# *Second Language Studies*

Volume 40 Issue 1

**TABLE OF CONTENTS**

# *SECOND LANGUAGE STUDIES* – PLANNING A TRANSITION

**DUSTIN CROWTHER**

***University of Hawai'i at Mānoa***

On behalf of the editorial team for *Second Language Studies*, we are pleased to share our 2022 issue, the largest yet under my time as editor! As University of Hawai'i transitions back towards normality with the lifting of pandemic restrictions, we are beginning to see once again the range of research possibilities available to us in the field of second language studies. Accompanying this move towards greater normality in our teaching and research practices comes an exciting transition for how *Second Language Studies* will be run moving forward.

## A STUDENT DRIVEN PUBLICATION

In 2020, I took over as editor of *Second Language Studies* from the recently retired Dr. JD Brown, with the goal to eventually give journal ownership over to the graduate student body of the Department of Second Language Studies. Since 2020, I have worked closely with three SLS PhD students, Ha Nguyen, Micah Mizukami, and Kristen Urada, with three more SLS PhD students joining us for the current issue: Ricky Larkin, Michol Miller, and Hitoshi Nishizawa. With an editorial board of six graduate students, all of whom have contributed to a growing vision for the journal, I feel that now is an appropriate time for the students of SLS to take control of the journal. While remaining as a journal advisor, I believe that the opportunity to run *Second Language Studies* will provide both current and future SLS student editors the opportunity to more fully experience and comprehend the publication process in a stress-free environment. This is in addition to maintaining a local outlet for student and faculty work on a range of topics relevant to second language studies. In the coming year, your new student editorial board will provide additional information on their mission statement and plans for future publication procedures!

## THE CURRENT ISSUE

For our fall 2022 issue, we present four papers, representing scholarly work at the MA, PhD, and faculty levels of the Department of Second Language Studies. Building off of previous issues, we now have work not only representative of keystone projects in SLS studies (i.e., MA scholarly papers, PhD dissertations), but also class level projects, highlighting the quality of SLS work being produced in classes such as SLS 490: *Second Language Testing* and SLS 671: *Research in Language Testing*. In addition to these four papers, we once again provide an overview of recent SLS MA/AGC scholarly papers and PhD dissertations.

Our first paper is a collaborative piece from SLS PhD students Yu-Tzu Chang, Ann Tai Choe, Dan Holden, and SLS faculty member Dan Isbell. Their paper, which developed out of an SLS 671 project, investigated the appropriateness of previous revisions made to the Hawai‘i English Language Program's (HELP) writing placement test, specifically in reference to the rubric used for assessing students' writing performance on the test.

The second paper, written by SLS faculty member Dustin Crowther and PhD student Kristen Urada, is an example of research adapting to restrictions due to COVID-19. In their paper, they compared listeners' perceptions of speech elicited in a face-to-face versus online setting, with the online speech samples originally collected as part of a pedagogically-oriented study postponed due to COVID-19 (see paper for more details).

The third paper is a test review written by Hikaru Ishiyama, a 2021-2022 SLS MA graduate. This paper, developed in SLS 490, reviewed the validity of the writing section in a well-used English proficiency exam in Japan, the *Eiken Test in Practical English Proficiency*.

Our final paper comes from Soo Jin Lim, another 2021-2022 SLS MA graduate, who considered the role of English kindergarten programs in South Korean. More specifically, she investigated the factors that influenced parents' decisions on which type of English kindergarten parents chose to enroll their children in. This paper served as her MA scholarly paper.

On behalf of our growing editorial board, we once again thank you for reading, and hope that you enjoy the 2022 issue of *Second Language Studies*. For those looking to publish in *Second Language Studies*, we are always open to submissions, including in-progress research, scholarship directly relevant to Hawai‘i and the Asia-Pacific region, needs analyses focused on

UH Mānoa-based language programs, theoretical papers, and other submissions relevant to second language studies!

Mahalo!

Dustin Crowther
Rickey Larkin
Michol Miller
Micah Mizukami
Thu Ha Nguyen
Hitoshi Nishizawa
Kristen Urada

# AN EVALUATION OF HELP'S WRITING PLACEMENT TEST RUBRIC USING MANY-FACET RASCH ANALYSIS

**YU-TZU CHANG, ANN TAI CHOE, DAN HOLDEN, AND DAN ISBELL**
***University of Hawai'i at Mānoa***

## ABSTRACT

This study is part of a larger project that explores the ongoing revisions of a writing rubric used in the Hawai'i English Language Program (HELP) at the University of Hawai'i at Mānoa. Previously, HELP revised its writing placement test, changing from an 8-category writing rubric with a 4-point scale ("old rubric," 2010–2016) to a 4-category writing rubric with a 20-point scale ("new rubric," 2017–2021). Because the rationale for revising the old rubric was based on anecdotal evidence (see Rock, 2016), whether the revision was justified from an empirical standpoint remains unknown. In this study, we aim to investigate the empirical support for the motivations and outcomes of the rubric revision using many-facet Rasch measurement (MFRM). Findings revealed that the old rubric was reliable in distinguishing test taker writing ability, but there were several critical issues with its functioning: (1) a redundancy of scale points, (2) rating criteria redundancy, and (3) a high number of misfitting raters. These findings lend support for HELP's previous decision to revise the old rubric. Implications for rubric revision are discussed.

## INTRODUCTION

Rubric revision and validation for performance-based language assessments are motivated by concerns toward reliability, measurement, and uses of scores for decision-making (Knoch & Chapelle, 2018). Compared to large-scale assessment contexts, however, test developers in local settings often have fewer resources and rely more on expert intuition or anecdotal evidence to inform revision decisions. Although the stakes are not as high as most large-scale tests, the consequences of local tests are not trivial, and may involve decisions such as placement in a

language program. In accordance with these stakes, validity evidence must not be in short supply (Kane, 2013). This study reports on efforts to validate a local writing placement test used in the Hawai‘i English Language Program (HELP) at the University of Hawai‘i at Mānoa (UHM). In 2016, HELP's writing placement exam underwent a substantial revision, sparked by input from and efforts of program administrators and teachers (who also served as raters; see Rock, 2016). As researchers affiliated with the university and HELP, we followed up on those test development efforts by carrying out systematic validation research using many-facet Rasch measurement (MFRM). Our aim is to show how rigorous validation techniques can be used to evaluate revisions made to local tests and motivate further changes.

## BACKGROUND

### *Decision Making in Rubric Design*

Test developers have a myriad of choices when designing or revising a test. In the context of performance-based assessments, these choices may include selecting the type of rating scale (holistic or analytic) and deciding the number of evaluation categories and score points to appear in the rubric. These choices concerning different aspects of the rating scale design must be carefully considered in order to ensure that the scoring procedure is consistent and appropriate for the assessment's purpose and the needs of the program. Take writing placement exams in local assessment contexts as an example. If practicality is a major concern due to a shortage of time or human resources, a holistic rating scale is more advantageous than an analytic rating scale because raters only need to assign a single score rather than parsing an essay several times to focus on different aspects of writing quality. Choosing a holistic rating scale would therefore reduce a significant amount of time needed to score the essays and place students into their corresponding placement levels. As Weigle (2002) cautioned, however, there are notable disadvantages of holistic rating scales, including the fact that a single score does not generate sufficient diagnostic information about a candidate's ability (e.g., grammar, vocabulary, and organization), and that the score may be less interpretable because raters may draw on different criteria in assigning the score. On the other hand, while analytic scoring may be more time-consuming, it can provide rich information about various aspects of a candidate's writing ability. As areas of second language (L2) writing tend to develop at different rates, analytic scoring is

potentially useful for making placement and pedagogical decisions in language programs. Another advantage of using analytic scoring is that reliability tends to increase when an essay is assigned with multiple scores (Weigle, 2002). Therefore, if reliability and accuracy are the primary concerns, adopting an analytic scoring rubric would generate reliable results about various aspects of a student's writing ability to inform placement or pedagogical decisions.

### *Rubric Revision and Validation*

Across both large- and small-scale assessment contexts, the quality of the test needs to be carefully validated. Validation is an iterative process as test developers accumulate new evidence from different stages of development to make revision decisions (Bachman & Palmer, 2010; Kane, 2013). Motivation for rubric revisions can arise from concerns raised in validation research related to the construct, measurement qualities, and the impact of using the rubric for decision making. Validation studies adopting the argument-based approach (Kane, 2006, 2013; Knoch & Chapelle, 2018) have conceptualized these concerns in terms of inferences, warrants, and assumptions. Each inference is specified by a claim, and each claim is evaluated based on warrants and a set of associated assumptions. The assumptions should be supported by evidence known as backing, which could be quantitative (e.g., statistical analysis, many-facet Rasch analysis) and/or qualitative (e.g., interviews with raters, analysis of rating sessions).

In our study, we focus on two inferences for rater-mediated assessments: *evaluation* and *generalization*. According to Knoch and Chapelle (2018), the evaluation inference (also known as "scoring," following Kane, 2006, 2013) "is justified if observations on assessment are evaluated using procedures that provide observed scores with intended characteristics" (p. 482). Concerns regarding the construct and measurement qualities can thus be conceptualized under the evaluation inference, with underlying warrants pertaining to rubric properties and measurement-related practices such as rater consistency and scale functioning. When evaluating the properties of an analytic scale, assumptions associated with the warrant include examining whether the scale steps are adequate in distinguishing among ability levels; assumptions to support the warrant on scoring could include raters' use of the scale in identifying distinct levels of performance, and rater consistency, especially when scores cannot practically be adjusted. Whereas the evaluation inference centers on scale properties and rater performance at the task level, the generalization inference focuses these issues at the test level and is warranted when the

scoring is consistent across different raters. Assumptions associated with the warrant include, for instance, assigning the same ratings at the test level and having a sufficient number of raters to generate a reliable score. To back the assumptions for each inference, validation studies have widely applied MFRM, which creates diagnostic information about the quality of a test and rater performance based on fit statistics (see details in Methods section).

## THE CURRENT STUDY

### *Context*

The current study draws on four years of operational testing data to investigate a major rubric revision for a writing placement test used in HELP at UHM. As part of the Department of Second Language Studies (SLS) at UHM, HELP is an intensive academic language program that prepares non-native speakers of English for academic work at UHM or other English-language institutions of higher education. The program's curriculum is divided into four levels (100 through 400). Students may enroll in an 8-week term during the spring or fall, or a 4- or 6-week term in the summer. Students who successfully complete two terms at the 400-level are eligible for conditional admission to UHM, raising the stakes of the outcomes associated with the institute's placement exam (i.e., higher placement can result in quicker university matriculation).

In many language programs, the placement tests are either bought from commercial publishing houses, adapted from other English as a second language (ESL) programs, or reproduced from current textbooks (Brown, 1996). Students enrolled in HELP are placed in their course levels on the intake day before classes begin, and their placement is based on a battery of tests including the Michigan English Placement Test (EPT), a writing sample, and an oral interview. While the University of Michigan EPT is a multiple-choice listening test, the writing sample and speaking test rely on raters for scoring. The writing and speaking tests are scored by trained teacher-raters and program administrators at HELP.

The old rubric (in use between 2010–2016) under investigation was developed locally at HELP by a group of writing instructors with guidance from the administration. The writing instrument asked incoming students to produce a hand-written, multi-paragraph essay in response to one of three prompts within a 30-minute timeframe:

1. Who is the most important person in your life? Why?
2. Write about an interesting tradition or holiday in your country.
3. Write about a time in your life when you had a problem. What was the problem? How did you find a solution to the problem?

The prompts aimed to elicit descriptive writing, and there was not any guidance in terms of paragraph organization. Students were encouraged to make a list of ideas before writing. They were also explicitly instructed to raise their hand if they have a question or need more paper, write legibly, and not to use a dictionary or talk to peers during the test. In terms of scoring, two instructors were asked to score eight categories (see Figure 1) according to numerical values that coincided with the four course levels at HELP (i.e., 100, 200, 300, or 400). Additionally, raters were asked to compute an average across all categories and make an overall level recommendation for the student.

In 2016, HELP decided to revise its rubric due to concerns raised by the teacher-raters. In this instance, the prompts did not elicit writing that was reflective of university writing tasks, such as formulating or evaluating an argument. Also, some teacher-raters were dissatisfied with this system as there had been confusion as to whether their scores should be considered analytic or holistic (Rock, 2016). If raters were working with analytic scores in mind, each of the eight categories should have been scored independently based on students' abilities; however, if raters chose to view the rubric holistically, they would either select uniform scores or scores that work out to a numerical average in line with their intended final placement score. Among raters who attempted to score each rubric category independently, some raters reported difficulty in distinguishing among eight different aspects of writing quality, some of which were seen as overlapping (e.g., "Organization" and "Flow/Cohesion"). Related to this potential for inconsistency in scoring among raters, it was found that some raters preferred to create their own 'desire paths' in placing the students by recording half scores (i.e., 150, 250, and 350) when they were unwilling to commit to providing a whole score.

Based on input from the teachers and administrators, in 2016, HELP revised the 8-category old rubric with a 4-point scale to a 4-category new rubric with a 20-point scale. The writing prompts were changed from eliciting descriptive writing to argumentative writing. Although aspects of the new rubric have been examined using MFRM in a pilot study (Rock, 2016), to this day, no prior research has investigated the quality of the old rubric and its rater performance.

There is therefore no point of reference for comparing the quality of the old rubric versus the new rubric and their impact on rater performance from an empirical standpoint.

**Figure 1**

*Old Writing Placement Test - Scoring Sheet & Rubric*

**Writing Placement Scoring Guidelines**

| | 100 | 200 | 300 | 400 |
|---|---|---|---|---|
| **Fluency** | Very little writing<br>Only several sentences in 30 min | | | Very fluent language use<br>Writes 1 ½ -2 pages in 30 min |
| **Organization**<br>*(not all essays will have paragraphs, even at high levels)* | No intro/conclusion<br>No thesis statement<br>May not have paragraphs | Some paragraphs have a clear main idea | Many paragraphs w/clear main idea<br>TOEFL-like, 5 paragraph organization | Most paragraphs have clear main idea |
| **Flow / Cohesion** | Few connectors (or none at all)<br>Ideas are choppy (not related) | Some simple connectors (first, next, finally) | Frequent use of connectors, more advanced, start sentences w/ adj clause | Sophisticated use of connectors |
| **Sentence Complexity** | Short, simple or incomplete sentences (4-5 words) | A few different sentence and clause types<br>Primarily writes using "I" as subject | Some use of complex sentence patterns | Variety of complex sentence patterns and embedded clauses |
| **Grammar/ Mechanics** | Basic grammar (SVO)<br>May not capitalize/punctuate correctly | Errors often interfere w/comprehension | | Errors seldom interfere w/ comprehension |
| **Vocabulary** | Limited/simple vocabulary<br>Frequently repeats same words | Limited vocabulary | Uses a variety of general vocabulary items with some low frequency words | Uses very specific, academic, or technical vocabulary |
| **Focused Topic/ Support** | Topic is not focused or slightly focused<br>Few (if any) supporting details and Examples | Somewhat focused w/ few details | Focused w/many details to support paragraph main ides | Very focused w/many details |
| **Depth of Content** | Lacks ability to express basic idea of exam question | Expresses ideas in very little detail; likely to address the question on a surface level | Expresses some advanced ideas surrounding the question | Expresses many advanced ideas surrounding the question clearly |

Highlighted = Most distinguishing features of each writing level. Used to make borderline decisions (see back side of handout)

**Writing Placement Test – Scoring Sheet**

Student name: _____

Rater name/initials: _____

| Category | 100 | 200 | 300 | 400 | Level |
|---|---|---|---|---|---|
| Fluency | | | | | |
| Organization | | | | | |
| Flow/ Cohesion | | | | | |
| Sentence Complexity | | | | | |
| Grammar/ Mechanics | | | | | |
| Vocabulary | | | | | |
| Focused Topic/ Support | | | | | |
| Depth of Content | | | | | |

**Total_____/8**

**Average _____**

**Recommended Level _____**

The current study aims to address this gap in research by investigating the quality of the old rubric and its rater performance. Specifically, the study is guided by the following research questions:

RQ1. In the context of the HELP writing placement test, how well was the rubric functioning?

