ETS TOEFL

Quality Beyond Measure.



*TOEFL*® Research Report
TOEFL–RR-98
ETS Research Report No. RR–22-05

# The Impact of Using Synthetically Generated Listening Stimuli on Test-Taker Performance: A Case Study With Multiple-Choice, Single-Selection Items

Ikkyu Choi

Jiyun Zu

December 2022

The *TOEFL*® test is the world's most widely respected English language assessment, used for admissions purposes in more than 130 countries including Australia, Canada, New Zealand, the United Kingdom, and the United States. Since its initial launch in 1964, the TOEFL test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the *TOEFL iBT*® test, contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments. In addition to the TOEFL iBT, the TOEFL Family of Assessments has expanded to provide high-quality English proficiency assessments for a variety of academic uses and contexts. The TOEFL Young Students Series (YSS) features the *TOEFL*® *Primary*™ and *TOEFL Junior*® tests, designed to help teachers and learners of English in school settings. The *TOEFL ITP*® Assessment Series offers colleges, universities, and others an affordable test for placement and progress monitoring within English programs.

Since the 1970s, the TOEFL tests have had a rigorous, productive, and far-ranging research program. ETS has made the establishment of a strong research base a consistent feature of the development and evolution of the TOEFL tests, because only through a rigorous program of research can a testing company demonstrate its forward-looking vision and substantiate claims about what test takers know or can do based on their test scores. In addition to the 20-30 TOEFL-related research projects conducted by ETS Research & Development staff each year, the TOEFL Committee of Examiners (COE), composed of distinguished language-learning and testing experts from the academic community, funds an annual program of research supporting the TOEFL family of assessments, including projects carried out by external researchers from all over the world.

To date, hundreds of studies on the TOEFL tests have been published in refereed academic journals and books. In addition, more than 300 peer-reviewed reports about TOEFL research have been published by ETS. These publications have appeared in several different series historically: TOEFL Monographs, TOEFL Technical Reports, TOEFL iBT Research Reports, and TOEFL Junior Research Reports. It is the purpose of the current TOEFL Research Report Series to serve as the primary venue for all ETS publications on research conducted in relation to all members of the TOEFL Family of Assessments.



To obtain more information about the TOEFL programs and services, use one of the following:

**E-mail:** toefl@ets.org **Web site:** www.ets.org/toefl


*ETS is an Equal Opportunity/Affirmative Action Employer.*

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

TOEFL Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

# The Impact of Using Synthetically Generated Listening Stimuli on Test-Taker Performance: A Case Study With Multiple-Choice, Single-Selection Items

Ikkyu Choi & Jiyun Zu

ETS, Princeton, NJ

Synthetically generated speech (SGS) has become an integral part of our oral communication in a wide variety of contexts. It can be generated instantly at a low cost and allows precise control over multiple aspects of output, all of which can be highly appealing to second language (L2) assessment developers who have traditionally relied upon human voice actors for recording audio materials. Nevertheless, SGS is not widely used in L2 assessments. One major concern in this use case lies in its potential impact on test-taker performance: Would the use of SGS (as opposed to using human voice actor recordings) change how test takers respond to an item? In this study, we investigated using SGS as stimuli for English L2 listening assessment items on test-taker performance. The data came from a pilot administration of multiple new task types and included 653 test takers' responses to two versions of the same 13 items, differing only in terms of their listening stimuli: a version using human voice actor recordings and the other version with SGS files. Multifaceted comparisons between test takers' responses across the two versions showed that the two versions elicited remarkably comparable performance. The comparability provides strong empirical evidence for the use of SGS as a viable alternative for human voice actor recordings in the immediate domain of L2 assessment as well as related domains such as learning material and research instrument development.

**Keywords** *TOEFL Essentials*™ test; item difficulty; item discrimination; L2 listening assessment; synthetically generated speech; test taker performance

doi:10.1002/ets2.12347

Synthetically generated speech (SGS) has become part of our everyday life. It now constitutes an important method for our interaction with many devices, from watches to cars, and its use cases include wide-ranging contexts. For example, Wagner et al. (2019) listed 12 SGS applications, ranging from static and asynchronous (e.g., announcements) to dynamic real-time (e.g., dialog systems). SGS is better and more available than ever. State-of-the-art systems have achieved performances that are almost comparable to natural speech (e.g., Elias et al., 2020). Open source SGS architectures exist (e.g., De Silva et al., 2018), and companies such as Google and Amazon provide pay-per-use access to their state-of-the-art technologies.

Despite the quality and availability of SGS, it has not been widely adopted in second or foreign language (L2) assessments. This limited usage may seem vexing to some, as several characteristics of SGS can be quite appealing to L2 assessment developers. L2 assessments, especially those measuring listening proficiency, rely heavily on recorded monologs and conversations for their listening stimuli. Most involve human voice actors, and as a result, the process of creating stimuli can be expensive and time-consuming. SGS has the potential to make the process cheaper and faster. It can eliminate the need to secure human voice actors, recording engineers, and required equipment, allowing assessment developers to instantly turn written scripts into listening stimuli. One major concern in this use case, however, lies in the impact on test-taker performance: What if using SGS as stimuli (as opposed to using human voice actor recordings) changes how test takers perform on assessment items and introduces construct-irrelevant variance? Unfortunately, little is known about whether and how the use of SGS affects test-taker performance.

In this paper, we report on an empirical investigation of the impact on test-taker performance of using SGS as stimuli for L2 listening assessment items. This impact was estimated by comparing the performance on two versions of the same items that differ only in terms of how their stimuli were created: a version using human voice actor recordings and the other version with SGS. Our finding has direct implications for assessment developers who are interested in utilizing SGS

*Corresponding author:* I. Choi, E-mail: ichoi001@ets.org

for L2 listening proficiency assessment. The remainder of this paper is organized as follows: We first review the relevant literature and situate within it the goal of this study. Next, we describe our method, including the source of data and analysis procedures, and present the results from the analysis. We then discuss the results and their implications and conclude with the limitations of this study and next steps.

## Review of Literature

SGS has a long history that spans multiple centuries (Story, 2019) and has evolved from its assistive origin (e.g., Euler, 1761) into a ubiquitous technology whose use cases include multiple aspects of everyday life (Wagner et al., 2019). One such use case that has attracted the interest of applied linguists is second or foreign language (L2) learning. Several researchers (Delmonte, 2008; Kılıçkaya, 2006; Sha, 2010) noted the flexibility and cost-effectiveness of SGS as well as its potential as a useful tool for L2 learning, for which empirical evidence has been reported across multiple domains, such as L2 pronunciation (e.g., Kılıçkaya, 2011; Liakin et al., 2017a, 2017b; Qian et al., 2018), listening comprehension (e.g., Sha, 2010), reading (Proctor et al., 2007; Volodina & Pijetlovic, 2015), and writing (Kirstein, 2006; Shadiev et al., 2018).

Despite such evidence, however, there exists a consensus that more empirical studies are needed to further and better understand the impact of SGS use for L2 learning (Cardoso, 2018; Liakin et al., 2017b; Soler Urzúa, 2011). A particularly interesting topic is the way learners perceive SGS and whether their perception affects their interaction with learning materials. Although multiple researchers have deemed the quality of SGS comparable to that of human-voiced materials (e.g., Azuma, 2008; Cardoso et al., 2015), the evaluation by learners has yielded somewhat mixed results. For example, although the participants in Kataoka et al. (2007) did not even notice that their listening materials included SGS, Kılıçkaya's (2011) participants noted during their interaction with SGS-based materials the lack of tonal variations and the resulting "unnaturalness."

