475–1479.

H. Giles, editor. (2016). *Communication Accommodation Theory: Negotiating Personal Relationships and Social Identities Across Contexts*. Cambridge University Press.

Gray, W. D. (2017). Game-XP: Action Games as Experimental Paradigms for Cognitive Science. *Topics in Cognitive Science*, 9(2):289–307.

Gruenstein, A., McGraw, I., and Sutherland, A. (2009). A self-transcribing speech corpus: Collecting continuous speech with an online educational game. In *ISCA International Workshop on Speech and Language Technology in Education (SLaTE)*, pages 109–112.

Habernal, I., Pauli, P., and Gurevych, I. (2018). Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Hawkins, G. E., Rae, B., Nesbitt, K. V., and Brown, S. D. (2012). Gamelike features might not improve data. *Behavior Research Methods*, 45(2):301–318.

Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, 47(4):966–976.

Howes, A. (2017). Games for psychological science. *Topics in Cognitive Science*, 9(2):533–536.

Järvelä, S., Ekman, I., Kivikangas, J. M., and Ravaja, N. (2014). A practical guide to using digital games as an experiment stimulus. *Transactions of the Digital Games Research Association*, 1(2):85–115.

Keil, J., Michel, A., Sticca, F., Leipold, K., Klein, A. M., Sierau, S., von Klitzing, K., and White, L. O. (2016). The Pizzagame: A virtual public goods game to assess cooperative behavior in children and adolescents. *Behavior Research Methods*.

Kicikoglu, D., Bartle, R., Chamberlain, J., and Poesio, M. (2019). Wormingo: a 'true gamification' approach to anaphoric annotation. In *FDG '19: Proceedings of the 14th International Conference on the Foundations of Digital Games*, pages 1–7. Association for Computing Machinery. Article No.: 75.

Kim, M., Horton, W. S., and Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Lab Phon*, 2(1).

Kimball, G., Cano, R., Feng, J., Feng, L., Hampson, E., Li, E., Christel, M. G., Holt, L. L., Lim, S.-j., Liu, R., and Lehet, M. (2013). Supporting research into sound and speech learning through a configurable computer game. In *IEEE International Games Innovation Conference (IGIC)*, pages 110–113.

Kirsh, D. and Maglio, P. (1994). On Distinguishing Epistemic from Pragmatic Action. *Cognitive Science*, 18(4):513–549.

Krüwald, J., Duran, D., and Lewandowski, N. (2018). Gamification in the phonology lab. Presented at: LabPhon16 – Variation, development and impairment: Between phonetics and phonology.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *J. of Stat. Software*, 82(13):1–26.

Lange, L., Pfeiffer, B., and Duran, D. (2015). ABIMS – auditory bewildered interaction measurement system. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1074–1075.

Levitan, R., Gravano, A., Willson, L., Beňuš, Š., Hirschberg, J., and Nenkova, A. (2012). Acoustic-prosodic entrainment and social behavior. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19.

Levitan, S. I., Shin, J., Chen, I., and Hirschberg, J. (2018). LieCatcher: Game framework for collecting human judgments of deceptive speech. In Jon Chamberlain, et al., editors, *Games4NLP – Games and Gamification for Natural Language Processing. Proceedings*, pages 12–16.

Lewandowski, N. and Duran, D. (2018). Testing speech perception today and tomorrow: serious computer games as perception tests. In André Berton, et al., editors, *Elektronische Sprachsignalverarbeitung 2018: Tagungsband der 29. Konferenz*, pages 232–239. TUDpress, Dresden.

Lewandowski, N. and Jilka, M. (2019). Phonetic convergence, language talent, personality and attention. *Frontiers in Communication*, 4:18.

Lewandowski, N. (2012). *Talent in nonnative phonetic convergence*. Doctoral dissertation, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Lewandowski, N. (2013). Phonetic convergence and individual differences in non-native dialogs. Montréal, Canada. Abstract presented at New Sounds.

Lim, S.-j. and Holt, L. L. (2011). Learning Foreign Sounds in an Alien World: Videogame Training Improves Non-Native Speech Categorization. *Cognitive Science*, 35(7):1390–1405, September.

Lindstedt, J. K. and Gray, W. D. (2015). Meta-T: Tetris as an experimental paradigm for cognitive skills research. *Behavior Research Methods*, 47(4):945–965.

Lumsden, J., Skinner, A., Woods, A. T., Lawrence, N. S., and Munafò, M. (2016). The effects of gamelike features and test location on cognitive test performance and participant enjoyment. *PeerJ*, 4:e2184.

Madge, C., Bartle, R., Chamberlain, J., Kruschwitz, U., and Poesio, M. (2019). Making text annotation fun with a clicker game. In *FDG '19: Proceedings of the 14th International Conference on the Foundations of Digital Games*, pages 1–6. Association for Computing Machinery. Article No.: 77.

Maglio, P. P. and Kirsh, D. (1996). Epistemic action increases with skill. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, pages 391–396.

Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., van Ruijven, T., Lo, J., Kortmann, R., and Wenzler, I. (2014). The research and evaluation of serious games: Toward a comprehensive methodology. *British Journal of Educational Technology*, 45(3):502–527.

McGraw, I., Gruenstein, A., and Sutherland, A. (2009). A self-labeling speech corpus: Collecting spoken words with an online educational game. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3031–3034.

McPherson, J. and Burns, N. R. (2007). Gs Invaders: Assessing a computer game-like test of processing speed. *Behavior Research Methods*, 39(4):876–883.

McPherson, J. and Burns, N. R. (2008). Assessing the validity of computer-game-like tests of processing speed and working memory. *Behavior Research Methods*, 40(4):969–981.

Namy, L. L., Nygaard, C. L., and Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *J Lang Soc Psychol*, 21:422–432.

Nelson, J. B., Navarro, A., and Sanjuan, M. d. C. (2014). Presentation and validation of "The Learning Game," a tool to study associative learning in humans. *Behavior Research Methods*, 46(4):1068–1078.

Newell, A. (1973). You can't play 20 questions with nature and win: projective comments on the papers of this symposium. Technical report, Carnegie Mellon University. Research Showcase at CMU.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, 39(2):132–142.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2):8–13, May.

Peterson, M. (2010). Computerized games and simulations in computer-assisted language learning: A meta-analysis of research. *Simulation & Gaming*, 41(1):72–93.

Picca, D., Jaccard, D., and Eberlé, G. (2015). Natural language processing in serious games: A state of the art. *International Journal of Serious Games*, 2(3):77–97.

Pickering, M. J. and Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(04):329–347.

Porter, D. B. (1995). Computer games: Paradigms of opportunity. *Behavior Research Methods, Instruments, & Computers*, 27(2):229–234.

Quinn, C. N. (1991). Computers for cognitive research: A HyperCard adventure game. *Behavior Research Methods, Instruments, & Computers*, 23(2):237–246.

R Core Team, (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rácz, P., Hay, J. B., and Pierrehumbert, J. B. (2017). Social salience discriminates learnability of contextual cues in an artificial language. *Frontiers in Psychology*, 8.

Rebenitsch, L. and Owen, C. (2016). Review on cybersickness in applications and visual displays. *Virtual Reality*, 20(2):101–125.

Sachdev, I. and Giles, H. (2006). Bilingual Accommodation. In Tej K. Bhatia et al., editors, *The handbook of bilingualism*, pages 353–378. Blackwell Publishing, Malden, MA, USA.

Schweitzer, A. and Lewandowski, N. (2013). Convergence of articulation rate in spontaneous speech. In *Proceedings of the 14th Interspeech, Lyon*, pages 525–529.

Schweitzer, A. and Lewandowski, N. (2014). Social factors in convergence of f1 and f2 in spontaneous speech. In *Proceedings of the 10th International Seminar on Speech Production, Cologne*.

Schweitzer, A., Lewandowski, N., and Duran, D. (2015). Attention, please! Expanding the GECO database. In The Scottish Consortium for ICPhS 2015, editor, *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK. Paper number 620.

Schweitzer, A., Lewandowski, N., and Duran, D. (2017). Social Attractiveness in Dialogs. In *Interspeech 2017*, pages 2243–2247.

Segalowitz, S. J. and Graves, R. E. (1990). Suitability of the IBM XT, AT, and PS/2 keyboard, mouse, and game port as response devices in reaction time paradigms. *Behavior Research Methods, Instruments, & Computers*, 22(3):283–289.

Segalowitz, N. (2007). Access Fluidity, Attention Control, and the Acquisition of Fluency in a Second Language. *TESOL Quarterly*, 41(1):181–186.

Singmann, H., Bolker, B., Westfall, J., and Aust, F., (2018). *afex: Analysis of Factorial Experiments*. R package version 0.19-1.

Trofimovich, P. and Kennedy, S. (2014). Interactive alignment between bilingual interlocutors: Evidence from two information-exchange tasks. *Bilingualism: Language and Cognition*, 17(04):822–836.

Unity Technologies. (2016). Unity. Computer program. Version 5.

Vais, J., Walsh, M., and Lewandowski, N. (2015). Investigating frequency of occurrence effects in l2 speakers: Talent matters. In The Scottish Consortium for ICPhS 2015, editor, *Proceedings of the 18th International Congress of Phonetic Sciences*, pages 1–5, Glasgow, UK. Paper number 723.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., and Bradlow, A. R. (2010). The Wildcat Corpus of Native-and Foreign-accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profiles. *Language and Speech*, 53(4):510–540.

von Ahn, L. (2006). Games with a purpose. *Computer*, 39(6):92–94.

Wade, T. and Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *The Journal of the Acoustical Society of America*, 118(4):2618–2633.

Ward, N. G. and Abu, S. (2016). Action-coordinating prosody. In *Proc. Speech Prosody*, pages 629–633.

Washburn, D. A. and Gulledge, J. P. (1995). Game-like tasks for comparative research: Leveling the playing field. *Behavior Research Methods, Instruments, & Computers*, 27(2):235–238.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.