RQ2. How did the raters perform using this rubric?

## METHODS

### *Data*

The research team accessed archived placement test data from 325 unique students who had been admitted to HELP between the Spring semester in 2012 and Fall semester in 2016.[1] Students who typically attend HELP tend to be college-aged (19–22 years old) and the vast majority tend to come from East Asian countries (e.g., Japan, Korea, and China), with a small minority of students from countries in the Middle East, Europe, and Southeast Asia. For the latter groups of students, the ages tended to skew higher into their 30s and 40s, and there were a few examples of students in their late 50s and early 60s.

A total of 53 teacher-raters were included in the Spring 2012 to Fall 2016 data. The raters at HELP are composed of a mix of longtime administrators, part/full time instructors, and newer instructors, some of whom also serve as graduate teaching assistants. These raters come from a variety of backgrounds, ages, and are equally mixed in gender. Some had joined the program with a long history of L2 English teaching experience, while others were relatively more novice. As each term at HELP was typically held for an 8-week period, there was an increase in instructor turnover throughout the year, particularly with the graduate teaching assistants. This also meant that raters varied in their experience scoring the writing placement test, and each administration featured a different pool of raters. Due to limited resources, especially given that placement decisions have to be made almost immediately after the students have taken the test, training opportunities for new raters are limited.

[1] Although the old rubric was implemented between 2010–2016, the research team was only able to obtain archived data between 2012–2016.

***Analysis***

An MFRM analysis was used to examine how well the old rubric of the placement writing task functioned. The MFRM measurement investigates multiple measurement facets (i.e., students, raters, categories, and prompts) and enables direct comparisons among them in log-odds units (*logits*). Diagnostic information on (a) the quality of measurement elements such as prompt and students fitting the model and (b) the interactions across facets (e.g., how raters rate different prompts) are also provided. MFRM creates two measures of fit statistics: Infit Mean Square (IMS) and Outfit Mean Square (OMS). Since IMS is more sensitive to non-extreme unexpected ratings because it is weighted by the variance of ratings, we focused on IMS and set the acceptable infit value range as between 0.60 and 1.30 (Myford & Wolfe, 2003). IMS values that are higher than the acceptable range indicate underfitting of the raters; in contrast, values lower than the acceptable range signal overfit[2]. MFRM also allows for detailed analysis of rating scales. Logit-scaled thresholds between score points and probability curves for each score point can be estimated to evaluate the functioning of scales. Please refer to Eckes (2011) for more information on MFRM.

To analyze the data, the FACETS software (version 3.83.6, Linacre, 2021b) was used. A Four-Facet Rating Scale model (RSM) consisting of measurement facets (i.e., students, raters, categories, prompts), were constructed for the old rubric with 7-point scales (original 4 points with consideration of half points used by the raters). Unidimensionality of the datasets was checked using infit values, variance explained by Rasch measurement, eigenvalues, and standardized model residuals (see Linacre, 2021a). As the test developers intended subscales for rubric criteria to be applied similarly, the RSM was chosen for performing the MFRM rather than partial-credit hybrid models. There was an issue with connectivity that was discovered in the prompt facet: Prompt three ("Write about a time in your life when you had a problem. What was the problem? How did you find a solution to the problem?") was less selected by the students than the other two prompts, resulting in a small subset of unconnected data. To resolve the problem, this prompt was specified as a 'dummy' facet (i.e., not estimated) in the RSM, as

[2] A misfit is indicated in MFRM if violations of the theoretical expectations of the model exceed a certain degree of deviations. Underfit means the data is unpredictable and not fitting the model's expectations (e.g., unexpected ratings from the rater). Overfit is the case in which the deviation from the model's assumption is less than expected. For example, when a rater's scores are too predictable (e.g., when generally using a narrow portion of the rating scale), they are considered to be overfitting.

prompts appeared to differ very little in terms of difficulty based on observed score averages and partial MFRM calibrations of the data.

## RESULTS

In the following, we present the results of the MFRM analysis of the rubric.

### *Many-facet Rasch Model fit*

The infit values of the measurement model ranged from 0.89 to 1.10 and 0.95 to 1.03 for the category facet and the prompt facet, respectively, indicating unidimensionality of the data based on Smith (1996). Additional evidence for unidimensionality was that 63.71% of variance was explained by the Rasch measure. According to Reckase (1979), above 20% of variance explained by Rasch measure can suggest unidimensionality. Furthermore, a principal components analysis of Rasch model residuals (PCAR) was used to evaluate unidimensionality, and an eigenvalue greater than 2.0 for a contrast was considered as potential evidence of non-unidimensionality (Linacre, 2016). Our results returned a first contrast eigenvalue of 0.30. Tallying standardized model residuals showed that 2% of residuals were over |2.0| and none were over |3.0|, showing good data-model fit (Linacre, 2002).

### *Measurement Summary*

A Wright map is presented to display the summary of the model with all measurement facets (Figure 2). The leftmost column (Measr) of the figure is a ruler indicating each element's associated logits in every facet. The first facet, student, had Rasch-estimated writing ability between -7 and +4 logits, where higher logits suggest greater writing ability. The estimate of Person separation reliability of the student facet was .95, showing the writing test reliably measured students' ability. With a separation reliability of .96, the rater separation index was 6.43, suggesting that raters could be reliably separated into at least 6 different strata in terms of severity. In addition, the maximum estimate of severity was 2.82 logits from the most severe rater, while the minimum estimate of severity, -2.65 logits, came from the most lenient rater; this range of 5 logits could yield a difference of up to 4 score points (of the 7-point scale; namely, 100, 150, 200, 250, 300, 350, 400) on a given category. In other words, for rater, the higher the

value of the logits, the stricter the rater is. The third column of Figure 2 corresponds to the 8 scoring categories, where higher logits indicate greater difficulty to earn a higher score on the category. "Vocabulary" was the most difficult category to receive higher scores (0.31 logits), whereas "Fluency" was the easiest (-0.31 logits). However, the difficulty range of rubric categories was not large, and several categories had nearly identical difficulties. The last column provides information on the scale of scoring points. Each horizontal line in the last column indicates the threshold in logits, that divides adjacent score points on the rating scale. For example, it was more likely for a student with an ability of 1 logit to receive 300 points in each category.

### *Rating Scale Analysis*

Table 1 shows that the (unsanctioned) half points 150, 250, and 350 were used far less often than adjacent rubric-allowed score points. These three score points were half points assigned by the raters, which were not provided in the rubric. Although the average measure and the threshold of score points on the scoring scale seems to be in order (Table 1), each score point did not reflect a distinct range of student ability. Score points 150, 250, and 350 were wholly subsumed by an adjacent point(s) (see category probability curves in Figure 3).

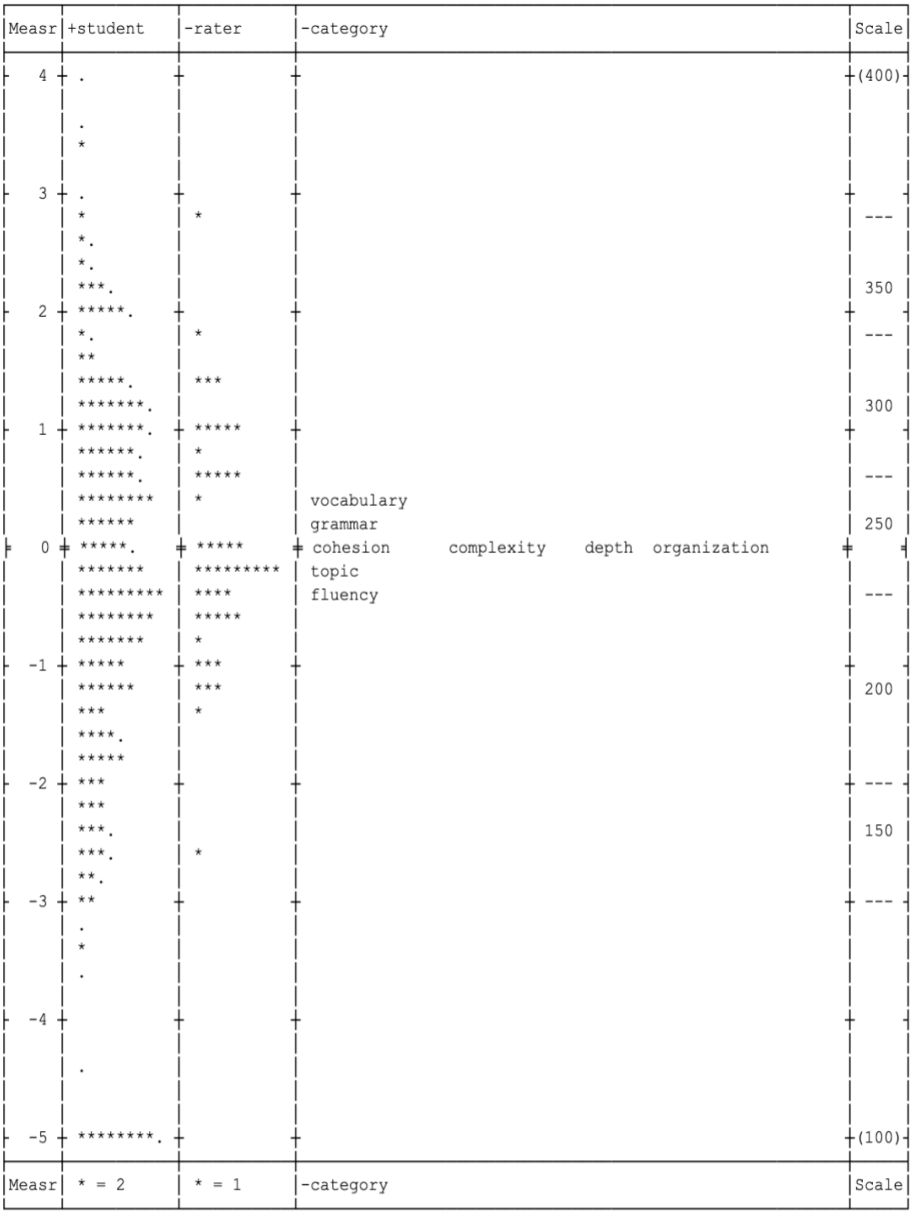
**Table 1**

*Score Point Summary – Old Rubric*

| Score | % of all scores | Average measure | Outfit Mean square | Rasch-Andrich Thresholds |
|---|---|---|---|---|
| 100 | 9% | -2.53 | 1.0 | |
| 150 | 2% | -1.88 | 0.8 | -4.13 |
| 200 | 42% | -0.76 | 1.2 | -3.97 |
| 250 | 5% | 0.05 | 0.8 | -1.41 |
| 300 | 32% | 0.85 | 1.0 | -0.67 |
| 350 | 1% | 1.88 | 0.4 | 0.2 |
| 400 | 9% | 2.35 | 1.0 | 1.92 |

**Figure 2**

*Variable Map of the Writing Placement Test using the Old Rubric*




### *Rater Fit*

Figure 4 illustrates the consistency of individual raters based on their infit and outfit statistics. Ten out of 49 raters (20%) had IMS values outside of the acceptable range, including three raters with values > 2.0, indicating an extreme level of misfit (Linacre, 2016). A somewhat larger proportion of raters had OMS values outside of the acceptable range, including five that were near or exceeding 2.0.

**Figure 3**

*Score Probability Curves for the Old Rubric*

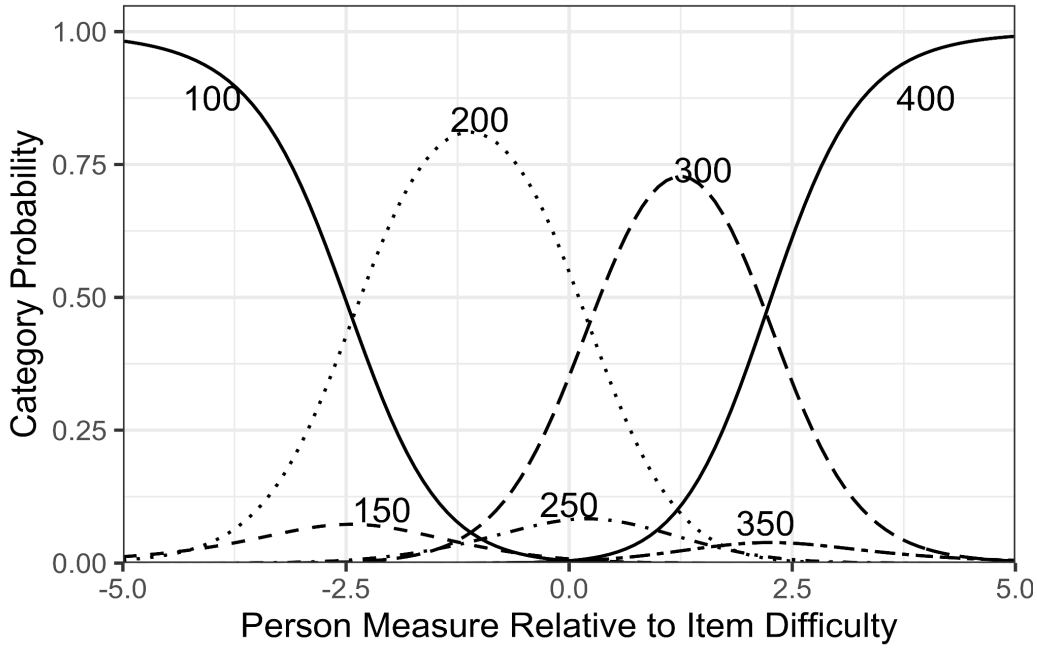
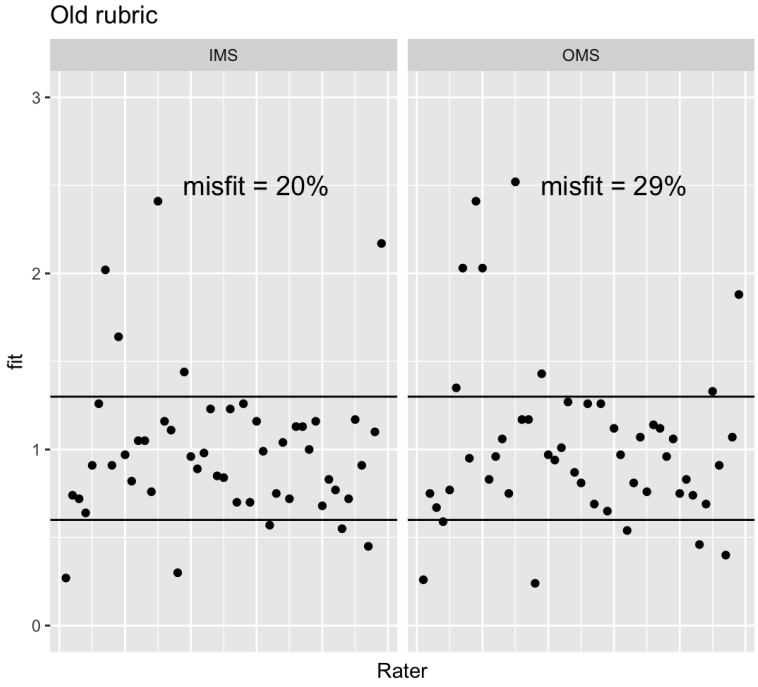



**Figure 4**

*Rater Fit of the Old Rubric*




*Note.* IMS = infit mean square, OMS = outfit mean square.