The quality of an SGS technology can have a major impact on learners' perception of their interaction with SGS. Studies on learner perception of SGS in the L2 learning literature (e.g., Bione et al., 2016; Kataoka et al., 2007; Kılıçkaya, 2011; Liakin et al., 2017b) were based on different SGS technologies at different time points; their results are conditional on the specific technology adopted. The literature does not provide a systematic view of how learners' perception has been affected or changed as the SGS technology has evolved over time. There is a consensus, with empirical backing, that the technology has improved dramatically over the last decade (e.g., Ning et al., 2019). The standard measure for evaluating SGS has been the mean opinion score (MOS; International Telecommunication Union, 2016), which is obtained via 5-point Likert scale items capturing the global impression of listeners. At the time of this writing, multiple state-of-the-art SGS systems have achieved the MOS that is comparable to that for a human voice (e.g., Kong et al., 2021; Li et al., 2019; Shen et al., 2018), a feat that was not achieved until a few years ago (e.g., Georgila, 2017).

Researchers have also explored and discussed the limitations of the MOS. A careful analysis of its content by Lewis (2001) revealed inconsistent grain sizes across the items, and investigations into its internal structure yielded mixed results (Kraft & Portele, 1995; Sonntag et al., 1999; Viswanathan & Viswanathan, 2005). Multiple researchers (Betz et al., 2018; Rosenberg & Ramabhadran, 2017; Wester et al., 2015) have also noted the potentially unstable nature of the MOS, especially considering how it is often obtained by presenting decontextualized sentences to a small number of listeners. Another fundamental limitation of the MOS is that, because of its focus on listener perception, it provides little information about whether and how SGS affects listener behaviors. In a recent review of the literature on SGS evaluation, Wagner et al. (2019) noted this limitation and called for evaluation based on the impact of SGS on listener behaviors.

In an L2 assessment context, an important question about the impact of SGS on listener behaviors has to do with their performance. Test-taker performance is captured in their responses to assessment items and item characteristics such as difficulty and discrimination. The question can thus be rephrased as the following: Would the use of SGS change how test takers respond to an item and how difficult or discriminating an item is? A prerequisite to answering this question is empirical knowledge about the use of SGS in L2 assessment contexts. However, to our knowledge, there is no such reported study. In fact, little is known about the use of SGS in L2 assessment in general. The only place in which SGS is mentioned in assessment-related materials appears to be accommodation documents (IELTS, n.d.).

This knowledge gap presents a major challenge for utilizing SGS in L2 assessment. Assessments are designed such that the resulting scores reflect test-taker proficiency without artifacts due to measurement methods (Crocker & Algina, 2006). Assessment developers are responsible for making and justifying with empirical backing a validity argument for their assessment (Bachman & Palmer, 2010; Kane, 2013). The impact of SGS use on test-taker performance is thus a necessary

piece of evidence that allows a meaningful evaluation of whether the use case can be justified. The lack of such evidence makes it difficult for assessment developers to utilize SGS despite its practical benefits.

## Research Questions

The goal of this study was to empirically investigate the impact on test-taker performance of using SGS audio files (as opposed to human voice actor recordings) as listening stimuli for English L2 assessment items. We focused on the impact on multiple-choice single-selection items, each with a listening stimulus (either in the form of SGS or human voice actor recording), a single key, and three distractors (four options total). Consequently, the resulting item responses were captured as one of the four options and scored dichotomously into correct or incorrect. Both the selected options and the dichotomous scores are pertinent to test-taker performance and were thus investigated in this study. Specifically, we aimed to estimate the magnitude and practical importance of the impact of using SGS (as opposed to human voice actor recordings) on test takers' response option selection and probability of responding correctly (which we call item response probability in the remainder of this paper). In sum, this study was guided by the following two research questions:

1. What is the impact of using SGS on test takers' response option selection for multiple-choice listening items?
2. What is the impact of using SGS on test takers' item response probability for multiple-choice listening items?

## Methods

### Data Collection Instruments

#### *Listening Tasks*

We evaluated the impact of using SGS with two listening task types, both of which involved multiple-choice items with a single key and three distractors and were considered to be included in the *TOEFL Essentials*™ test (Papageorgiou et al., 2021). The first type is based on a two-turn spoken dialog between two interlocutors. The opening turn by the first interlocutor is either a question or a statement and is presented to test takers via audio only, without any text. Test takers are then asked to select one out of four written options (that are not read out) as the most appropriate response. The second type also involves two interlocutors but is situated in a longer conversation in which each interlocutor takes two or three turns. The entirety of a conversation is presented to test takers via audio only except for the last turn. Then, test takers are presented with two single-selection (out of four written options), multiple-choice items about the conversation. For convenience, the first and second task types will be called the LR (short for "listen and reply") and LC (short for "listen to a conversation") tasks, respectively, in the remainder of this paper.

We used five LR tasks and four LC tasks in the data collection. Each LR task consisted of a single item, whereas each LC task included two items. Therefore, there were five LR items and eight LC items, making the total number of Items 13. All 13 items were prepared with two versions of audio. In one version, the audio stimuli were read by professional voice actors who had extensive experience in reading scripts for English proficiency assessments. The other version involved SGS files based on the same scripts. The process of generating the SGS files is described in detail in the following subsection. For convenience, we will call these two versions the human-voiced version and the SGS version, respectively, in the remainder of this paper. All items were scored dichotomously. For the purpose of this study, we used both the response option selection data (i.e., which option test takers chose) and score data (i.e., whether test takers responded correctly or not).

#### *Preparation of Synthetically Generated Speech Files for Listening Tasks*

We created all SGS files using Amazon Web Services (AWS) Polly and a custom data management pipeline written in Python. AWS Polly is an SGS engine that takes a speech synthesis markup language (SSML) file as input and turns it into an SGS audio file. It provides multiple voice options across ages, accents, and genders. In this study, we utilized adult female and male voices with the North American accent to match the professional voice actors who read the scripts for the human-voiced version. The two SGS voices differed in terms of their default pace and loudness. We used a set of custom pace and loudness for each of the two voices to match those of the corresponding human voice actor.

The custom data management pipeline was used to prepare input for AWS Polly and process its output files. The input preparation stage involved transforming a plain-text script into a proper SSML file such that it could be fed into AWS Polly.

All AWS Polly output files were returned as 16-bit linear PCM files, which we then converted into WAVE audio files. For the LR stimuli, each of which involved only a single sentence (the first turn), we used the resulting WAVE files without further processing. On the other hand, the LC stimuli, which consisted of three or four turns from two interlocutors, required another processing step. Because AWS Polly generates speech from one voice at a time, each turn of the LC stimuli was generated as a separate file. We then concatenated files that belonged to the same LC stimulus (with a 0.7-second pause between turns, as requested by test developers) to form a single SGS file per LC stimulus.

All SGS audio files were reviewed by test developers with extensive experience in writing scripts and evaluating listening stimuli for English proficiency assessments. The reviewers evaluated the SGS files in terms of technical (e.g., sound distortion) as well as linguistic (e.g., stress and intonation) qualities. None of the files were flagged for technical issues, but several were identified as having a somewhat unnatural pause or intonation. Those issues were addressed by making changes to the corresponding SSML files and regenerating the SGS files with the revised SSML files. The resulting SGS files satisfied the reviewers in terms of both technical and linguistic qualities.