## DISCUSSION AND CONCLUSION

When taking into consideration that the purpose of the HELP writing exam is meant to divide students into the four distinct levels associated with the core curriculum, the old rubric was actually functioning well from a broad perspective. However, when looking more closely at the results of the MFRM analysis, there were potential issues regarding the eight categories and sporadic use of half points that needed to be addressed.

In support of the anecdotal evidence reported in Rock (2016), it seems as though the teacher-raters had a difficult time making a meaningful distinction between the categories of "Flow/Cohesion," "Sentence Complexity," "Depth of Content," and "Organization." Looking more closely at "Sentence Complexity" and "Flow/Cohesion," for instance, the rubric descriptors for the 300-level stating "some use of complex sentence patterns" and "more advanced, starts sentences with adjective clauses" would likely be intrinsically linked, particularly when "some" and "more advanced" are open for interpretation from the teacher-raters. This seems to be the case, in reviewing the Rasch difficulty estimates for this pair, they were treated at almost exactly the same level. If the raters were unable to treat these categories as being distinct, their continued inclusion as part of the rubric would not have resulted in meaningful placement decisions. Additionally, it may be worth noting that the "Fluency" category may have been seen as the "easiest" category to score as it was entirely related to sentences/page length, and interestingly was rather unspecified on the rubric, lacking descriptors for both the 200 and 300 levels; considering this, it is unclear how teacher raters would treat these categories for students who fell into the middle of the range at the 200 or 300 level.

The other potential issue was the use of half points by a few teacher-raters in the data set. It was not such a wide-spread phenomenon that it impacted the ordering of the scoring scale, but rather, the appearance of half points is a reflection of the indecision of particular teacher-raters regarding each category's description. Because the 100 to 400 level on the *Writing Scoring Guidelines* (see Figure 1) were visualized as a range rather than distinct levels, there is the possibility that some teacher raters felt more confident to mark a score that fit within that range. However, since students are ultimately placed within one of the four curriculum levels, these half points were not useful in accomplishing this goal and ultimately undermined consistent criteria and procedures for assigning scores to writing performances.

In regard to how the teacher-raters were performing using this rubric outside of the half point issue, the majority of teacher raters were demonstrating good fit; however, there was a large group of raters that were misfitting. As previously explained, HELP typically has an increased likelihood of turnover rate regarding their instructors each semester and regularly-scheduled placement training sessions were limited when compared to large-scale tests. Considering this, there is evidence to suggest that some teacher-raters may have had difficulties in making meaningful distinctions among the eight categories of the rubric, as well as had difficulties in connecting the descriptors of the categories to specific levels of performance/ability aligning with placements without knowledge of the curriculum. This may have resulted in either the use of half points or the decision to mark the same score across multiple categories (e.g. marking "300" for both "Sentence Complexity" and "Flow/Cohesion").

Therefore, while the initial motivation for revising HELP's rubric in 2016 was based on anecdotal evidence, our empirical analysis also suggests that a revision was necessary and justified. While there is evidence that the rubric used from 2012–2016 was functioning well and the majority of teacher raters were using it as intended, there were notable issues related to likely confusion surrounding the distinctiveness of the eight categories, the potential for the use of half points, and the higher number of misfitting raters that undermine the consistency and meaningfulness of individual scores.

## REFERENCES

Bachman, L., & Palmer, A. (2010). *Language assessment in practice.* Oxford University Press.

Brown, J. D. (1996). *Testing in language programs.* Prentice Hall Regents.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Peter Lang D.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

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

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2016). Winsteps® (Version 3.92.1) [Computer Software]. Beaverton, OR: Winsteps.com. Retrieved from http://www.winsteps.com/

Linacre, J. M. (2021a) Fit diagnosis: infit outfit mean-square standardized, Retrieved on December 6, 2021 from https://www.winsteps.com/winman/misfitdiagnosis.htm

Linacre, J. M. (2021b). Winsteps® (Version 5.1.1) [Computer Software]. Portland, Oregon: Winsteps.com. Available from https://www.winsteps.com/

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics*, *4*(3), 207–230. https://doi.org/10.3102%2F10769986004003207

Rock, K. (2016). *Development and analysis of a writing rubric for an IEP.* Department of Second Language Studies, University of Hawai‘i (Unpublished manuscript).

Smith, R. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling*, *3*(1), 25–40. https://doi.org/10.1080/10705519609540027

Weigle, S. C. (2002). *Assessing writing.* Cambridge University Press.

# FACE-TO-FACE VERSUS REMOTE L2 SPEECH ELICITATION: LISTENERS' PERCEPTIONS OF SOUND CLARITY

**DUSTIN CROWTHER & KRISTEN URADA**
***University of Hawai'i at Mānoa***

## ABSTRACT

Research into listeners' global speech perceptions (e.g., accentedness, comprehensibility) has primarily emphasized face-to-face (F2F) speech elicitation that allows researchers to maintain direct control over recording procedures. Given advancements in digital technology (and recent restrictions placed on researchers during the COVID-19 pandemic), it is imperative to investigate the extent to which speech elicited remotely (i.e., through the use of computer and/or mobile applications) is comparable to speech elicited F2F. Fourteen English learners completed a pair of tasks in the presence of a researcher (i.e., F2F), while 14 other learners did so remotely through the application Extempore. Thirty native listeners rated each recording for accentedness, comprehensibility, and sound clarity. Listeners indicated relatively high reliability across all three dimensions, and no significant differences in sound clarity were found between F2F and remote speech. Despite initial positive implications of these findings, several additional concerns remain. Guidelines for eliciting second language speech remotely are provided.

Broadly defined, second language (L2) Intelligibility refers to listeners' ability to understand an L2 utterance (Levis, 2020). To avoid possible confusion, we refer to this broad conceptualization using an upper case "I". More narrowly, Levis (2020) discussed Intelligibility in reference to both accuracy of listeners' understanding of an L2 utterance (*intelligibility*, represented here with a lower case "i") and listeners' degree of effort required to understand the same utterance (*comprehensibility*). These two narrow definitions of Intelligibility were established in Munro and Derwing (1995), alongside a third dimension of L2 speech, *accentedness*, or the extent to which an L2 speaker can approximate the sounds of a target

community. As scholarly interest into these three dimensions continues to see significant growth (see Crowther et al., 2022; Derwing & Munro, 2011 for a pair of research timelines), it is not only necessary to ensure rigorous methodological practice, but to simultaneously explore areas for methodological growth. Two of the aforementioned dimensions, comprehensibility and accentedness, target listeners' impressionistic judgments of L2 speech, and are traditionally operationalized using listeners' scalar ratings (Munro & Derwing, 2015).[3] While the methodological appropriateness of scalar ratings in reference to both dimensions has received scholarly consideration (e.g., Isaacs & Thomson, 2013; Isbell, 2018; Southwood & Flege, 1999), little emphasis has been given in regards to the technical procedures behind the recordings to be rated. While the vast majority of L2 comprehensibility and accentedness studies report face-to-face (F2F) speech elicitation, advances in digital technology have provided numerous tools allowing for remote, application (app)-based speech elicitation (e.g., Extempore, Flipgrid, JotForm). However, given that such an app-based approach would lessen a researcher's ability to control environmental factors during recording, it is necessary to establish a) whether remote-elicited speech can be comparable to F2F-elicited speech, and, if possible, b) a set of guidelines to ensure comparability between remote- and F2F-elicted speech. To address these two points, we present a brief report which compares 30 native English listeners' perceptions of two groups of intensive English program (IEP) students, one group which recorded their speech in a F2F environment (i.e., with the researcher present), and one group which recorded their speech remotely (i.e., using the app Extempore). Listeners assigned Likert scale ratings for not only comprehensibility and accentedness, but also sound clarity. We draw upon the 30 listeners' ratings of sound clarity to investigate *to what extent L2 speech elicited in a F2F environment is comparable to L2 speech elicited through remote procedures in reference to recording quality.*

## GLOBAL SPEECH RATING PROCEDURES

In eliciting L2 speech, existing research indicates a range of both theoretical and methodological considerations. By theoretical, we refer to decisions tied directly to the objectives of the researcher(s). For example, interest in how learners' first language (L1) may

[3] A third dimension not considered in the current study, *fluency*, has frequently been included as well, typically defined as the ease of flow of L2 speech, in reference to the presence of pauses and other dysfluency markers (Derwing & Munro, 2015).

influence how listeners perceive their speech has led researchers to elicit speech from different L1 groups (e.g., Crowther et al., 2015; Derwing & Munro, 1997). Similarly, researchers have considered the role of L2 proficiency in listener perception by eliciting speech across a range of proficiencies (e.g., Isbell et al., 2019; Saito et al., 2016). Or, recognition that task specifications can affect the language necessary for completion has led studies to elicit participants' speech on a range of tasks (e.g., Crowther, 2020; Crowther et al., 2018). These are only a few examples of theoretically motivated decisions when eliciting L2 speech. Of particular interest to the current study, however, are those decisions which are methodological in nature.

L2 Intelligibility research has not been lacking in regards to methodological review. For example, Kang et al. (2018) investigated several measures of intelligibility to determine which may best represent the target construct (i.e., accuracy of understanding). For the impressionistic measures of comprehensibility and accentedness, Isaacs and Thomson (2013) considered the appropriateness of 5-point versus 9-point Likert scales for eliciting listeners' perceptions of L2 speech (see also Isbell, 2018 and Southwood & Flege, 1999), while Nagle and Rehman (2021) discussed the appropriateness of eliciting such perceptions online as opposed to in person. Though such Likert scale ratings appear to be the norm (Munro & Derwing, 2015), recent studies have employed a few additional measures. For example, Saito et al. (2017) utilized a 1000-point sliding scale for similar purposes (see also Crowther et al., 2015, 2018, amongst several others), while Nagle et al. (2019) made use of MacIntyre's (2012) Idiodynamic Software to enable real time adjustments of comprehensibility ratings (i.e., listeners could upgrade/downgrade a speaker's comprehensibility throughout an utterance). With evidence that indicates that 20- to 60-second recordings allow for reliable ratings of L2 speech (Munro et al., 2010), existing research has frequently settled on 30-second utterances (e.g., Crowther et al., 2015, 2018). Yet, with mounting evidence that global perceptions, specifically in regards to comprehensibility, are dynamic (e.g., Nagle et al., 2019; Trofimovich et al., 2020), research into stimuli length is growing (e.g., Moran Wilson et al., 2019; Suzuki & Kormos, 2020). One final example is that of O'Brien (2016), who investigated a) whether a difference existed when rating perceptual measures simultaneously versus separately, and, if separately, b) if the order of measures (i.e., comprehensibility, accentedness, fluency) mattered.

Despite clear methodological interest, surprisingly little emphasis has been placed on actual recording procedures. The vast majority of L2 global speech studies indicate an emphasis

on F2F speech elicitation, which is not surprising given that F2F procedures allow for the necessary experimental control identified in Munro and Derwing (2015). Making reference to Murphy's law (anything that can go wrong will go wrong), Munro and Derwing (2015) highlighted the importance of "equipment quality, quiet recording environments, and post-collection processing of audio files" (p. 24), amongst other important considerations. While recent research has indeed pushed beyond traditional laboratory and classroom settings for speech elicitation, even video-based teleconferencing studies (e.g., Akiyama & Saito, 2016; Saito & Akiyama, 2017) still enabled some degree of researcher control. However, recent years have seen an increased pool of apps that allow for remote recording of speech, many which allow for immediate and automatic upload to a pre-designated destination. Saito et al. (2018) utilized ALC Press Inc.'s Telephone Standard Speaking Test (https://tsst.alc.co.jp/biz/en/), in which participants responded to several speaking prompts over the phone. Saito et al's analyses included eliciting ratings of comprehensibility. Inceoglu (2019), during a classroom-based intervention study, required participants to record their speech using an audio Dropbox. These recordings were not editable, allowing her to analyze participants' initial attempts at the speaking task. These analyses included eliciting ratings of comprehensibility, accentedness, and fluency.

A number of additional app-based tools are now available for use in eliciting L2 speech. Given expected technological advancements, we here simply reference several tools considered for our study (conducted in spring 2020). *FlipGrid* (info.flipgrid.com) was a free mobile or PC based program designed for educators. For the purposes of speech elicitation research, FlipGrid could allow participants to submit video-based responses to a range of prompts via their phone or computer, which were then immediately available for analysis by the researcher. The drawback (at the time of our study) was that FlipGrid did not provide an option for video-free elicitation. *JotForm* (jotform.com) was a freemium survey-based program which could allow for a range of data collection procedures. However, for speech elicitation purposes, participants would have to first record their responses before uploading them to JotForm, adding an additional complication to the process. Finally, *Extempore* (extemporeapp.com) was a freemium mobile and PC-based program. The advantage to using Extempore was that participants could record both audio + video and audio-only responses (and even textual, if desired by the researcher). Despite this already extensive range of tools available to L2 researchers for app-based speech elicitation, little research exists considering issues of experimental control, such as equipment quality, recording

environments, and audio files processing. Though convenient, the use of new, technologically-enhanced procedures must be considered with the same care previously raised by Munro and Derwing (2015) for F2F settings.