### *Other Pilot Tasks*

This study was part of a larger pilot of new task types and their scoring methods for the TOEFL Essentials test.[1] The instrument for the pilot included, in addition to the LR and LC items, several other tasks with different target skills. There were three reading comprehension testlets, each consisting of a passage followed by a few selected response items. Moreover, there were six speaking tasks that elicited a series of spoken responses. Also included were four writing tasks eliciting open-ended responses. The reading, speaking, and writing tasks were all designed to address research questions about the way information is presented (reading) or responses are scored (speaking and writing). Detailed descriptions about the reading, speaking, and writing tasks, as well as those for the associated research questions, are beyond the scope of this paper and are provided elsewhere (Papageorgiou et al., 2021). Last, there were two C-test sets, each of which consisted of a single paragraph with 20 partial words (with only the first two or three characters showing to represent the first half of each partial word) for test takers to fill in. We note that not all of the task types included in the pilot study would become part of the TOEFL Essentials test. Some were included to be a proxy measure for overall English proficiency for the purpose of the pilot study, and others were considered at the time of the study but did not end up being included in the test.

### *Forms*

All the items were divided into two forms, which we will call Forms 1 and 2, respectively, in the remainder of this paper. The forms were assembled such that they would be structurally equivalent but differ in task design aspects relevant to the research questions of the larger study (including those of this study). The total number of items as well as the order of item presentation were thus identical across the two forms. For the purpose of this study, both forms included the full set of 13 listening items (five LRs and eight LCs), but the key difference between the two forms lay in the voices that read the scripts. In Form 1, the SGS versions were used for the LR items, whereas the human-voiced versions were used for the LC items. The version assignment was reversed in Form 2 such that the SGS versions were used only for the LC items. Both forms included the same two C-test sets such that test-taker performance on those sets could serve as a proxy for their overall English proficiency level.

## Participants and Procedures

A total of 653 English learners participated in the pilot as test takers. They were recruited from multiple countries and age groups to ensure a diverse and representative pool in terms of first languages (L1s) and English learning backgrounds. As a result, the test takers came from 10 different countries and spoke 26 different languages. The three largest origin countries were mainland China (226), Korea (90), and Japan (65), and the three largest L1 groups were Chinese (260), Korean (90), and Spanish (61).

All but four test takers (649 out of 653) submitted additional information about their age, education, gender, and English learning experience via a background information questionnaire. Most (561 out of 649) were between 18 and 30 years of age. Eight test takers were younger than 18, and 80 were older than 30. As expected from the age distribution, most test

takers (606 out of 649) were either undertaking their undergraduate (358) or graduate (82) study or had already attained a bachelor's degree (129) or higher (37) at the time of the data collection. There were more female test takers (437) than males (207), and a small number of test takers (5) preferred not to provide their gender information. The English learning experience of the test takers was measured using five bands defined as the following: (a) 2 years or less, (b) 3 – 5 years, (c) 6 – 8 years, (d) 9 – 11 years, and (e) 12 years or more. Progressively more test takers belonged to the later bands, with the numbers of test takers in Bands 1 through 5 being 45, 67, 127, 178, and 232, respectively.

Each test taker was given access to a custom web platform that delivered the background information questionnaire and both forms. The test takers were randomly assigned to one of the two forms and, when they accessed the platform, were automatically routed to the assigned form.[2] The platform was accessible via multiple operating systems and web browsers, including mobile ones. Other than the network bandwidth, the only hardware requirement was a working microphone for the speaking tasks. The platform was also open at all times during the data collection period such that the test takers could access it at their convenience. The test takers were asked to complete their assigned form in one seating and did so in approximately 1 hour.

## Analysis

### *Research Question 1*

To address the first research question (What is the impact of using SGS on test takers' response option selection for multiple-choice listening items?), we compared the test takers' responses to the LR and LC items across the two forms. As mentioned earlier, the two forms were designed such that their LR and LC items differed only in terms of the listening stimulus version. If the use of the SGS version (as opposed to the human-voiced version) affected test-taker performances, that impact would show up as between-form differences in their selection of response options. We thus focused on the comparison between the two forms in terms of the distribution of selected response options.

We summarized the frequency counts of selected response options into a three-way, $13 \times 4 \times 2$ table. The first, second, and third dimensions of this table corresponded to items, response options, and forms, respectively. If test takers' performances were not affected by which listening stimulus version they were assigned to, the item and response option variables would be independent of the form variable; that is, under the hypothesis of no difference in response option selection between the two versions, the probability of any given response option in any given item should be the same across the two forms. This hypothesis corresponds to the log-linear model with all three main effects (items, response options, and forms) and the two-way interaction effect between items and response options. This model, which we call the between-form equiprobability model, is highly restrictive for the three-way table in that it does not have any of the interaction terms involving the form variable. Fitting the between-form equiprobability model to the three-way table thus amounted to a literal and stringent test of the zero impact due to the SGS usage.

### *Research Question 2*

We addressed the second research question (What is the impact of using SGS on test takers' item response probability for multiple-choice listening items?) by comparing item response probabilities across the two versions. The item response probabilities were estimated by fitting two-parameter logistic (2PL) IRT models to the test-taker response data formatted as a $653 \times 66$ score matrix. Each row in this matrix represented a test taker, and each column represented an item. There were 13 items in each version, making the total number of LR and LC items 26. The remaining 40 items were from the two C-test sets (each consisting of 20 items) that were included in both forms.[3] The cells in this matrix took one of the three values: 0 (indicating an incorrect response), 1 (indicating a correct response), and NA (missing). The structure of this score matrix is visually summarized in Figure 1.

The matrix was sparse in that none of the test takers took all 66 items; as mentioned earlier, the test takers were randomly assigned to one of the two forms and responded to only one version of the LR and LC items, by design. Because of this random assignment, the missing data satisfied the assumption of missing completely at random (Rubin, 1976) and ignorable. The C-test items provided the linkage between the two forms, which ensured a common scale for the resulting item response probabilities.

We fit multiple 2PL IRT models representing different levels of equivalence between the two versions. The most flexible model (denoted by IRT.M1) included separate discrimination and difficulty parameters for the same LR or LC item across

| | | Items only in Form 1 | | | | | | Items in Forms 1 & 2 | | | Items only in Form 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LR1S | ... | LR5S | LC1H | ... | LC8H | C1 | ... | C40 | LR1H | ... | LR5H | LC1S | ... | LC8S |
| Form 1 Test Takers | TT1 | 1 | ... | 0 | 1 | ... | 1 | 1 | ... | 1 | NA | ... | NA | NA | ... | NA |
| | TT2 | 0 | ... | 1 | 1 | ... | 1 | 0 | ... | 1 | NA | ... | NA | NA | ... | NA |
| | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | NA | ... | NA | NA | ... | NA |
| | TT335 | 0 | ... | 0 | 0 | ... | 1 | 0 | ... | 0 | NA | ... | NA | NA | ... | NA |
| Form 2 Test Takers | TT336 | NA | ... | NA | NA | ... | NA | 0 | ... | 0 | 0 | ... | 1 | 1 | ... | 1 |
| | TT337 | NA | ... | NA | NA | ... | NA | 0 | ... | 1 | 1 | ... | 0 | 1 | ... | 1 |
| | ... | NA | ... | NA | NA | ... | NA | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| | TT653 | NA | ... | NA | NA | ... | NA | 1 | ... | 1 | 1 | ... | 1 | 1 | ... | 1 |

**Figure 1** The structure of the score matrix. The number following TT represents the test-taker ID number (e.g., TT1 is the first test taker). The letters LR, LC, and C represent the LR, LC, and C-test items, respectively. The number following the task type indicates the item ID (e.g., LR1 is the first LR item). Last, the letters S and H at the end of the LR and LC items represent the synthetically generated speech and human-voiced versions, respectively.

the two versions. The model IRT.M1 thus represented the assumption of no relationship between the two versions. On the other hand, the most constrained model (denoted by IRT.M4) assumed that the two versions of the same item would have identical discrimination and difficulty parameters and thus represented the full equivalence between the two versions. Also considered were two models based on partial equivalence: the equal discrimination model (denoted by IRT.M2) with equal discrimination parameters but separate difficulty parameters across the two versions, and the equal difficulty model (denoted by IRT.M3) with equal difficulty parameters but separate discrimination parameters across the two versions.

The full and partial equivalence models were obtained by imposing equality constraints on the model IRT.M1. Fixing the discrimination and difficulty parameters of the LR and LC items to be equal across the two versions led to the full equivalence model IRT.M4. The models IRT.M2 and IRT.M3 were obtained in a similar manner, with equality constraints on the discrimination and difficulty parameters, respectively.

We compared the estimated item response probabilities at both the overall level and the individual item level. The overall level comparison involved examining the fit of the four models using the likelihood ratio statistics, the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). The result of this comparison would indicate how much impact the full and partial equivalence assumptions had on model fit relative to that of the model IRT.M1. The goal of the item-level analysis was then to help understand where the impact of the full and partial equivalence assumption (in terms of model fit) came from and what the impact meant in terms of item response probabilities. We first examined the estimated item parameters and their standard errors from the model IRT.M1. Specifically, we constructed 95% interval estimates for each of the item parameters and checked whether and how much the resulting intervals overlapped across the two versions. Last, we examined the practical impact of potential differences across the two versions by comparing the item characteristic curves and the test characteristic curves (i.e., the expected sum score based on the 13 items) from the same model IRT.M1.

## Results

### Research Question 1

The three-way, $13 \times 4 \times 2$ table is given in Table 1. The table consists of 104 cells. The between-form equiprobability model, on the other hand, included 52 parameters: 1 intercept term, 12 item main effect terms for the 13 items, 3 response option main effect terms for the four options, and 36 ($12 \times 3$) item by response option interaction terms. Consequently, the model had 52 degrees of freedom. Fitting this model to the table, we obtained the likelihood ratio test statistic (denoted by $L^2$ in the remainder of this paper) of 84.8. This value exceeds the 95th percentile of a chi-square variable with 52 degrees of freedom (which is approximately 69.8), indicating a significant discrepancy between the data and the model.

To understand the source of the model – data discrepancy, we examined adjusted residuals (Haberman, 1978). Figure 2 shows the adjusted residuals grouped by items. Because the model did not include any effect due to the form difference,

**Table 1** Response Option Frequency Counts in a Three-Way Table

| | Form 1 | | | | | Form 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Version | a | b | c | d | Version | a | b | c | d |
| LR1 | SGS | 55 | 46 | 78 | 152 | Human | 45 | 47 | 73 | 150 |
| LR2 | SGS | 45 | 180 | 16 | 93 | Human | 32 | 185 | 10 | 90 |
| LR3 | SGS | 72 | 144 | 61 | 59 | Human | 57 | 203 | 28 | 28 |
| LR4 | SGS | 76 | 27 | 209 | 23 | Human | 63 | 13 | 207 | 32 |
| LR5 | SGS | 26 | 37 | 197 | 75 | Human | 16 | 32 | 196 | 73 |
| LC1 | Human | 23 | 217 | 62 | 34 | SGS | 21 | 191 | 47 | 57 |
| LC2 | Human | 312 | 10 | 10 | 4 | SGS | 305 | 3 | 6 | 2 |
| LC3 | Human | 25 | 79 | 20 | 211 | SGS | 16 | 64 | 15 | 222 |
| LC4 | Human | 4 | 96 | 209 | 25 | SGS | 2 | 83 | 210 | 22 |
| LC5 | Human | 17 | 242 | 55 | 22 | SGS | 17 | 231 | 53 | 16 |
| LC6 | Human | 34 | 20 | 23 | 258 | SGS | 27 | 29 | 22 | 239 |
| LC7 | Human | 8 | 258 | 36 | 32 | SGS | 12 | 228 | 38 | 39 |
| LC8 | Human | 62 | 38 | 214 | 22 | SGS | 48 | 31 | 223 | 15 |

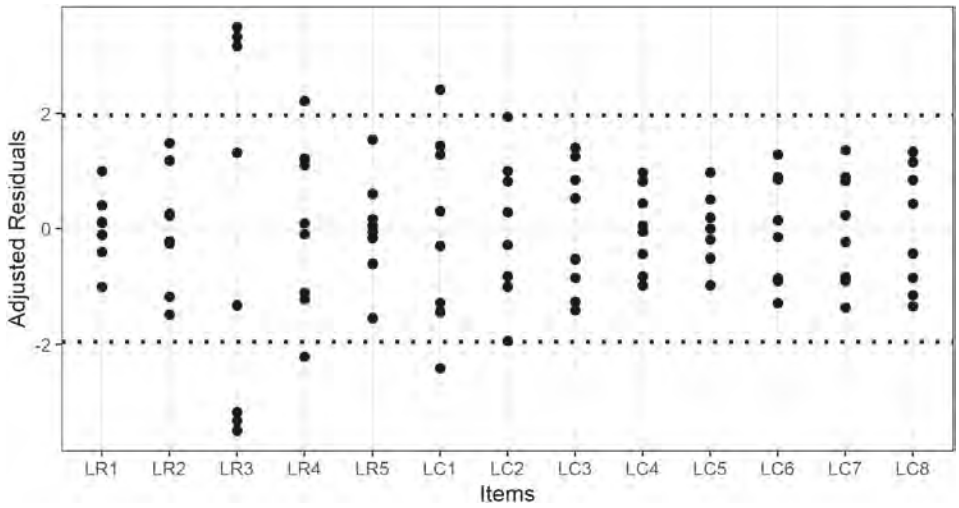


**Figure 2** Adjusted residuals from the between-form equiprobability model. The dashed guidelines correspond to the 2.Fifth and 97.Fifth percentile values of the standard normal distributions.

the adjusted residuals from the two forms were mirror images of one another, with their signs switched. The adjusted residuals asymptotically follow the standard normal distribution and can thus be interpreted against its quantiles. The dashed guideline in Figure 2 corresponds to the 2.5th and 97.5th percentile values of the standard normal distribution. There were 10 adjusted residuals whose values lay outside the 95% interval. If the adjusted residuals were standard normal variables, we would expect to see about five ($104 \times 0.05$) such extreme values, which is fewer than what we observed.

It is clear from Figure 2 that the item that contributed the most to the model – data discrepancy was the third LR item; it was responsible for six of the 10 extreme values, and those six were the most extreme as well. Had it not been for the third LR item, the total number of adjusted residuals outside of the 95% interval would have been four, which is in line with what's expected out of 96 (104 – 8) standard normal variables. We thus tested whether it would be plausible to attribute the model misfit solely to the third LR item by fitting the same between-form equiprobability model to the $12 \times 4 \times 2$ subtable with the 12 remaining items. The resulting $L^2$ statistic became 49.1 with 48 degrees of freedom, which was much smaller than the corresponding 95th percentile value of 65.2 ($p \approx .57$). This indicates that the between-form equiprobability model accounts well for the data for the remaining 12 items and thus isolates the third QC item as the source of the misfit.