## THE CURRENT STUDY

We present a comparison of two sets of speech data, collected under different circumstances. The first set of data features the speech of 14 L1 Japanese speakers who completed a series of English-speaking tasks in a F2F environment, and serves as a subset of data originally reported in Crowther (2020). The second set of data again consists of 14 L1 Japanese speakers, but who completed their pair of English-speaking tasks using the app Extempore. We note that the initial impetus to use app-based speech elicitation was due to the sudden restrictions placed on F2F data collection due to the growing threat of COVID-19 in spring 2020.[4] Recognizing that the participants in both data sets were enrolled in IEP courses at the time of data collection, represented the same L1, and had completed comparable speaking tasks, we decided to pool our data as a means to investigate the comparability of F2F- versus remote-elicited speech in reference to global dimensions (i.e., comprehensibility, accentedness) of L2 speech. The research question that guided our study was as follows:

1. To what extent is L2 speech elicited remotely comparable to L2 speech elicited in a F2F environment?

To address this question, we elicited the Likert scale ratings of 30 native-English listeners for comprehensibility, accentedness, and sound clarity. To determine comparability, we first report on the reliability of listeners' perception of sound clarity, followed by a comparison of sound clarity ratings between the F2F and Extempore groups. Finally, though neither the F2F or Extempore speech data were initially intended for comparison, we briefly report on group comparisons of comprehensibility and accentedness simply to reinforce the potential of remote-

[4] The initial study was an IEP-based pedagogical intervention drawing upon principles of Rose & Galloway's (2019) Global Englishes Language Teaching. While an attempt was made to continue coursework online, the class participants soon returned to their home country and discontinued their studies.

elicited speech data. We conclude our paper by presenting proposed guidelines to consider when eliciting participant speech remotely.

## METHODOLOGY

### *Participants*

We recruited 30 undergraduate students (*age* = 22.5, *SD* = 4.75, *range* = 18–34) enrolled in a US Pacific university to serve as Listeners. Due to COVID-19 concerns, only 20 Listeners actually resided in the US, while the remaining 10 lived abroad: Korea (5), China (1), Indonesia (1), Japan (1), Philippines (1), and Switzerland (1). Seventeen Listeners reported English as their L1, while 11 more reported Korean (5), Ilocano (2), Japanese (2), Chendu Dialect (1), and Indonesian (1). The final two Listeners reported as being bilingual. Twenty-five Listeners (83%) reported familiarity with at least one additional language.

### *Materials*

We elicited speech from two sets of IEP students, both consisting of 14 L1 Japanese speakers of English. Speakers in set one provided their speech in a F2F environment (and are hereafter referred to as F2F Speakers), while speakers in set two provided their speech using the app Extempore (hereafter Extempore Speakers).

**F2F speakers**. The 14 F2F Speakers (7 females, 7 males) were attending an IEP at a Midwest US university. Their mean age was 24.07 (*SD* = 5.95), and they had been studying English for 12.54 years (*SD* = 5.49). An in-house placement test placed them in levels 3 (*N = 3*) and 4 (*N* = 11) of a 4–level program. The Speakers each completed a picture narrative task and long turn task. The picture narrative task was Derwing et al's (2009) "Suitcase Story", which shows a man and a woman who bump into each other, accidently exchange their identical suitcases, and realize their mistake upon returning to their respective residences (available through the IRIS Digital Repository). For the long turn task, Speakers, in 1–2 minutes, described either a restaurant or party that they previously enjoyed (with prompts derived from International English Language Testing System, 2009, 2011).

**Extempore speakers.** The 14 Extempore Speakers (4 female, 10 male) were enrolled in an IEP at a US Pacific university and had a mean age of 19.43 (*SD* = 0.73). On average, these

Speakers had been studying English for 7.93 years (*SD* = 2.19) and an in-house placement test indicated intermediate level proficiency. Each Speaker completed a picture narrative task and long turn task. The six-frame picture narrative task portrayed a girl's birthday party, including a trip to the zoo where the absence of the bears resulted in the birthday girl crying (available via the IRIS Digital Repository, under Mackey, 1999, 2002). For the long turn task, Speakers responded to the same party prompt used for F2F Speakers.

### *Procedures*

**F2F speakers.** Each Speaker met individually with the first author in a private on-campus office. They each completed a total of three tasks in a counterbalanced order, though only two, the picture narrative task and long turn task, are considered in this study. The first author provided instructions on how to complete each task and addressed any questions before Speakers recorded their responses. Speakers' responses were recorded on a Sony ICD-PX333 digital recorder.

**Extempore speakers.** Each Extempore Speaker completed their recordings during the first week of their IEP study. Due to the unexpected shift to online learning caused by the onset of the COVID-19 pandemic, the second author asked Speakers to record their responses remotely using Extempore. Speakers received an image-based PDF document providing in-depth instructions on how to access and navigate the Extempore interface, and were asked to complete their recordings in a quiet location. Upon accessing Extempore, Speakers could choose which of the three tasks to begin with (a read aloud task is not included in the current study), and upon choosing a task received their speaking prompt (picture narrative or long turn). When ready, Speakers would click "Record", and when finished, Speakers would click "Submit Attempt". All recordings were saved automatically, and were immediately available to us for analyses. In several instances, we were required to unlock students' accounts due to technical issues with Extempore (see discussion).

**Speech rating.** We edited all recordings (F2F and Extempore) down to the initial 30 seconds of speech, with initial dysfluencies (e.g., uh, um) and false starts removed (e.g., Saito et al., 2016; Trofimovich & Isaacs, 2012). The 56 recordings (28 speakers X 2 tasks) were then presented to Listeners using the online survey software Qualtrics (www.qualtrics.com). Listeners rated all recordings using 9-point Likert scales in a two-block session, with blocks

counterbalanced across Listeners based on tasks (i.e., picture narrative-long turn, long turn-picture narrative). F2F and Extempore recordings were alternatively distributed within each block, with four different randomizations per block. Prior to rating, Listeners received instructions (through Qualtrics) on all rating constructs and completed two practice ratings per block. Twenty-nine listeners reported they completed the speech rating on a laptop and used headphones. All 30 listeners indicated they completed their ratings in a quiet location. Listeners assigned ratings for three dimensions:

- Accentedness – similarity of speech to a speaker of North American English (1 = highly accented, 9 = not accented at all)
- Comprehensibility – ease or difficulty in understanding the speaker (1 = hard to understand, 9 = easy to understand)
- Sound Clarity – extent to which recording clarity impacted ability to assess speakers' accentedness and comprehensibility (1 = very poor sound clarity, 9 = excellent sound clarity)

We piloted the entire rating process with a pair of listeners, and made revisions as needed. Specifically, the instructions for Sound Clarity were refined, including revising the dimension title from Sound Quality to Sound Clarity. For all recordings, Listeners were able to rate accentedness and comprehensibility simultaneously while listening; however, sound clarity ratings were elicited on a following page after the recording was finished.

## RESULTS

To address the extent to which L2 speech elicited remotely (i.e., through Extempore) is comparable to L2 speech elicited in a F2F environment we conducted two analyses. The first analysis measured Listener reliability per rated construct, though with an emphasis on sound clarity. The second analysis compared Listeners' ratings of sound clarity between the F2F and Extempore groups.

### *Reliability*

We begin with reliability, as high interrater reliability indicates that "listeners share a similar perceptual experience of the phenomenon at hand" (Munro & Derwing, 2015, p. 30). Following

common practice (Munro & Derwing, 2015), we began by calculating Cronbach's alpha (α) for accentedness, comprehensibility, and sound clarity. We report both overall and elicitation medium-specific alpha values in Table 1. We note that when calculating alpha coefficients for sound clarity, one listener was flagged and removed due to a lack of variance in their ratings (i.e., they assigned a "9" to every recording). Reliability ratings all exceeded the oft cited threshold of .70–.80 (Larson-Hall, 2016), and were in line with much previous research considering accentedness and comprehensibility ratings (Isaacs & Thomson, 2013). Interestingly, speech ratings elicited in the app-based context appeared slightly more reliable than those in the F2F context, though values for both contexts exceed the .70–.80 threshold.

**Table 1**

*Cronbach's Alpha (α) Values for 30 Listeners' Ratings of Speech*

| Rating Category | Overall | F2F | Extempore |
|---|---|---|---|
| Accentedness | 0.93 | 0.81 | 0.95 |
| Comprehensibility | 0.92 | 0.88 | 0.93 |
| Sound Clarity[a] | 0.88 | 0.80 | 0.91 |

*Note*. [a] One Listener was removed due to lack of variance in ratings.

As relying on a single measure of reliability can be misleading (Isaacs & Thomson, 2015; Stemler & Tsia, 2008), we additionally considered reliability using the intraclass correlation coefficient (ICC) 2*k* variant. To provide a bit more clarity, unlike Cronbach's alpha, which determines reliability based on similarities between listeners in the rank ordering of speakers (i.e., without consideration of whether listeners are actually using the range of scores similarly), ICC additionally models rater variation in determining reliability (Isbell, 2018), with the 2*k* variant appropriate when "averaged scores from a random group of raters are to be used for interpretation" (Isbell, 2018, p. 96).[5] As before, we report both overall and elicitation-based values in Table 2. Though not as high as the Cronbach's alpha values described above, generally reliability ranges from good (.60–.74) through excellent (.75–1.00) (Cicchetti, 1994).

[5] The ICC 2*k* variant is used when *k* number of raters are randomly selected from a larger population, their ratings will be averaged to determine a single score per item (in our case, Speaker), and absolute agreement between raters is of interest.

Interestingly, the one exception is that Listener reliability for the sound clarity of F2F recordings was only fair (.40–.59).

**Table 2**

*ICC (2k) Values for 30 Listeners' Ratings of Speech*

| Rating Category | Overall | F2F | Extempore |
|---|---|---|---|
| Accentedness | 0.89 | 0.68 | 0.93 |
| Comprehensibility | 0.83 | 0.75 | 0.86 |
| Sound Clarity[a] | 0.73 | 0.53 | 0.81 |

*Note*. [a] One Listener was removed due to lack of variance in ratings.

### *Group Comparisons*

We report mean and standard deviations for sound clarity ratings in Table 3. For both tasks, there appeared to be minimal difference in Listeners' ratings of sound clarity for F2F- and Extempore-elicited speech. Given both the low sample size and a generally negative skew in ratings, we ran a pair of non-parametric Mann-Whitney U tests. Our results indicated Listeners did not perceive a significant difference in sound clarity between the F2F and Extempore speech samples for either the picture narrative (W = 130.50, $p = .141$, $r = .28$) or long turn (W = 123.50, $p = .260$, $r = .21$) tasks[6], though a weak effect for the picture narrative existed.[7]

**Accentedness and comprehensibility comparisons.** Given the high Cronbach's alpha values for sound clarity, and the lack of group differences in sound clarity ratings, it would seem reasonable to move forward with more traditional analyses of accentedness and comprehensibility. We provide the brief analysis below simply for demonstrative purposes, given that both sets of Speakers are drawn from unrelated studies.

We present mean and standard deviations for accentedness and comprehensibility per group by task in Table 4. First, Spearman's rank correlations indicated a moderate to strong association between accentedness and comprehensibility for both Picture Narrative ($\rho_{Extempore} = .86$, $\rho_{F2F}$

[6] We conducted group comparisons at the task level as a comparison of all F2F recordings against all Extempore recordings would have violated the assumption of independence (i.e., each Speaker per group would have contributed two recordings).

[7] Interpretation of effect sizes follow Plonsky & Oswald's (2014) proposed guidelines for interpreting effect sizes in second language acquisition research.

= .52) and Long Turn ($\rho_{Extempore}$ = .72, $\rho_{F2F}$ = .43) performance. A series of Wilcoxon Signed-Ranks tests found that speakers were more comprehensible than they were nativelike in their speech ($p < .05$), with consistently strong effects ($r > .80$). However, Mann-Whitney U tests indicated no differences between F2F and Extempore Speakers in regards to accentedness or comprehensibility ($p > .05$), though weak effects were found for the picture narrative task ($r_{com}$ = .34, $r_{acc}$ = .26).

**Table 3**

*Mean Ratings (with Standard Deviations) for Sound Clarity by Task and Group (N =14 each)*

| | Picture Narrative | | Long Turn | |
|---|---|---|---|---|
| | F2F (N =14) | Extempore | F2F | Extempore |
| Rating | 7.26 (0.38) | 7.46 (0.66) | 7.30 (0.39) | 7.40 (0.71) |

**Table 4**

*Mean Ratings (with Standard Deviations) for Accentedness and Comprehensibility*

| | Picture Narrative | | Long Turn | |
|---|---|---|---|---|
| | F2F | Extempore | F2F | Extempore |
| Accentedness | 3.38 (0.57) | 3.86 (1.19) | 3.67 (0.44) | 5.88 (0.49) |
| Comprehensibility | 5.24 (0.77) | 6.02 (1.03) | 3.99 (1.25) | 6.27 (0.84) |

## DISCUSSION

In this study, we set out to investigate to what extent L2 speech elicited remotely is comparable to L2 speech elicited in a F2F environment. A review of two reliability coefficients (Cronbach's $\alpha$, ICC 2$k$) indicated generally strong reliability for both remote and F2F speech, with Cronbach's values quite high and on par with previous speech judgment research (Isaacs & Thomson, 2013). In addition, listeners indicated no difference in their perception of sound clarity between F2F and remote speech, with remote speech even receiving slightly more positive scores. Based on these two analyses, it seems initially tempting to assume that L2 speech elicited remotely is indeed comparable to L2 speech elicited in a F2F environment. However, we note caution is still necessary.

For our data, we first calculated correlations between Listeners' ratings of accentedness and comprehensibility across Speakers and tasks, followed by running the non-parametric equivalents to paired- and independent-samples *t*-tests. These are common analyses employed in global speech research (e.g., Crowther et al., 2015, 2018), and, given the acceptable reliability coefficients, justifiable analyses to run. However, it has become common in this line of research to consider the extent to which characteristics of the speech stream (e.g., segmental accuracy, word stress accuracy, speech rate) may inform listeners' perception of the same speech (e.g., Kang, 2010; Kang et al., 2010). To do so, researchers typically draw on correlation and regression analyses that allow them to consider the association between global dimensions and specific acoustic measures of the speech stream. It is at this point that we caution researchers regarding a specific limitation of remote-elicited speech. While our focus in this paper has been at the global level, with an emphasis on listeners' impressionistic judgments, our findings say nothing regarding the comparability of the acoustic signal for more fine-grained analyses (e.g., through the use of the speech analysis software Praat; https://www.fon.hum.uva.nl/praat/). With a focus on linguistic fieldwork, Sanker et al. (2021) investigated this very concern. Their research team simultaneously recorded speakers on 6 different devices (two Macs, a tablet, two cellular phones, a handheld recorder). Unlike our current study, this design allowed Sanker et al. to compare the acoustic properties of the same speech as recorded on multiple devices. Examples of acoustic measures compared across devices included consonant duration, vowel duration, peak f0 timing, and F1 and F2 values. In short, Sanker et al. found that "both device and software altered the recording in ways that affected the retrieval of measurements and the instantiation of contrasts" (p. 25) with some effects being large enough that they could produce misleading phonetic results. Their main implication was "that it will be difficult to directly combine or compare data gathered 'in person' in the fieldwork with data gathered remotely, even if recorded from the same speaker" (p. 26). While remote-elicited speech may still work well for larger, discourse level analyses, Sanker et al's findings raise concerns regarding to what extent remote-elicited speech can allow for the same consistency in discrete-level acoustic analyses compared L2 speech elicited in F2F environment. However, further investigation is necessary.