We then examined in detail the response patterns from the third LR item. Table 2 gives the proportions of the four response options within each form. This item was quite difficult (in fact, the most difficult among the 13 items) in that only about one fifth of the test takers responded correctly. The correct response proportions were comparable across the

**Table 2** The Proportions of the Four Response Options on the Third QC Item Within Each of the Two Forms

| | Form 1 (SGS) | Form 2 (human) |
|---|---|---|
| a | 0.21 | 0.18 |
| b | 0.43 | 0.64 |
| c | 0.18 | 0.09 |
| d | 0.18 | 0.09 |

**Table 3** Global Fit Results From the IRT Models

| | | | | Comparison to IRT.M4 in terms of $L^2$ | | |
|---|---|---|---|---|---|---|
| | AIC | BIC | $L^2$ | $L^2$ Difference | Deg. of freedom difference | $p$-value |
| IRT.M4 | 32,168.03 | 32,643.08 | 31,956.03 | — | — | — |
| IRT.M3 | 32,168.05 | 32,701.36 | 31,930.05 | 25.99 | 13 | .02 |
| IRT.M2 | 32,174.02 | 32,707.33 | 31,936.02 | 20.01 | 13 | .09 |
| IRT.M1 | 32,180.19 | 32,771.76 | 31,916.19 | 39.84 | 26 | .04 |

two forms. The adjusted residuals corresponding to these correct response cells were within the 95% interval, indicating that the observed difference between the correct-response cells could have come from chance alone and was not the source of the model misfit. To formally test this observation, we fit the between-form equiprobability model to the original data from all 13 items with the wrong response options (three per item) collapsed into the single "wrong" category (yielding a $13 \times 2 \times 2$ table). The resulting $L^2$ statistic was 25.77, with 26 degrees of freedom (52 cells with 1 intercept, 12 item parameters, 1 response parameter, and 12 item by response parameters), which was much smaller than the corresponding 95th percentile value of 38.89 ($p \approx .52$) and thus added statistical backing for the observation.

The major difference between the two forms came from the most popular distractor (Option B). In Form 2 (the human-voiced version), that option was chosen by more than 60% of the test takers, whereas the corresponding proportion in Form 1 (the SGS version) was about 20 percentage points lower. Those who chose neither the key nor the most attractive distractor were evenly split between the other options across the two forms. The differential attractiveness of the third LR item's most popular distractor was thus the sole source of the overall model–data discrepancy.

## Research Question 2

The global fit results of the four IRT models are summarized in Table 3. Under the most constrained model IRT.M4, the 13 LR and LC items were assumed to have identical discrimination and difficulty parameters across the two versions. Therefore, there were 26 ($13 + 13$) more degrees of freedom under that model compared to those under the model IRT.M1 without any equality constraints. Similarly, the partial equivalence models IRT.M2 and IRT.M3 had 13 more degrees of freedom than the model IRT.M1 because of their equality constrains on the 13 discrimination and difficulty parameters, respectively.

The comparison between the most flexible IRT.M1 and the most constrained IRT.M4 led to a disagreement between the $L^2$ and AIC/BIC statistics. Because the AIC and BIC statistics always agreed, we focus on the AIC statistics in the remainder of this section. The fit difference between the two models yielded the $L^2$ statistic of 39.8, which was slightly larger than the 95th percentile value of a chi-square variate with 26 degrees of freedom (38.9). On the other hand, the model IRT.M4 was associated with the smaller AIC value. This disagreement has to do with the difference between what the two statistics measure. The $L^2$ statistic is a summary of a model's fit to the data at hand, whereas the AIC is an estimate of a model's performance on a future data set (i.e., another random sample from the same population). The disagreement thus indicates that the advantage of relaxing the full equivalence assumption was noticeable in the current data set but not large enough to ensure the same on a future data set.

The other comparisons in Table 3 provide information on whether the impact on model fit came from the equivalence assumptions involving the difficulty parameters, the discrimination parameters, or both, respectively. The comparison between the models IRT.M4 and IRT.M3 represented a test of the equivalence assumption on the difficulty parameters.

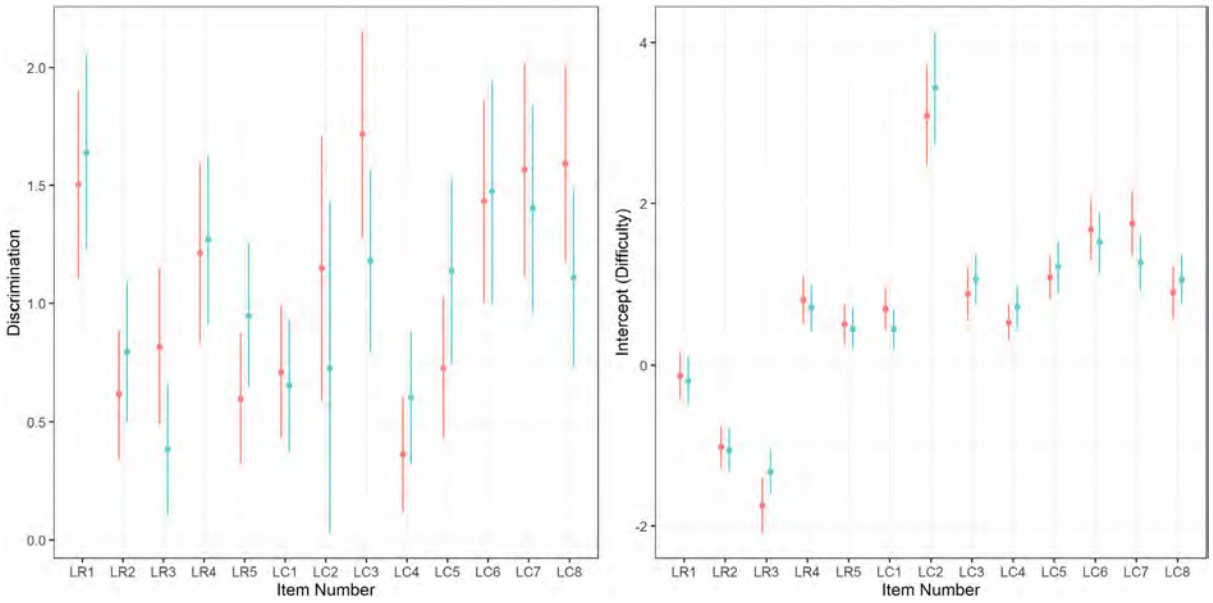


**Figure 3** Item-level comparisons between the versions in terms of interval estimates from the model IRT.M1. For both panels, the estimates from the human-voiced versions are represented with the red dots and lines, and those from the synthetically generated speech versions are represented with the green dots and lines.

The model IRT.M3 (without the equal difficulty assumption) yielded a significantly better fit than the model IRT.M4 did in terms of the $L^2$ statistic, although the former was preferred over the latter in terms of the AIC. The next comparison, between the models IRT.M4 and IRT.M2, had to do with the equivalence assumption on the discrimination parameters. Here, the two statistics led to the same conclusion: The equal discrimination assumption did not affect the overall model fit. These results suggest that the impact on the model fit (in terms of the $L^2$ statistics) resulting from the full equivalence assumption was mainly due to the potential version differences in item difficulty.

We complemented the above global fit comparisons with comparisons between estimated parameters at the individual item level. Specifically, we examined the interval estimates of the discrimination and intercept parameters from the model IRT.M1 (in which no equality assumption was made across the two versions), which are presented visually in Figure 3.[4] The left panel shows the estimated 95% confidence intervals of the discrimination parameters, and the right panel provides the same information for the intercept parameters. The estimated 95% confidence intervals overlapped across all item pairs for both parameters. This indicates that, at the individual item level, there was no single item (or a small number of items) whose parameters differed substantially between the two versions. This also suggests the lack of any large violations from the full equivalence assumption.

We also examined the item and test characteristic curves of the human-voiced and SGS versions from the model IRT.M1. As expected from the overlapping interval estimates, the item characteristic curves from the two versions were highly comparable across all 13 items. Consequently, the two test characteristic curves were also highly comparable, as can be shown in Figure 4. They differed by no more than 0.19 across the entire range from −4 to 4. This maximum difference occurred at the proficiency value of −4, which is a highly unlikely value (3 in 100,000 would have −4 or below under the standard normal distribution). In our sample, no test taker was estimated to be at or below that level. The proficiency range among the 653 test takers was from −2.7 to 2.5 (two vertical dotted lines in Figure 4). Within that range, the maximum absolute difference between the two curves was 0.1. This difference amounted to less than 1% of the scale size (from 0 to 13). This result indicates that, even under the model without any equality assumption across the two versions, the practical impact of the version on a test taker's score was trivial.

## Discussion and Implications

We addressed the two research questions by comparing the patterns in response option selection and response probabilities across the human-voiced and SGS versions of 13 multiple-choice single-selection listening items. The log-linear model

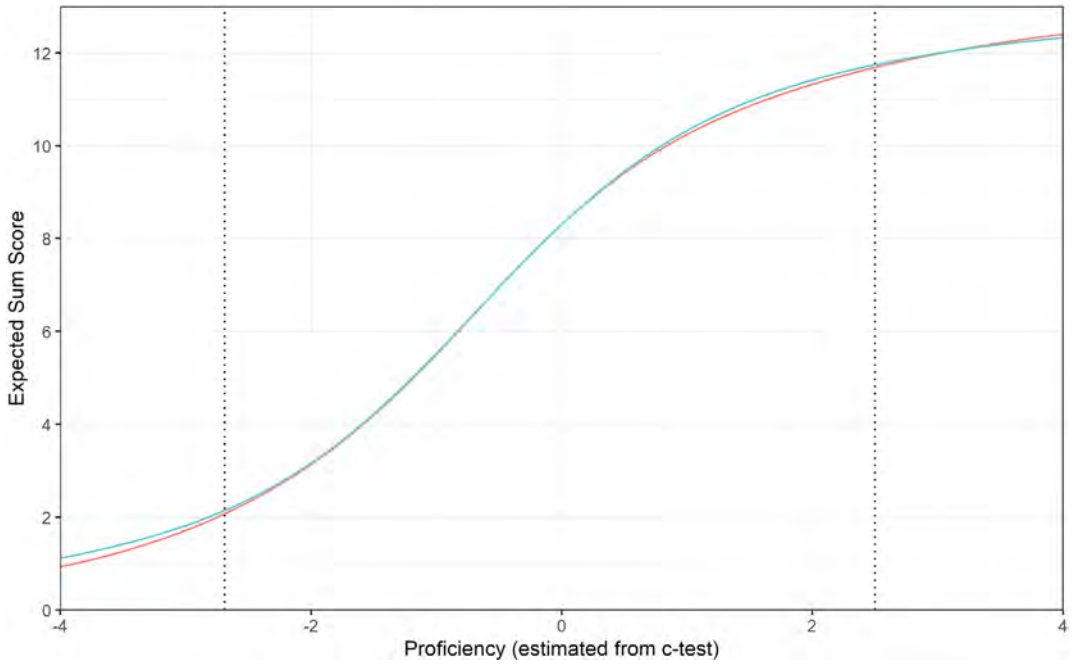


**Figure 4** Comparison between the test characteristic curves across the two versions from the model IRT.M1. The red curve is from the human-voiced version, and the green curve is from the synthetically generated speech version. The minimum and maximum proficiency levels from the sample are represented with the vertical dotted lines.

assuming the equiprobability of item response options across the two versions successfully accounted for the response option selection patterns of 12 out of the 13 items. The sole exception came from the third LR item, which was the most difficult item. Specifically, its most popular distractor was more attractive under the human-voiced version than under the SGS version. As for the second research question, the model IRT.M4, which assumed the full equivalence in the item parameters across the two versions, was preferred over the model IRT.M1 without any equality assumption in terms of the AIC and BIC statistics. On the other hand, the full equivalence assumption led to a significantly worse fit in terms of the $L^2$ statistics, which motivated several additional analyses of item response probabilities. Although we identified the potential difference in the difficulty parameters across the two versions at the global level, we did not find any item that differed significantly across the versions. Moreover, the practical impact of the version difference, even under the model IRT.M1, was negligible. The test characteristic curves from the two versions did not differ more than 0.1 across the entire range, which amounted to less than 1% of the score scale.

The natural question around the third LR item was whether the differential attractiveness of the most popular distractor could be attributed to the version difference. Its SGS stimulus was not flagged for any issues during the original review process. We reviewed its human-voiced and SGS stimuli again after observing it as the sole source of departure from the full equivalence, but even then failed to find any meaningful difference that could have caused the differential attractiveness. Specifically, we did not notice any unnatural or unexpected segmental or suprasegmental features in either version. There may be complicated interactions between the version and the distractor. It is also possible that the general difficulty of estimating multinomial probabilities had to do with this phenomenon. Without additional data, we currently do not have an empirical way of differentiating among these multiple possibilities. We thus view the cause of the differential attractiveness of the most attractive distractor inconclusive at this point.

Another potential discrepancy from the full equivalence between the two versions had to do with the item difficulty parameters. The equal difficulty assumption made a noticeable impact on the global model fit in terms of the $L^2$ statistics. However, subsequent item-level analyses did not find any item showing a major departure from that assumption. We also observed that the practical impact of the departure on the total score was negligible. Therefore, we believe that the multiple pieces of evidence for the between-version equivalence, including the global model fit measured by the AIC/BIC and the item-level analysis results, outweigh those that came from the global fit comparison in terms of the $L^2$ statistic.

Although we thoroughly examined a small number of departures from the full equivalence between the human-voiced and SGS versions, the predominant finding was the remarkable comparability between the two versions. It is worthwhile to reiterate the restrictive nature of the models used to test the equivalence. The underlying assumption for both the between-form equiprobability log-linear model and the full equivalence model (IRT.M4) was that the test-taker performances on the two versions were, for all intents and purposes, identical. This is an extraordinarily difficult assumption to hold in practice. Despite this restrictive assumption, the between-form equiprobability model fit all but one item well, and the full equivalence model IRT.M4 was preferred in terms of the AIC and the BIC over the most flexible model IRT.M1 that included 26 more parameters. Moreover, subsequent investigations into the departures did not show any clear indication against the comparability between the two versions. We believe that all these findings point to the same conclusion: The two versions are indeed highly comparable in terms of test-taker performance.