Additional practical concerns require consideration as well. As referenced earlier, it was necessary to "unlock" Speakers' accounts when they inadvertently locked themselves out of a specific task. Despite being provided step-by-step visual directions on how to navigate

Extempore, some Speakers would still make technical errors, such as clicking the "Stop Recording" button too early or forgetting to finalize the submission of their recording before logging out. While unlocking Speakers' accounts is a simple task, in several cases it allowed Speakers to complete a task multiple times, which in turn raises concerns regarding an unexpected potential effect of task repetition in speech performance (e.g., Lambert et al., 2016). Saito et al. (2018) reported similar technical concerns in their study as well. Saito et al., who made use of the Telephone Standard Speaking Test, referenced dropped calls during their study. The equivalent in our own study, and other potential remote-oriented studies, would be issues with Wi-Fi strength and consistency. These issues, among many others, speak to the potential loss of control referenced in Munro and Derwing (2015), and thus L2 speech researchers considering the use of remote-elicitation must keep in mind several important steps when planning and carrying out their research.

### *Guidelines for Participant Instruction*

As referenced throughout our discussion, remote speech elicitation requires careful planning. We below highlight what we view as an initial list of important guidelines to be considered, with reference to study conceptualization, project design, and participant support. Conceptually, while our findings indicate listener reliability on par with previous research into global dimensions, such as accentedness and comprehensibility, we make no claims regarding the comparability of audio quality and possible acoustic analyses (see Shankar et al., 2021). As such, we caution researchers against assuming that high listener reliability entails the appropriateness of correlation/regression-oriented analyses common to studies interested in how the speech stream informs listener judgments (e.g., Kang, 2010; Kang et al., 2010). Researchers who wish to pursue such analyses with remote-elicited speech are encouraged to be clear in their write ups of the potential limitations of such an approach. It is also important for researchers to consider at this stage whether their target population has both access to the technology necessary to record speech remotely and an environment that provides the necessary conditions for adequate remote elicitation (e.g., quiet environment, strong internet connection).

Given the range of devices that participants may use to record their speech, it is vital to consider project design in reference to multiple platforms (e.g., PC, iPad, Android). It is necessary for researchers to either a) limit participation to those using a single platform type,

which can ensure consistency in the interface across participants, or b) pilot their project across platforms to ensure interface comparability (an issue previously discussed for online questionnaire development; see Dewaele, 2018). Smaller design considerations come into play as well. For example, if participants are required to complete a picture narrative task, researchers need to make sure the details in the picture are large enough to see on the remote interface (i.e., are pictures large and clear enough for participants to accurately describe the narrative? is clarity consistent across different interfaces?). Additional variables of interest include submission medium (e.g., audio, audio + video), submission type (e.g., automatic through app, participant upload), task planning time (e.g., free, controlled), randomization (e.g., is it possible to randomize order of tasks for counterbalancing purposes?), amongst many additional considerations. Key for researchers in this regard is to familiarize themselves with the affordances provided by different apps that allow for speech elicitation, and choose the app that best supports the research objective(s) of their study.

Finally, it is important to understand that by asking participants to submit their speech remotely, we are, in essence, placing an extra burden on them as part of their participation. As such, researchers need to provide as much support as possible to ease participants' experience. We suggest providing an instruction guide with screenshots of each step in the recording, from app registration through recording submission. Depending on participants' proficiency level, we also suggest marking where and when to click to move on to the next screen. In addition, while our study did not include a practice task, given the number of participants who reported issues in completing their recordings, we strongly recommend that a practice task be included before participants are asked to complete any official speaking tasks intended for analysis (e.g., *To complete this task, please say "This is a Test" and click submit*). To ensure participant comfort with engaging with the chosen interface, several rounds of pilot testing and revision are necessary, ideally with members of the target population.

## CONCLUSION

Our goal in this study has not been to advocate for or against the use of speech elicited remotely. Indeed, while our findings indicate initial positive returns in reference to eliciting L2 speech remotely, additional concerns beyond those addressed in our study exist. We hope that by

shedding light on when and how remote elicitation might be utilized in L2 speech research we can help researchers to make informed decisions regarding their own data collection procedures and data analysis practices. While we would still recommend F2F speech elicitation due to the greater degree of control provided to researchers (Munro & Derwing, 2015), it is important that we recognize the growing trends in digital technologies, their associated affordances provided to researchers, and the strengths and weaknesses of these advances. And in some cases (e.g., COVID-19 restrictions), innovation is necessary, though such innovation must be considered critically.

## REFERENCES

Akiyama, Y., & Saito, K. (2016). Development of comprehensibility and its linguistic correlates: A longitudinal study of video-mediated telecollaboration. *The Modern Language Journal*, *100*(3), 585-609. https://doi.org/10.1111/modl.12338

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Crowther, D. (2020). Rating L2 speaker comprehensibility on monologic vs. interactive tasks: What is the effect of speaking task type? *Journal of Second Language Pronunciation, 6*(1), 96-121. https://doi.org/10.1075/jslp.19019.cro

Crowther, D., Holden, D., & Urada, K. (2022). Second language speech comprehensibility. *Language Teaching.* Published online 12 May 2022. https://doi.org/10.1017/S0261444821000537

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015). Second language comprehensibility revisited: Investing the effects of learner background. *TESOL Quarterly, 49*(4), 814-837. https://doi.org/10.1002/tesq.203

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition, 40*(2), 443-457. https://doi.org/10.1017/S027226311700016X

Derwing, T. M. & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *20*(1), 1-16. https://doi.org/10.1017/S0272263197001010

Derwing, T. M., & Munro, M. J. (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching, 44*(3), 316-327. https://doi.org/10.1017/S0261444811000103

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*(4), 533-557. https://doi.org/10.1017/S0272263109990015

Dewaele, J.-M. (2018). Online questionnaires. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 269-286. Palgrave Macmillan.

International English Language Testing System. (2009). *Cambridge IELTS 7: Examination papers from University of Cambridge ESOL Examinations: English for speakers of other languages.* Cambridge University Press.

International English Language Testing System. (2011). *Cambridge IELTS 8: Examination papers from University of Cambridge ESOL Examinations: English for speakers of other languages.* Cambridge University Press.

Inceoglu, S. (2019). Exploring the effects of instruction on L2 French learner pronunciation, accentedness, comprehensibility, and fluency: An online classroom study. *Journal of Second Language Pronunciation*, *5*(2), 223-246. https://doi.org/10.1075/jslp.18004.inc

Isaacs, T., & Thomson, R. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*(2), 135-159. https://doi.org/10.1080/15434303.2013.769545

Isbell, D. (2018). Assessing pronunciation for research purposes with listener-based numerical scales. In O. Kang & A. Ginther (Eds.), *Assessment of second language pronunciation* (pp. 89-112). Routledge.

Isbell, D., Park, O.–S., & Lee, K. (2019). Learning Korean pronunciation: Effects of instruction, proficiency, and L1. *Journal of Second Language Pronunciation, 5*(1), 13-48. https://doi.org/10.1075/jslp.17010.isb

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, *38*(2), 301–315. https://doi.org/10.1016/j.system.2010.01.005

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*(4), 554-566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, *68*(1), 115-146. https://doi.org/10.1111/lang.12270

Lambert, C., Kormos, J., & Minn, D. (2016). Task repetition and second language speech processing. *Studies in Second Language Acquisition, 39*(1), 167-196. https://doi.org/10.1017/S0272263116000085

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). Routledge.

Levis, J. (2020). Changes in L2 prionunciation: 25 years of intelligibility, comprehensibility, and accentedness. *Journal of Second Language Pronunciation, 6*(3), 277-282. https://doi.org/10.1075/jslp.20054.lev

MacIntyre, P. D. (2012). The idiodynamic method: A closer look at the dynamics of communication traits. *Communication Research Reports, 29*(4), 361–367. https://doi.org/10.1080/08824096.2012.723274

Mackey, A. (1999). Input, interaction, and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition*, *21*(4), 557-587. https://doi.org/10.1017/S0272263199004027

Mackey, A. (2002). Beyond production: Learners' perceptions about interactional processes. *International Journal of Educational Research*, *37*(3-4), 379-394. https://doi.org/10.1016/S0883-0355(03)00011-9

Moran Wilson, M., Kang, O., & Kermad, A. (2019). *The effects of length of speech stimuli on listener perception in speech evaluation research*. Paper presented at the annual American Association for Applied Linguistics conference, Atlanta, Georgia, USA.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, intelligibility, and comprehensibility in the speech of second language learners. *Language Learning, 45*(1), 73-97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (2015). A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation*, *1*(1), 11-42. https://doi.org/10.1075/jslp.1.1.01mun

Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication, 52*(5-7), 626-637. https://doi.org/10.1016/j.specom.2010.02.013

Nagle, C., & Rehman, I. (2021). Doing L2 speech research online: Why and how to collect online ratings data. *Studies in Second Language Acquisition, 43*(4), 916-939. https://doi.org/10.1017/S0272263121000292

Nagle, C., Trofimovich, P., & Bergeron, A. (2019). Toward a dynamic view of second language comprehensibility. *Studies in Second Language Acquisition, 41*(4), 647-672. https://doi.org/10.1017/S0272263119000044

O'Brien, M. G. (2016). Methodological choices in rating speech samples. *Studies in Second Language Acquisition, 38*(3), 587-605. https://doi.org/10.1017/S0272263115000418

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878-912. https://doi.org/10.1111/lang.12079

Rose, H., & Galloway, N. (2019). *Global Englishes for language teaching*. Cambridge University Press.

Saito, K., & Akiyama, Y. (2017). Video-based interaction, negotiation for comprehensibility, and second language speech learning: A longitudinal study. *Language Learning*, *67*(1), 43-74. https://doi.org/10.1111/lang.12184

Saito, K., Dewaele, J. M., Abe, M., & In'nami, Y. (2018). Motivation, emotion, learning experience, and second language comprehensibility development in classroom settings: A cross-sectional and longitudinal study. *Language Learning*, *68*(3), 709-743. https://doi.org/10.1111/lang.12297

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics 37*(2), 217-240. https://doi.org/10.1017/S0142716414000502

Saito, K., Trofimovich, P., Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics, 38*(4), 439-462. https://doi.org/10.1093/applin/amv047

Sanker, C., Babinski, S., Burns, R., Evans, M., Johns, J., Kim, J., Smith, S., Weber, N., & Bowen, C. (2021). (Don't) try this at home! The effects of recording devices and software on phonetic analysis. LingBuzz. https://lingbuzz.net/lingbuzz/005748

Southwood, M. H., & Flege, J. E. (1999). Scaling foreign accent: Direct magnitude estimation versus interval scaling. *Linguistics & Phonetics, 13*(5), 335-349. https://doi.org/10.1080/026992099299013

Stemler, S. E., & Tsai, J. (2008). Best practices in estimating interrater reliability. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Sage.

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143-167. https://doi.org/10.1017/S0272263119000421

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition, 15*(4), 905-016. https://doi.org/10.1017/S1366728912000168

Trofimovich, P., Nagle, C. L., O'Brien, M. G., Kennedy, S., Taylor Reid, K., & Strachan, L. (2020). Second language comprehensibility as a dynamic construct. *Journal of Second Language Pronunciation, 6*(3), 430-457. https://doi.org/10.1075/jslp.20003.tro

# TEST REVIEW: THE VALIDITY OF WRITING SECTIONS ACROSS FIVE GRADES IN THE EIKEN TEST

**HIKARU ISHIYAMA**

***University of Hawai'i at Mānoa***

## INTRODUCTION

Despite the widespread use of English standardized tests such as the *Test of English as a Foreign Language* (TOEFL) and the *International English Language Testing System* (IELTS), some countries in the so-called outer and expanding circles (Kachru, 1985) still have their own English language proficiency tests. The *Eiken Test in Practical English Proficiency* (実用英語技能検定 *jitsuyo eigo ginou kentei*; hereafter Eiken*)* was developed in 1963 by the *Society for Testing English Proficiency* (STEP), a Japanese non-profit organization affiliated with the Ministry of Education, Culture, Sports, Science, and Technology (MEXT). With the governmental support, the test has achieved prominence in Japan. The number of test takers has increased as the use of the test results has been expanded. During the 2020 school year, the number of test takers reached more than three million (more information available at https://www.eiken.or.jp/eiken/merit/situation/). The upper levels among the seven levels of Eiken called *Grade* have been used as a language proficiency test for university admissions not only for Japanese universities but also for the University of Hawaii and some other English-dominant universities overseas. Given that tests can give a considerable influence on what is taught at school (i.e., washback; Bailey, 1996), it is imperative to examine whether it measures what it is supposed to measure as well as if the test use is appropriate in the target context (Chapelle, 2012). Previous literature examining Eiken tests predominately focused on the reading and listening sections (Chujo & Oghigian, 2009; Hamada, 2015; Miura & Beglar, 2002; Piggin, 2011; Plumb & Watanabe, 2016). This may reflect the Japanese examination culture, where receptive skills (i.e., reading and listening) are prioritized over productive skills (i.e., speaking and writing). To fill the lacuna in the review of Eiken tests, the present paper focuses on the validity of writing sections across levels and provides information for stakeholders in terms of the use of the test results for their purposes.

## TEST PURPOSE AND USE

Eiken covers a range of proficiency levels, from elementary to advanced, divided into seven distinctive levels called *Grade*: Grade 5, 4, 3, Pre-2, 2, Pre-1, and 1. Each Grade measures English proficiency in four skills (i.e., listening, reading, speaking, and writing) except for the lowest two levels, which measure only listening and reading. STEP provides the correspondence of each Grade to school grades on its website (https://www.eiken.or.jp/eiken/exam/about/). As the targets of test takers are different depending on Grades, the use of each test varies from an achievement test for school graders to a proficiency test for college admissions. Because of the association between school curriculum and Eiken Grades, a large number of Japanese secondary schools have conventionally used Eiken as achievement tests up to Grade 2 (high school graduation level). According to the report provided by STEP (https://www.eiken.or.jp/eiken/merit/situation/), among all test takers in 2020, the number of "secondary school students" was the highest (2,911,389), followed by "others" (383,598), "elementary and below" (325,390), and "college students" (57,784).