The comparability between the two versions has direct implications for the development of listening proficiency assessment content. Currently, most standardized language assessments rely on human voice actors for listening stimuli. This is not only expensive and time-consuming but also makes it difficult to scale and/or personalize the content. For example, when every listening stimulus needs to be prerecorded by human voice actors, even a simple personalization activity, such as addressing test takers by their name, can be prohibitively difficult to implement. With SGS, assessment developers can have listening stimuli for their content instantly without any time lag due to the scheduling, recording, and editing activities required for human voice actors. Moreover, they can, without worrying about prerecording everything, come up with personalization devices that help establish an environment in which test takers can perform at their best. Our findings provide empirical evidence that such promising possibilities can be realized without introducing construct-irrelevant variance in test-taker performance.

We believe that the comparability has implications beyond the immediate domain of assessment development. A closely related domain that can also benefit from adopting SGS is learning content development. As researchers have pointed out (e.g., Kılıçkaya, 2006; Liakin et al., 2017a; Sha, 2010), SGS has multiple advantages in this domain. The low cost, immediacy, and scalability of SGS are certainly relevant and can be attractive to learning content developers. The personalization potential we discussed above is arguably more important in the learning content development domain, as backed by a well-established literature on the benefits of personalization in learning content (see Walkington & Bernacki, 2020, for a recent review of this literature, and Chapelle, 2001, for the importance of personalized technology in L2 learning). Another major advantage of SGS lies in the potential for the fine-grained control of output. SGS allows a user to generate speech with varying levels of pace or emphasis. This level of control available with SGS may be an additional benefit for learning content developers as well as for researchers who design instruments for experiments. There have already been studies that utilized SGS for research instrument development (e.g., Ji et al., 2019), and we believe that our findings provide additional empirical justifications for such studies.

## Conclusions

In this study, we investigated the impact of using SGS on test-taker performance. The focus of the investigation was on multiple-choice, single-selection items, which is one of the most popular task types in L2 listening proficiency assessments. The data came from a pilot administration of multiple new task types and included test-taker responses to two versions of the same 13 items, differing only in terms of their listening stimuli: a version using human voice actor recordings and the other version with SGS files. Multifaceted comparisons between test-taker responses across the two versions showed that the two versions elicited remarkably comparable performance. The comparability provides strong empirical evidence for the use of SGS as a viable alternative for human voice actor recordings in the immediate domain of language assessment as well as related domains such as learning material and research instrument development.

We acknowledge that this study was limited in terms of its scope and size. We focused on multiple-choice, single-selection items with short listening stimuli (ranging from one to four turns). There are many other task types that elicit different types of responses (e.g., multiple selection, matching, ordering, constructed responses) and can involve longer listening stimuli (e.g., announcements, lectures). Moreover, the per-version sample size of this study (∼300) was relatively small to accurately estimate the item parameters separately for each version. These limitations provide natural starting points for future research. A clear next step is to extend the scope to different task

types involving longer listening stimuli, ideally at a larger scale, to allow even more accurate estimation of all item parameters.

We view this study as a first step toward a comprehensive set of investigations into how the use of SGS affects test takers and how learners interact with language tasks. Our findings suggest that a current state-of-the-art SGS implementation already has a clear use case. The underlying technology is also continuously improving at a rapid pace. Even between the time of the pilot data collection and the time of this writing, our SGS engine of choice (AWS Polly) has added new voices that, had they existed at the time of the pilot, would have been considered for use for the purpose of this study. We expect that such improvements will continue and bring about more and more use cases. There is thus a strong case for continuous research efforts to gather empirical evidence for (or against) the use of SGS across the growing range of L2 assessment and learning use cases.

## Acknowledgments

We thank Larry Davis and John Norris for their helpful feedback on the study design, and Claudia Hauck and Shuhong Li for generously sharing cleaned and prepared data. We also thank John Davis, Spiros Papageorgiou, and Tongyun Li for their careful review of an earlier version of the manuscript and their helpful comments. Any remaining errors are, of course, our own.

## Notes

1 Because a detailed description of the pilot study and the test is available elsewhere (Papageorgiou et al., 2021), we only describe in this subsection what we consider necessary to make this paper self-contained.

2 As expected from this random assignment, the two groups of test takers (grouped by their form assignment) were highly comparable in terms of their background information as well as their scores on the two common C-test sets. In particular, the two groups differed by less than a half score (27.1 vs. 27.5) in their respective means and yielded almost identical first (22 vs. 23), second (28.5 vs. 28), and third (34 vs. 34) quartiles.

3 This concurrent calibration is based on the unidimensionality assumption that all three task types are measuring the same construct. Although it is difficult to argue that the C-test items measure exactly the same construct as the LR and LC items do, there is theoretical (e.g., Babaii & Ansary, 2001) and empirical (e.g., McKay et al., 2021) evidence that establishes the effectiveness of the C-test items as a measure of general language proficiency. We also examined the factor structure of both forms using the single-factor confirmatory model and observed acceptable fit in both cases (RMSEA of 0.058 and 0.059 for Forms 1 and 2, respectively). We thus believe that this assumption would not have any major practical impact on our findings.

4 The differences in the intercept point estimates between the two versions ranged from −0.5 to 0.4 with mean 0.02 and standard deviation 0.25. The corresponding numbers for the differences in the discrimination point estimates were − 0.54 (minimum), 0.41 (maximum), −0.05 (mean), and 0.33 (standard deviation).

## References

Azuma, J. (2008). Applying TTS technology to foreign language teaching. In F. Zhang & B. Barber (Eds.), *Handbook of research on computer-enhanced language acquisition and learning* (pp. 497–506). Information Science Reference. https://doi.org/10.4018/978-1-59904-895-6.ch029

Babaii, E., & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System*, *29*(2), 209–219. https://doi.org/10.1016/S0346-251X(01)00012-4

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Betz, S., Carlmeyer, B., Wagner, P., & Wrede, B. (2018). Interactive hesitation synthesis: Modelling and evaluation. *Multimodal Technologies and Interaction*, *2*(1), 9–29. https://doi.org/10.3390/mti2010009

Bione, T., Grimshaw, J., & Cardoso, W. (2016). An evaluation of text-to-speech synthesizers in the foreign language classroom: Learners' perceptions. In S. Papadima-Sophocleous, L. Bradley, & S. S. Thouësny (Eds.), *CALL communities and culture – Short papers from EUROCALL 2016* (pp. 50–54). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.eurocall2016.537

Cardoso, W. (2018). Learning L2 pronunciation with a text-to-speech synthesizer. In P. Taalas, J. Jalkanen, L. Bradley., & S. Thouësny (Eds.), *Future-proof CALL: Language learning as exploration and encounters—Short papers from EUROCALL 2018* (pp. 16–21). Research-publishing.net. https://doi.org/10.14705/rpnet.2018.26.806

Cardoso, W., Smith, G., & Garcia Fuentes, C. (2015). Evaluating text-to-speech synthesizers. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds.), *Critical CALL—Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 108–113). Research-publishing.net. https://doi.org/10.14705/rpnet.2015.000318

Chapelle, C. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research*. Cambridge University Press.

Crocker, L. M., & Algina, J. (2006). *Introduction to classical and modern test theory*. Holt, Rinehart, & Winston.

De Silva, P., Wattanavekin, T., Hao, T., & Pipatsrisawat, K. (2018). Voice builder: A tool for building text-to-speech voices. *Proceedings of the 11th International Conference on Language Resources and Evaluation*, 2241–2245. LREC.

Delmonte, R. (2008). Speech synthesis for language tutoring systems. In M. Holland & P. Fisher (Eds.), *The path of speech technologies in computer-assisted language learning: From research towards practice* (pp. 123–150). Routledge.