Since STEP promoted Eiken use as an academic proficiency test in 2007 at the JALT conference, a number of tertiary schools have accepted the top three Grades as language proficiency certificates for university admissions (Piggin, 2011). For example, the applicants who have a certain Grade are given additional points to their scores of a college entrance exam or are even exempted from taking entrance exams. Not only Japanese universities but also universities overseas have recognized Eiken as a good indicator of applicants' language proficiency levels. As of April 2021, over a hundred four-year universities in the U.S., as well as a large number of public high schools in Australia, have accepted the upper three Grades (i.e., Grade 2, Pre-1, and 1) (https://www.eiken.or.jp/eiken/abroad/about/). Likewise, since the University of Hawaii-West O'ahu accepted Eiken as a language assessment for international applicants in 2018, all campuses of the University of Hawaii have been using Eiken results as evidence of English proficiency. In addition to advantages for college admissions, the holders of the upper Grades benefit in career opportunities as well. For example, the Japan Tourism Agency supported by the Japanese government allows Grade 1 holders to be exempted from taking a language test (Japan National Tourism Organization, 2022).

## TEST METHOD

As Table 1 shows, each grade has a different test structure (STEP, n.d.). The number bolded indicates the number of items in each task and the time in parenthesis shows the time allotted for each task. For example, Grade Pre-2 has 37 items of vocabulary and reading comprehension in the reading section, one independent writing task in the writing section, 30 items in the listening section, and five items in the speaking section. Reading and writing are integrated as one section, so examinees need to complete both sections in 75 minutes. All questions in the reading and listening sections are multiple-choice and examinees mark one of four options on their answer sheet. As for the writing section, the examinees write with a pencil on the back of their mark sheet. According to information provided on the website, writing is graded by a trained rater with an analytical rating (STEP, n.d.). The examinees first receive the result of the first stage comprised of reading, writing, and listening. Only those who passed the first stage can take the speaking test about a month later at an appointed test center. The speaking test is conducted face-to-face with one examiner except for Grade 1, to which two examiners are assigned.

**Table 1**

*The Eiken Test Format (adapted from https://www.eiken.or.jp/eiken/exam/)*

| | Reading + Writing | Listening | Speaking | Total |
|---|---|---|---|---|
| 1 | **41** + **1** Essay Writing (100 min) | **27** (35 min) | **1** (10 min) | 145 min |
| Pre-1 | **41** + **1** Essay Writing (90 min) | **29** (30 min) | **5** (8 min) | 128 min |
| 2 | **38** + **1** Essay Writing (85 min) | **30** (25 min) | **5** (7 min) | 117 min |
| Pre-2 | **37** + **1** Sentence Writing (75 min) | **30** (25 min) | **6** (6 min) | 106 min |
| 3 | **40** + **1** Sentence Writing (50 min) | **30** (25 min) | **6** (5 min) | 80 min |
| 4 | **35** (35 min) No writing | **30** (30 min) | N/A | 65 min |
| 5 | **25** (25 min) No writing | **25** (25 min) | N/A | 50 min |

The test is administered three times a year, January, May, and October. While student examinees normally take the test in their own classrooms, non-school-based examinees take the test at a test center. More than 400 test centers are placed across Japan. Moreover, due to the expanded use of the test for university admissions overseas, Eiken tests are now administered in London, New York, Los Angeles, and Honolulu.

## SCORES

As for the scoring system, STEP provides comprehensive information on its website (available at https://www.eiken.or.jp/cse/). The Eiken test is a pass-or-fail system based on the *Common Scale of English* (CSE), a scoring system originally developed by STEP in 2014. The CSE score sets 4,000 points at the highest language proficiency level and zero at the lowest. Grade Pre-1 holders, for example, are in the range of 2,305 points to 3,000 points. STEP also assigns a range of points to each of the bands in *Common European Framework of Reference for Languages* (CEFR). For example, C2 level is from 3,300 points to 4,000 points, C1 is from 2,600 points to 3,300 points, B2 is from 2,300 points to 2,600 points, and it goes on to A1. The association of the Eiken CSE score and the CEFR made different language tests comparable to each other. For instance, Grade Pre-1 holders have at least 2,305 points in the CSE score, which corresponds to B2 in CEFR. As CEFR bands correspond to other standardized tests, Eiken's results can be compared to scores from other standardized tests such as TOEFL and IELTS.

Since the implementation of CSE scores into Eiken, the result for a pass or a fail has been decided based on a cut-score. The cut score is calculated from tests administered previously and is different depending on the Grade but fixed at each Grade. For example, the cut-score for the first stage (i.e., reading, listening, and writing) of Grade 3 is 1103 points, and the one for the second stage (i.e., speaking) is 353 points. It means that the cut-score of Grade 3 is 1456 points in total out of 2200 points. Each of the four skills has the same full score so that the same value is weighed for each skill. Grade 1, for example, has 3,400 points in total and 850 points for each skill. The test takers receive their scores in each task as well as the pass-or-fail result.

## TEST COST

The cost is different depending on the Grade: 11,800 yen for Grade 1; 9,800 yen for Grade Pre-1; 8,400 yen for Grade 2; 7,900 yen for Grade Pre-2; 6,400 yen for Grade 3; 4,500 yen for Grade 4; 3,900 yen for Grade 5 (as of June 2022; https://www.eiken.or.jp/eiken/schedule/). It is relatively reasonable compared to other language tests. Even the highest cost for Grade 1 is less than half the price of TOEFL iBT. This explains one of the reasons Eiken is still popular among

Japanese people. More information about the test is available both in Japanese (https://www.Eiken.or.jp/Eiken/) and in English (https://www.Eiken.or.jp/Eiken/en/).

## VALIDITY

Against the backdrop of expanding the use of results from the Eiken examination, the validity of the Eiken test has been examined both by its developers (Brown et al., 2010) and researchers (Chujo & Oghigian, 2009; Hamada, 2015; Miura & Beglar, 2002; Piggin, 2011; Plumb & Watanabe, 2016). Most of these studies, however, focused on vocabulary, reading, and listening sections. No review has been conducted on the writing section of the Eiken examination to the extent of my knowledge. Moreover, regardless of the seven-staged examination, no discussion has been published to examine how the levels of each grade are structured and distinguished from each other.

To fill in the gap in the previous literature, the rest of the present paper will be devoted to considering the validity of the writing sections across levels. Validity in language testing can be broadly defined as "an evaluation of the credibility, or plausibility, of the proposed interpretations and uses of test scores" (Kane, 2010, p. 180). To establish sound validity, test makers should explicitly propose what the test measures as well as how the results of the test should be used for any decision making. In a book overviewing writing tasks in the Cambridge suite of examinations, Shaw and Weir (2007) proposed five key factors for the writing test validation. They include *cognitive validity, context validity, criterion-related validity, scoring validity,* and *consequential validity*. In this paper, *cognitive validity* and *criterion-related validity* will be excluded from the focus as both need the examination of the examinees' actual writing performance. *Cognitive validity* is defined as "how closely [a writing task] represents the cognitive processing involved in writing contexts beyond the test itself" (Shaw & Weir, 2007, p. 34). To validate it, researchers need to investigate how examinees employ cognitive processing such as macro-planning, organization, micro-planning, translation, monitoring, and revising while they perform a writing task (Guapacha Chamorro, 2022; Shaw & Weir, 2007). *Criterion-related validity* is concerned with "the extent to which test scores correlate with a suitable external criterion of performance with established properties" (Shaw & Weir, 2007, p. 6). Criterion-related validity also requires the examinees' performance. Thus, the present paper will

only review the other three aspects of validity (i.e., context validity, scoring validity, and consequential validity).

### *Context Validity*

*Context validity* is the term employed by Shaw and Weir (2007), referring to the traditional notion of *content validity*. In their definition, *context validity* "relates to the linguistic and content demands that must be met for successful task realization and to features of the task setting that serve to describe the performance required" (Shaw & Weir, 2007, p. 63). Linguistic demands refer to test takers' ability to demonstrate lexical and structural resources, discourse mode, and content knowledge. Task settings refer to the conditions such as instructions, writing purpose, text length, time allotment, writer-reader relationship, and physical conditions (Shaw & Weir, 2007). Each grade in Eiken is created in accordance with a criterion of writing skill (STEP, n.d.), as is shown in Table 2. The upper two levels, for example, have the terms *logically* in the statement. This indicates that the examinees are expected to write an essay with a certain amount of words with coherence and cohesion. Although each writing construct has different words, it is hard to differentiate from one another. For example, the only difference in the constructs between Grade 1 and Grade Pre 1 is the word "*wide*." This may not be sufficient to proclaim that Grade 1 holders have different levels of writing ability from Grade-Pre 1 holders. STEP should provide more concrete writing constructs for the different five grades. The discussion below is going to address how each criterion is operationalized in writing tasks of each grade.

**Table 2**

*Criteria of Writing Skills of Eiken Grades (adapted from https://www.eiken.or.jp/eiken/exam/criteria/)*

| | |
|---|---|
| 1 | Can write logically on a wide range of socially relevant topics |
| Pre-1 | Can write logically on a range of socially relevant topics |
| 2 | Can write on socially relevant topics |
| Pre-2 | Can write on topics from everyday life |
| 3 | Can write simple texts about himself/herself |

**Linguistic Demands.** Five levels of writing tasks are individualized in the writing prompts. The prompts shown in Table 3 were taken from the previous Eiken tests held in January 2021 (Previous Eiken tests are available on its website; https://www.eiken.or.jp/eiken/exam/). The main difference between the upper two levels and the lower three levels is the expected writing genre. Examinees of Grade 1 and Grade Pre-1 are specifically instructed to write an argumentative essay with a traditional essay structure: introduction, main body, and conclusion. Given that most secondary school students rarely have a chance to learn an essay structure either in their first language or second language (Kobayashi & Rinnert, 2002), it is reasonable that only the top two grades require a clear essay structure. For the lower three levels, examinees are only expected to provide their opinion with two reasons to support it.

The level of vocabulary used in prompts and the *topic* of each grade also differentiate the upper grades from the lower. The *topic* for the Grade 1 examination held in January 2021, for example, was, "Are economic sanctions a useful foreign-policy tool?" If examinees do not know the meaning of *sanctions*, they may not be able to write anything. In addition to broad lexical knowledge, they also need to have content knowledge about politics. As for Grade Pre-1, the degree of difficulty of the prompt type is similar to Grade 1. However, it provides some scaffoldings called *points* that may help examinees come up with their answers.

Levels for lower grades seem to vary in difficulty depending on to what extent the question is personal. The prompt in Grade 2 asks examinees' opinions about a non-personal topic such as the food waste from restaurants and supermarkets. On the contrary, the prompt in Grade 3 asks about personal experience and can be answered without background knowledge.

**Setting of the Writing Tasks.** As for the settings of writing tasks, levels are constructed with instructional language, text length, and time allotment. Instructions are provided in English for the upper three levels while in Japanese for the lower two levels. As for the suggested text length, it varies from 25-35 words for Grade 3 to 200-240 words for Grade 1. Given that all grades are paper-and-pencil instead of computer-based writing, Eiken does not seem to differentiate levels based on test formats.

**Table 3**

*Writing Tasks in Eiken (adapted from https://www.eiken.or.jp/eiken/exam/)*

| | Prompts and instructions provided in the writing test administered in January 2021 | Suggested length | Instruction |
|---|---|---|---|
| 1 | *Topic:* "Are economic sanctions a useful foreign-policy tool?"<br>Write an essay on the given TOPIC.<br>Give THREE reasons to support your answer.<br>Structure: introduction, main body, and conclusion | 200- 240 words | English |
| Pre-1 | *Topic:* "Agree or disagree: Big companies have a positive effect on society"<br>Write an essay on the given TOPIC.<br>Use TWO of the POINTS below to support your answer. [Points: products, the economy, the environment, work-life balance]<br>Structure: introduction, main body, and conclusion | 120- 150 words | English |
| 2 | *Topic:* "It is often said that restaurants and supermarkets should try to reduce the amount of food that they throw away. Do you agree with this opinion?"<br>Write your opinions about the topic and provide two reasons to support your answer.<br>*Points* are provided to help you with your writing. You may write from other perspectives.<br>[Points: Cost, health and safety, and the environment]<br>When you do not answer the *topic*, your writing may not receive a grade. Please read the topic carefully. | 80- 100 words | Japanese |
| Pre-2 | *Question:* "Do you think it is a good idea for people to have a car?"<br>You are asked about the *question* by your foreign friend.<br>Write your opinions about the *question* and provide two reasons in English.<br>When you do not answer the *question*, your writing may not receive a grade. Please read the topic carefully. | 50- 60 words | Japanese |
| 3 | *Question:* "Where do you like to go shopping?"<br>You are asked about the *question* by your foreign friend.<br>Write your opinions about the *question* and provide two reasons in English.<br>When you do not answer the *question*, your writing may not receive a grade. Please read the topic carefully. | 25- 35 words | Japanese with Furigana (i.e., Japanese reading aid) |

***Scoring Validity***

Scoring validity is related to "all the aspects of the testing process that can impact on the reliability of test scores" (Shaw & Weir, 2007, p. 143). It can be established when scoring procedures such as the scoring guide accurately reflect the writing construct being measured (Weigle, 2002).

For the evaluation of writing performance in Eiken, a similar rubric consisting of four criteria is used at all five Grades: Grade 3 (https://www.eiken.or.jp/eiken/exam/2017scoring_3w_info.html), Grade Pre 2 (https://www.eiken.or.jp/eiken/exam/2017scoring_p2w_info.html), and the other three upper levels (https://www.eiken.or.jp/eiken/exam/2016scoring_w_info.html). The criteria include *content, organization, vocabulary,* and *grammar. Content* pertains to clearness, persuasiveness, and concreteness of the content. *Organization* is judged with the effective use of expressions such as conjunctions. *Vocabulary* is evaluated with accuracy of spelling and meaning for the lower two levels (i.e., Grade 3 and Grade Pre-2) and the appropriate use of words and a variety of words for the upper three levels (i.e., Grade 2, Grade Pre-1, and Grade 1). *Grammar* is about grammatical accuracy and a variety of sentence patterns for all levels except for Grade 3, which is only evaluated with grammatical accuracy. Each criterion has zero to four points and thus the full score of a writing section is sixteen. Grade 1 has the same four criteria but each criterion has eight points so the full score is 32. The score is then calculated to a CES score.

The rubric for writing assessment used in Eiken is similar to the one used for an independent writing task in other standardized language tests. A rubric for an independent writing task in IELTS, for example, has criteria of *task response*, *coherence and cohesion*, *lexical resource*, and *grammatical range and accuracy*. The distinctive feature of Eiken is, however, its adoption of an analytic scoring approach. It allows raters to assign separate scores to each criterion in a writing rubric. The strength of this scoring approach is its detailed information about a test taker's performance (Weigle, 2002). On the other hand, TOEFL and IELTS employ a holistic scoring approach in which raters provide only one score for overall writing (Crusan, 2013). Because of the cost- and time effectiveness, holistic scoring has been preferred for large-scale and high-stakes testing settings.

The reliability of a scoring approach seems to vary depending on writing task types. Ohta et al. (2018) found that a holistic approach yields higher reliability for integrated writing tasks

while earlier studies concluded that an holistic scoring approach was more reliable for independent writing tasks (Barkaoui, 2007; Schoonen, 2005). Since Eiken only has an independent writing task for all Grades, holistic scoring might be more appropriate to adopt. In terms of the scoring approach for the lower grades, however, analytic scoring would be more beneficial given that the lower grades are utilized as an achievement test at school. When the result is used for an educational purpose, detailed information about writing performance would help test takers develop their writing skills. Currently, STEP only reports a total score that encompasses all four criteria. To make the test more instructional, the separate scores in each dimension should be reported so that the examinees can use them as feedback.

### *Consequential Validity*

The validity of tests should also be considered from the consequential aspects including "evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term" (Messick, 1996, p. 251). Because the consequence of the test results is more significant for Grade Pre-1 and Grade 1, this section focuses on the validity of the test use for the top two grades.

According to CSE score, Grade 1 is equivalent to TOEFL 95-120, as well as IELTS Band 7.0-8.0. Grade Pre-1 corresponds to TOEFL 72-94 and IELTS 5.5-6.5. When writing sections in Eiken Grade 1 and Grade Pre-1 are compared to TOEFL and IELTS, which were specifically developed for university admissions, Eiken seems less demanding in the following three points. Firstly, Eiken has only one independent writing task (i.e., examinees write their response to a statement or a question) while the other two tests have two different tasks with two different rubrics. TOEFL, for example, has an integrated writing task in addition to an independent test. The integrated writing task requires test takers to read or listen to a certain passage and synthesize the information into their writing. This test is more demanding compared to the other in that it entails additional skills other than writing skills. Given that college students write a paper based on lectures and readings, an integrated writing task reflects real-world situations in an academic environment more than an independent writing task. Eiken's writing section with only an independent writing task, thus, may not be able to infer test takers' ability to write in an academic environment accurately.

Secondly, the suggested word length in the Eiken writing section is much shorter than in TOEFL and IELTS. Eiken requires 200-240 in Grade 1 and 120-150 words in Grade Pre-1 while TOEFL and IELTS suggests at least 300 words and 250 words respectively. Since TOEFL and IELTS have two sets of writing questions, examinees of the tests write more than 500 words. Although Eiken tries to measure examinees' writing skills to structure an essay with introduction, body, and conclusion, the short amount of writing would not likely be enough to achieve an accurate measurement.

Thirdly, all writing sections in Eiken are provided with hand-writing. This can also be a limitation to infer test takers' writing performance in academic settings. Given the growing exposure to technology, handwriting skills may thwart fair and valid assessment of writing (Barkaoui & Knouzi, 2018). If the results of higher levels such as Grades 1 and Grade Pre-1 are used as a language certificate for college entrance, it should also consider the implementation of computer-based testing. When it comes to the use of Eiken for university admissions, thus, the writing task format should be reexamined. It should take into consideration real-world needs that examinees would face in academic environments.

## Conclusion

Since its foundation in 1963, Eiken has endeavored to accommodate test structures to socio-temporal demands in the globalized era. It added two more grades between Grade 3 and Grade 2, and Grade 2 and Grade 1, and also introduced writing sections to lower grades. Thanks to the flexible changes in its format and scoring system, the test has now been used not only as an achievement test but also as a proficiency test for universities in Japan and overseas.

This paper reviewed the validity of writing sections across grades. Writing sections have a gradual increase in their linguistic demands and task settings. However, in terms of consequential validity, upper grades still need to be revised. To make the two grades more reliable and valid, the length of words, the task type, and the writing mode (handwriting vs. computer-based writing) should be reexamined.

## REFERENCES

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257–279. https://doi.org/10.1177%2F026553229601300303

Barakaoui, K., & Knouzi, I. (2018). The effects of writing mode and computer ability on L2 test-takers' essay characteristics and scores. *Assessing Writing, 36*, 19–31. https://doi.org/10.1016/j.asw.2018.02.005

Brown, J. D., Davis, J.M., Takahashi, C., & Nakamura, K. (2010). Upper-level Eiken examinations: Linking, validating, and predicting TOEFL iBT scores at advanced proficiency Eiken levels. *Eiken Foundation of Japan*. Retrieved from https://www.eiken.or.jp/eiken/group/result/pdf/eiken-toeflibt-report.pdf

Chapelle, C. A. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 21–33). Routledge.

Chujo, K., & Oghigian, K. (2009). How many words do you need to know to understand TOEIC, TOEFL & EIKEN? An examination of text coverage and high frequency vocabulary. *Journal of AsiaTEFL*, *6*(2), 1–231.

Crusan, D. (2013). Assessing Writing. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 201–215). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla067

Guapacha Chamorro, M. E. (2022). Cognitive validity evidence of computer- and paper-based writing tests and differences in the impact on EFL test-takers in classroom assessment. *Assessing Writing*, *51*, 100594. https://doi.org/10.1016/j.asw.2021.100594

Hamada, A. (2015). Linguistic variables determining the difficulty of Eiken reading passages. *JLTA Journal*, *18*(0), 57–77. https://doi.org/10.20622/jltajournal.18.0_57

Japan National Tourism Organization. (2022). The examination guidelines for the national translators (lit. Zenkoku Tsuyaku Annaishi Shiken). https://www.jnto.go.jp/jpn/projects/visitor_support/interpreter_guide_exams/operation_2022.pdf

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the world:*

*Teaching and learning the language and literatures* (pp. 11–30). Cambridge University Press.

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182. https://doi.org/10.1177/0265532209349467

Kobayashi, H., & Rinnert, C. (2002). High school student perceptions of first language literacy instruction: Implications for second language writing. *Journal of Second Language Writing*, *11*(2), 91–116. https://doi.org/10.1016/S1060-3743(02)00067-X

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256. https://doi.org/10.1177/026553229601300302

Miura, T., & Beglar, D. (2002). The Eiken vocabulary section: An analysis and recommendations for change. *JALT Journal*, *24*(2). https://doi.org/10.37546/JALTJJ24.2-1

Ohta, R., Plakans, L. M., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing*, *38*, 21–36. https://doi.org/10.1016/j.asw.2018.08.001

Piggin, G. (2011). An evaluative commentary of the grade 1 EIKEN test. *Language Testing in Asia*, *1*(4), 144–167. https://doi.org/10.1186/2229-0443-1-4-144

Plumb, C., & Watanabe, D. (2016). A critique of the Grade 2 Eiken test reading section: Analysis and suggestions. *Shiken*, *20*(1), 12–17.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing.* Cambridge University Press.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1–30. https://doi.org/10.1191/0265532205lt295oa

Society for Testing English Proficiency. (n.d.-a). Test structures of each grade. https://www.eiken.or.jp/eiken/exam/

Society for Testing English Proficiency. (n.d.-b). Writing assessment policy https://www.eiken.or.jp/eiken/exam/eiken-scoringquality.html

Society for Testing English Proficiency. (n.d.-c). Eiken criteria https://www.eiken.or.jp/eiken/exam/criteria/

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

# PARENTS' PERCEPTIONS AND EXPERIENCES OF EARLY ENGLISH EDUCATION IN SOUTH KOREA: A FOCUS ON ENGLISH KINDERGARTENS

**SOO JIN LIM**
**University of Hawai'i at Mānoa**

## ABSTRACT

The purpose of this study is to investigate parents' perceptions and experiences of early childhood English education in South Korea based on their children's private English kindergarten education. It aims to explore parental opinions about early English language learning. In addition, the study intends to discover key factors that influence their decisions in selecting private English kindergartens. Finally, it attempts to determine their level of satisfaction with their children's English kindergarten experiences and language learning outcomes. A combination of quantitative and qualitative data gathering methods were used in the study. The data were collected through a questionnaire given to 30 participants and semi-structured interviews with 8 participants. As a result, participants perceived that learning English at an early age is beneficial for children's future academic and career success. The findings suggest that the current trend of private English kindergartens, highly accessible only to children from wealthy families, reflects the strong demands of parents for early English education as the Korean education system does not allow teaching English in kindergartens. To provide equal opportunities for all children to learn English in the current era of globalization, it is necessary to integrate English education into the national kindergarten curriculum.

*Keywords:* early English education, second language learning, private English kindergartens in Korea, academic-oriented learning, play-based learning

As English has become the modern-day lingua franca, early childhood English education has been gaining much attention in South Korea (Korea hereafter). There is a widespread belief among parents that young learners can acquire a second or foreign language better than adults do, and more and more Korean parents want their children to start learning English early.

However, despite parents' demands for early English education, the national kindergarten curriculum does not include English language education. The Korean curriculum mandates that English must be taught from the third grade in public education in order to normalize public education and to prohibit teaching ahead of the regular public curriculum. Nevertheless, Korean parents continue to provide English-learning opportunities for their children through private English education, an emerging market today.

The present study introduces a unique educational institution called 'English kindergarten,' a private institution that specializes in teaching English to young children in Korea. Private English kindergartens are run by either individuals or educational companies, not bound by the regulations of the Ministry of Education. A growing number of parents enroll their children in private English kindergartens with high aspirations and expectations for their children's English language learning. In particular, early English learning through private English kindergarten education is an upward trend among parents in Gangnam and Seocho Districts, the most affluent parts of Seoul, the capital of Korea.

Nevertheless, there is little research that examines the effectiveness of early English education on young learners' English language learning or that explores parents' experiences. Therefore, this paper aims to investigate parents' perceptions of early English education and analyze in-depth their perceptions and opinions based on their children's English kindergarten education, particularly focusing on the need and goals for English language learning, factors that influence parents' decisions in selecting English kindergartens, and parents' satisfaction with the English kindergarten programs and their children's English language outcomes after completing the programs.

## LITERATURE REVIEW

### *Social Background of Early English Education in Korea*

**English Education Fever in Korea: Historical Background.** Rising from one of the poorest countries in the world with almost no natural resources to a developed, high-income country in just a few decades, the Republic of Korea has overcome several national crises and shown rapid economic development since its first establishment in 1948. It is important to note that education was one of the few ways to gain status and power in society at that time and has

played a significant role in the development of Korea. Based on these social and cultural factors, education has been valued for centuries in Korea. The importance of education grew into Korean parents' educational aspirations, a so called 'education fever', or "the national obsession with education (Seth, 2002, p. 9)," for their children's future lives. Korea ranked 9th in the 2020 global gross domestic product (GDP) rankings (Choi, 2020) with one of the world's highest-educated labor forces among the Organisation for Economic Co-operation and Development (OECD) countries (OECD, 2018). The historical background of Korea's economic development explains the current pursuit of education in Korea.

During the 1960s and 1970s, Koreans who left for study abroad in English-speaking countries returned to Korea and became the mainstream of politics, economy, and society, leading to an increase in the study abroad population for university degrees in the 1980s (Park, 2009). This phenomenon contributed to the process in which English was recognized as an important social resource in Korean society. In the 1990s, when the Korean government decided to include English Language as a mandatory subject for the national college entrance exam, College Scholastic Ability Test (CSAT), along with Korean and Mathematics, English education emerged as a serious social issue (Kwon et al., 2017).

With an aim to develop students' communicative competence by fostering their interest in English language and culture, the Korea National English Curriculum also included English in regular curriculum for primary English education in 1997 (Korean Ministry of Education). Since then, the number of Korean students going abroad to English-speaking countries for early English education rapidly increased. This trend of early study abroad even created a new form of "wild-geese family," in which children study overseas, accompanied by mothers to take care of them, and fathers stay in Korea to work and provide financial support for the family (Koo & Lee, 2006).

**The Emergence and Growth of Private English Kindergarten Market in Korea.** In today's Korean society, English proficiency is acknowledged as a key to success in academics and jobs, as a decent job or high social position requires a good educational background and a strong professional network of colleagues and coworkers. Thus, most Korean students desire to enroll in elite universities such as Seoul National University, Korea University, and Yonsei University (collectively referred to with the acronym 'SKY'), the top three universities in Korea. In the study of Lee et al. (2020), Korean mothers indicated that the most important reason for

their children's English education is to get a good score on the CSAT. As a result, English proficiency has been regarded as a critical asset for academic achievement that led to occupational success in Korea and created a huge private English education market. According to the data compiled by Statistics Korea (2021), the total expenditure on private education for English language amounted to around 6.14 trillion Korean won ($5.2 billion US dollars).

However, despite the 'English education fever' and a high demand for early English education among Korean parents, the Ministry of Education does not allow teaching English in kindergartens and in the first and second grade curricula in public education due to the concern that teaching young learners Korean and English simultaneously could hinder their development of Korean proficiency (Jung, 2019). However, while there are parallel debates about the benefits of early foreign language education (Muñoz, 2014), there is very little evidence to support the idea that learning a second language early has a negative effect on the first language development. Nevertheless, the current Korean curriculum mandates that English must be taught from the third grade in public education. The governmental ban on English education for young children and parents' concerns about their children's future led to the emergence of a unique educational institution system called 'private English kindergarten' in the early 2000s.

The so-called English kindergarten is a private institution that specializes in teaching English to young children without being bound by the regulations of the Ministry of Education. The number of English kindergartens has been rapidly increasing over the past 20 years, especially in the past decade. In 2009, there were 181 English kindergartens across the country and 66 of them were located in Seoul, the capital city of Korea. As of January 2020, there are about 558 English kindergartens nationwide,[8] and 84 out of 288 in Seoul are located in Gangnam and Seocho Districts, the most affluent parts of the city (The World Without Concerns for Private Education, 2020). These private English kindergartens require tuition fees that are higher than the average college tuition fee.[9] Therefore, access to this type of education is very limited and depends on the socioeconomic status of the family. It is generally available to children from middle- and upper-class families, as well as children whose parents are white-collar workers or entrepreneurs.[10]

[8] The number of Korean kindergartens is 8,659 (Korean Educational Statistics Service, 2021).
[9] The average of monthly tuitions for private English kindergartens in Gangnam and Seocho Districts is $1,882 (The World Without Concerns for Private Education, 2020).
[10] In fact, among the participants who indicated the current occupations of both parents in this study, there were 11 homemakers (36.7%), five professional workers (16.7%), and two entrepreneurs (6.7%) among mothers, and six office workers (20%), five professional workers (16.7%), and nine entrepreneurs (30%) among fathers.

Despite criticism that private English language early childhood education disrupts the starting line of compulsory education and creates a learning gap between children from wealthy and low-income families (Jeon, 2012), the demand for early English education continues to grow.