Elias, I., Zen, H., Shen, J., Zhang, Y., Jia, Y., Weiss, R. J., & Wu, Y. (2020). *Parallel Tacotron: Non-autoregressive and controllable TTS* [arXiv preprint]. IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414718

Euler, L. (1761). The wonders of the human voice. In D. Brewster (Ed.), *Letters of Euler on different subjects in natural philosophy addressed to German princess* (pp. 76–79). J. J. Harper.

Georgila, K. (2017). Speech synthesis: State of the art and challenges for the future. In J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic., & A. Vinciarelli (Eds.), *Social signal processing* (pp. 257–272). Cambridge University Press. https://doi.org/10.1017/9781316676202.019

Haberman, S. J. (1978). *Analysis of qualitative data* (Vol. 1). Academic Press.

International English Language Testing System (IELTS®). (n.d.) *Visual difficulties*. https://www.ielts.org/en-us/for-test-takers/special-requirements/visual-difficulties

International Telecommunication Union. (2016). *ITU-T P.800.1 Mean opinion score (MOS) terminology*. www.itu.int/rec/T-REC-P.800.1-201607-I

Ji, W., Liu, R., & Lee, S. (2019). Do drivers prefer female voice for guidance? An interaction design about information type and speaker gender for autonomous driving car. In H. Krömker (Ed.), *HCI in mobility, transport, and automotive systems* (pp. 208–224). Springer. https://doi.org/10.1007/978-3-030-22666-4_15

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kataoka, H., Funakoshi, Y., & Kitamura, Y. (2007). Utilizing text-to-speech synthesis for English education. *Proceedings of the 23rd Annual Conference of Japan Society for Educational Technology*, 429–430. JSET.

Kılıçkaya, F. (2006). Text-to-speech technology: What does it offer to foreign language learners? *CALL-EJ Online*, *7*(2), 17–24.

Kılıçkaya, F. (2011). Improving pronunciation via accent reduction and text-to-speech software. In M. Levy., F. Blin, C. B. Siskin, & O. Takeuchi (Eds.), *WorldCALL: International perspectives on computer-assisted language learning* (pp. 85–96). Routledge.

Kirstein, M. (2006). *Universalizing universal design: Applying text-to-speech technology to English language learners' process writing* [Unpublished doctoral dissertation]. University of Massachusetts.

Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2021). *DiffWave: A versatile diffusion model for audio synthesis*. ArXiv. https://arxiv.org/abs/2009.09761v3

Kraft, V., & Portele, T. (1995). Quality evaluation of five German speech synthesis systems. *Acta Acustica*, *3*, 351–365.

Lewis, J. R. (2001). *The revised mean opinion scale (MOS-R): Preliminary psychometric evaluation* (IBM Technical Report TR 29.3414). IBM Voice Systems.

Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. (2019). Neural speech synthesis with transformer network. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 6706–6713. AAAI. https://doi.org/10.1609/aaai.v33i01.33016706

Liakin, D., Cardoso, W., & Liakina, N. (2017a). The pedagogical use of mobile speech synthesis (TTS): Focus on French liaison. *Computer Assisted Language Learning*, *30*(3–4), 325–342. https://doi.org/10.1080/09588221.2017.1312463

Liakin, D., Cardoso, W., & Liakina, N. (2017b). Mobilizing instruction in a second-language context: Learners' perceptions of two speech technologies. *Language*, *2*(3), 11–32. https://doi.org/10.3390/languages2030011

McKay, T. H., Teimouri, Y., Sağdıç, A., Salen, B., Reagan, D., & Malone, M. E. (2021). The cagey C-test construct: Some evidence from a meta-analysis of correlation coefficients. *System*, *99*, 102526. https://doi.org/10.1016/j.system.2021.102526

Ning, Y., He, S., Wu, Z., Xing, C., & Zhang, L-J. (2019). A review of deep learning based speech synthesis. *Applied Science*, *9*, 4050–4065. https://doi.org/10.3390/app9194050

Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the* TOEFL Essentials™ *test* (Research Memorandum No. RM-21-03). ETS.

Proctor, C. P., Dalton, B., & Grisham, D. L. (2007). Scaffolding English language learners and struggling readers in a universal literacy environment with embedded strategy instruction and vocabulary support. *Journal of Literacy Research*, *39*(1), 71–93. https://doi.org/10.1080/10862960709336758

Qian, M., Chukharev-Hudilainen, E., & Levis, J. M. (2018). A system for adaptive high-variability segmental perceptual training: Implementation, effectiveness, transfer. *Language Learning & Technology*, *22*(1), 69–96. https://dr.lib.iastate.edu/bitstreams/b2bedbeb-fdcb-4db7-be7e-4400cdea1297/download

Rosenberg, A., & Ramabhadran, B. (2017). Bias and statistical significance in evaluating speech synthesis with mean opinion scores. *Proceedings of Interspeech*, 3976–3980. https://doi.org/10.21437/Interspeech.2017-479

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Sha, G. (2010). Using TTS voices to develop audio materials for listening comprehension: A digital approach. *British Journal of Educational Technology*, *41*(4), 632–641. https://doi.org/10.1111/j.1467-8535.2009.01025.x

Shadiev, R., Hwang, W. Y., & Liu, T. Y. (2018). A study of the use of wearable devices for healthy and enjoyable English as a foreign language learning in authentic contexts. *Journal of Educational Technology & Society*, *21*(4), 217–231.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z, Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions. IEEE. *International Conference on Acoustics, Speech and Signal Processing*, 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368

Soler Urzúa, F. (2011). *The acquisition of English /ɪ/ by Spanish speakers via text-to-speech synthesizers: A quasi-experimental study* [Unpublished master's thesis]. Concordia University.

Sonntag, G.P., Portele, T., Haas, F., & Kohler, J. (1999). Comparative evaluation of six German TTS systems. *Proceedings of Eurospeech*, 251–254. Technical University of Budapest.

Story, B. H. (2019). History of speech synthesis. In W. F. Katz & P. F. Assmann (Eds.), *The Routledge handbook of phonetics* (pp. 9–32). Routledge. https://doi.org/10.4324/9780429056253-2

Viswanathan, M., & Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech and Language*, *19*, 55–83. https://doi.org/10.1016/j.csl.2003.12.001

Volodina, E., & Pijetlovic, D. (2015). Lark trills for language drills: Text-to-speech technology for language learners. *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, 107–117. https://doi.org/10.3115/v1/W15-0613

Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z, Székely, É., Tånnander, C., & Voße, J. (2019). Speech synthesis evaluation—State-of-the-art assessment and suggestion for a novel research program. *Proceedings of the 10th Speech Synthesis Workshop*, 105–110. https://doi.org/10.21437/SSW.2019-19

Walkington, C., & Bernacki, M. L. (2020). Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions. *Journal of Research on Technology in Education*, *52*, 235–252, https://doi.org/10.1080/15391523.2020.1747757

Wester, M., Valentini-Botinhao, C., & Henter, G. E. (2015). Are we using enough listeners? No!—An empirically-supported critique of Interspeech 2014 TTS evaluations. *Proceedings of Interspeech 2015*, 3476–3480. https://doi.org/10.21437/Interspeech.2015-689

## Suggested citation:

Choi, I., & Zu, J. (2022). *The impact of using synthetically generated listening stimuli on test-taker performance: A case study with multiple-choice, single-selection items* (TOEFL Research Report No. RR-98). ETS. https://doi.org/10.1002/ets2.12347

**Action Editor:** John Davis

**Reviewers:** Spiros Papageorgiou and Tongyun Li