Although kindergarten education is not compulsory in Korea, its importance has been increasingly recognized in recent years (Kwon et al., 2017). Children usually enter kindergarten at the age of 3 and continue for three to four years before entering the first grade of elementary school. Since the Ministry of Education does not allow teaching English in kindergartens, English kindergartens are not classified as kindergartens, but as private institutions, known as 'hagwons.' English kindergartens typically run a full-day kindergarten program – an average of 5 hours a day – targeting children between the ages of 2 to 6 years. The class size is usually small, with 10-15 children in a class. Two teachers, a native English-speaking teacher and a Korean teacher fluent in English, guide students' day using only English. Private English kindergarten operators set their own curriculum. Some create their own and others use curriculum drawn from the US, Canada, UK, etc. For instance, one English kindergarten in Gangnam applies the Common Core State Standards of the US and another English kindergarten in Seocho applies the British National Curriculum of the UK. Most private English kindergartens use English textbooks and materials imported from the countries listed above.

There are three types of private English kindergartens in Korea: academic-oriented, play-based, and mixed. The academic-oriented program is teacher directed and managed. This means that teachers present students with new learning material and transmit information in an explicit, direct way. It offers a wide range of academic subjects[11] taught in English and students typically learn content beyond their age level. With an aim to prepare students for higher-level English education in the future through repeated practice in a structured setting, it has very high academic standards focusing on the development of literacy. For instance, students learn everything from basic grammar to more complicated writing techniques to improve their writing quality and memorize a set number of vocabulary words each week for weekly vocabulary tests. An academic-oriented English kindergarten often creates a competitive atmosphere with an emphasis on academic achievement.

[11] For example, the subjects taught in one English kindergarten in Gangnam District included reading, writing, speaking, storytelling, vocabulary, grammar, art, math, science, social studies, etc.

Meanwhile, the play-based program promotes play as central to children's growth and development based on the idea that children learn through play. This type of kindergarten aims to help students grow into independent human beings by developing the creativity and problem-solving skills necessary for entering elementary school. Classes are taught in English and include a variety of activities, such as creative play, dramatic play, and arts and crafts. The goal is to enable students to learn English in a fun and enjoyable environment by communicating with each other and expressing themselves creatively.

Lastly, the mixed-type program combines the two styles of teaching approaches mentioned above. They set goals for students' English language development while pursuing a play-based approach. Nevertheless, the curriculum of the mixed-type program is more closely related to that of an academic-oriented program and thus often classified as an academic-oriented English kindergarten. However, these classifications are not strictly pedagogically based, but rather terms used by English kindergartens to promote themselves. Therefore, although statistics are not reported, it would appear that there are far more English kindergartens that are academic-oriented than play-based.

### *Previous Studies on Early Childhood English Education*

The value of play in children's learning and development has been widely recognized by developmental psychologists (Piaget, 1962; Vygotsky, 1978). As young learners tend to have a lot of physical energy but short attention spans, many language researchers support the use of gross motor activities as effective language learning tools (Richards & Rodgers, 2001). According to Paterson (2020), more recent studies have demonstrated that the inclusion of play and playful activities within school settings supports and enhances various aspects of children's development and academic performance, such as self-regulation, language, creativity, cognition, and social skills.

Drawing inspiration from the American National Association for the Education of Young Children (NAEYC, 1986), Korean early childhood education emphasizes developmentally appropriate practices for young children where knowledge can be constructed through authentic child-centered experiences (Shim & Herwig, 1997). The current national kindergarten curriculum also emphasizes a child-centered, play oriented, integrated teaching approach focusing on children's holistic development (Korea Institute of Child Care and Education, 2019).

However, academic-oriented English kindergartens often focus on building students' academic skills through teacher-directed subject teaching to meet parents' demands driven by the pragmatics of preparation for formal schooling and children's later success (Sharpe, 2002). Moreover, the teacher-centered, knowledge and skills-based approach has been long embraced in the Korean context, which makes it difficult for teachers to adapt to the more child-centered emphasis of Western philosophies in early years education (Lim & Torr, 2008).

Studies on early English education show that activities conducted at the private English kindergartens are inappropriate according to the Early Childhood Education Act. In the study of Kwon (2002), Korean preschool educators criticized many aspects of academic-oriented programs as inappropriate for early years education, such as teacher-directed, highly structured, and paper-and-pencil tasks using workbooks and worksheets. Nevertheless, it shows that the private kindergartens are influenced by parental demands when planning lessons, because the institutions are dependent on the high tuition fees paid by parents. Therefore, despite the discrepancies between the beliefs of Korean teachers and the actual practices in academic-oriented English kindergartens (Kwon, 2002), parents tend to choose programs based on their perceptions of effective educational practices for improving their children's English language skills.

In recent years, there has been an increasing number of research on the early childhood English education in the field of early childhood education in Korea. The study of Yi and Yang (2009) reviewed previous research on the current situation and awareness of early English education in Korea. Prior studies on early English education at Korean English kindergartens investigated the influence of early English education on children's first language development, bilingual language development, and social development (Ma, 2003). Although there is still controversy about the effectiveness of early education, some studies have pointed out the issues of excessive early English education in Korea (Shin, 2002; Woo et al., 2002).

In contrast, studies by Seo et al. (2003) and Yoon (2008) have shown the positive effects of prior learning in English kindergartens on the acquisition of English in elementary schools. The study of Hwang (2004), which examined the influence of early English education on children's bilingual language development, has also indicated that early exposure to English improved children's English language development. However, there are few prior studies that investigated the long-term effects of early English education on the development of young learners in the

second language environment and how various contextual factors affect young learners' English learning, thus evidence is insufficient to draw conclusions about the influence of early English education on children in Korea.

Moreover, there is little, if any, research that looked into parental perceptions of need for early English education in the EFL environment and in-depth analysis of their experiences through the unique English programs of private English kindergarten in Korea. In Korea's education system, where English language education is not included in the national kindergarten curriculum, this study examines why parents want their children to start learning English early, what type of kindergarten education they prefer, and whether their expectations are met through the English education. In other words, the current study aims to analyze the education market that facilitates and supports the implementation of early childhood English education by investigating Korean parents' perceptions and experiences of early childhood English education based on their children's private English kindergarten education experiences.

The research objectives are as follows:

1. To investigate South Korean parents' perceptions of the need for early English education and their goals for their children's English language education
2. To investigate factors that influence parents' decisions in selecting English kindergartens, such as learning types and qualifications of English teachers.
3. To investigate parents' satisfaction levels with private English kindergarten programs and their children's English language outcomes after completing the programs

## METHOD

**Participants.** The participants of this study are 30 Korean parents who enrolled their children in private English kindergartens in Korea. Twenty-seven mothers and three fathers participated in the survey and eight of the mothers participated in follow-up interviews. The ages ranged mainly from 30s to 50s; 22 participants in their 30s, seven in their 40s, and one in her 50s.

Considering the sensitivity of the questions, some respondents skipped two background questions related to occupations and education. As a result, 19 participants (63.3%) responded to the item regarding occupation and 25 participants (83.3%) indicated the highest level of

education they have completed. The occupations of the participants are homemakers (10 participants), professional workers (five participants; one lawyer, one judge, two instructors, and one freelancer), and entrepreneurs (four participants). In terms of education, all 25 participants who responded to the item appeared to be highly educated, with a bachelor's degree or higher: 18 participants have obtained a bachelor's degree, and seven participants have earned a master's and/or a doctorate degree. Of the total participants, 13 participants (43.3%) had study abroad experience. The level of English proficiency the participants rated themselves for indicated that 10 participants (33.3%) were 'good,' 14 participants (46.7%) were 'average,' and six participants (20%) were 'poor.' As a result, 80% of the participants indicated their English proficiency as at least average.

Analysis of the type of English kindergarten that the participants enrolled their child in showed that 20 participants (66.7%) sent their child to an academic-oriented English kindergarten, five participants (16.7%) to a play-based English kindergarten, and the remaining five participants (16.7%) to a mixed-type English kindergarten. As for the locations of the English kindergarten the participants' children attended, 23 English kindergartens (76.7%) were in Seoul and seven (23.3%) were in different cities. In particular, 18 out of 23 English kindergartens in Seoul were located in the Gangnam District. Lastly, with regard to English kindergarten education tuition, 66.7% of the participants spent more than 1.5 million won ($1,300 US dollars) per month; five participants (16.7%) spent between 2 and 2.5 million won ($1,700 - $2,100 US dollars) per month, 15 participants (50%) spent between 1.5 and 2 million won ($1,300 - $1,700 US dollars), seven participants (23.3%) spent between 1 and 1.5 million won ($850 - $1,300 US dollars), and three participants (10%) spent 1 million won ($850 US dollars) or less.[12]

**Instruments.** This study used a combination of quantitative and qualitative data gathering methods including an online survey using Google Forms and semi-structured interviews via phone call. The survey was used to investigate Korean parents' perceptions of early English education in general (see Appendix A) and the interviews were carried out with selected participants to examine the in-depth thoughts and opinions based on their experiences (see Appendix B). The survey consisted of four background questions, 30 six-point Likert scale

[12] The researcher converted Korean Won (KRW) to US Dollar (USD) according to the dollar exchange rate in October 2021.

question items, and eight demographic questions. The Likert scale items asked participants to indicate their level of agreement, from strongly agree to strongly disagree, on six items about the need and goals for early English education, 15 items about reasons for choosing an English kindergarten and factors that influenced their decisions, and nine items about their satisfaction after their children completed the program. The survey was given in the native language of the participants, Korean. The interview was semi-structured with four open-ended questions about participants' perceptions of early English education and several questions based on responses to the survey. The researcher conducted each interview over the phone and interviews lasted approximately 45 minutes. For accuracy and convenience of the communication, interviews were conducted in Korean and audio-recorded with the consent of the participants.

**Data Collection and Analysis.** Prior to primary data collection, a pilot study was held in early June 2021 with two people to improve the quality of the outcome of the present study. No major problems were found in the questionnaire and several individual interview questions were developed after the pilot test. As for the main study, the researcher first contacted several parents with whom she had personal relationships and knew had sent their children to English kindergartens in Seoul and requested their participation in the online survey with a brief background information about the purpose of the survey. After the initial round of recruitment, the participants introduced the researcher to other parents. The data were collected from a total of 30 participants from mid-June to the end of June through this snowball sampling approach. After the survey, the researcher contacted eight of the participants who answered "yes" to the question "Are you willing to participate in a follow-up interview?" and carried out in-depth individual interviews with them. To ensure participants' ability to respond meaningfully, the subjects were parents of children who attended English kindergarten for more than one year and graduated from English kindergarten within the last three years. The interviews were conducted during July.

Data analysis in this study involved quantitative analysis to measure participants' perceptions, experiences, and satisfaction, and qualitative analysis to understand the background of their perceptions, experiences, and opinions. The quantitative data were collected through the questionnaire with 30 Likert-scale items conducted with 30 participants. Responses to these items were numerically coded; strongly disagree = 1, moderately disagree = 2, slightly disagree = 3, slightly agree = 4, moderately agree = 5, strongly agree = 6. The qualitative data were

collected through the responses to the four open-ended questions in in-depth interviews with eight participants drawn from the total of 30. The interviews were first transcribed and then translated into English by the researcher. These translated transcripts were organized into patterns for analysis as the researcher generated codes based on the topics and applied them to sections relevant to the research objectives.

## RESULTS

### *Parents' Perceptions of Early English Education*

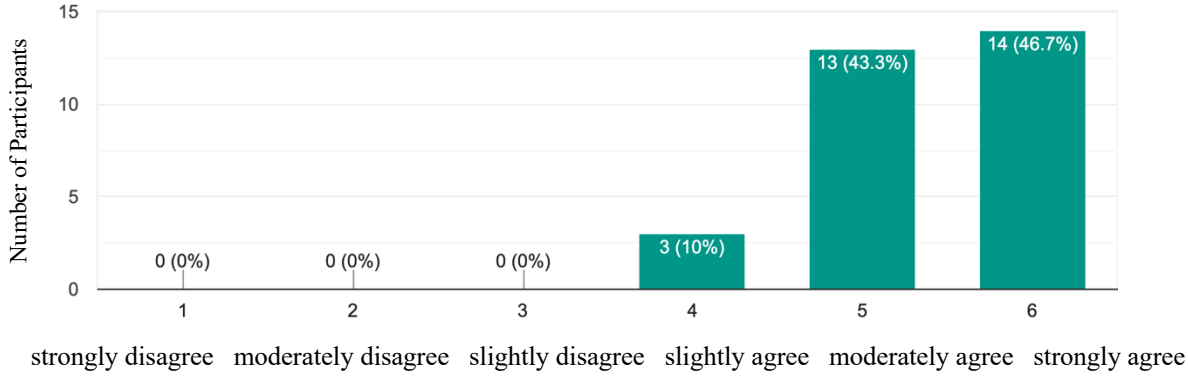
With an aim to investigate the perceptions of Korean parents about early English education, the first section of the questionnaire included items focusing on the need for early English education and their goals for their children's English language learning. The responses to the items revealed both general and specific beliefs about the importance of learning English from infancy.

**Need for Early English Education.** First, there was a general belief among participants that English skills are necessary in the era of globalization, in which all 30 participants selected the positive items on the 6-point Likert scale. As shown in Figure 1, 14 participants (46.7%) responded, 'strongly agree,' 13 participants (43.3%) said, 'moderately agree,' and the remaining three participants (10%) chose 'slightly agree.' Participant C, who has a 4-year-old son, described English language skills as "a necessary social communication tool whether traveling or doing business," and already expressed concern about the disadvantages her son would face if he does not speak English well in the future.

Participant's perceptions of the need for early English education are based on the popular belief that 'the younger the better' when it comes to learning a second language. The vast majority of participants (90%) believed that starting early yields linguistic advantages. According to the demographic data, 20 of the participants' children (66.7%) started attending English kindergartens at the age of 4 or earlier; two children (6.7%) at the age of 2, nine (30%) at the age of 3, and another nine (30%) at the age of 4. Of the remaining ten (33.3%) who started after the age of 4, four (13.3%) started at the age of 5 and six (20%) at the age of 6. The following quote is from a participant who enrolled her 2-year-old daughter in an English

**Figure 1**

*English Proficiency is Necessary in the Era of Globalization.*




kindergarten and responded 'strongly agree' to all three items related to the need for early English education.

> I think English is a critical component to living in a globalized world as it is the common language for travel, commerce, and technology. By learning English, my children will develop an excellent understanding of cultural diversity and be capable of communicating with people from all around the world. I also think that learning English is easier for young children because they soak up languages like sponges, so they should be able to learn it effortlessly. (Participant A)

Participants also commonly believed that young children learn English easily and accent-free. For example, Participant E said, "Children are less anxious and less inhibited than older language learners and thus learning English can be easy and fun for them" and Participant F said, "Children who receive early English education become confident in speaking English with native-like pronunciation and intonation."

More importantly, the high hopes of Korean parents for their children's academic well-being in the competitive education system and their wish to contribute to that success are noted as part of the reasons participants enroll their child in English kindergartens. Participant C, who described early English education as a "trend" among parents of preschool children, said, "everyone in Gangnam starts learning English early these days, so I could not help but join the trend for my child's academic success." On the other hand, Participant H criticized Korea's

current English education situation as "crazily overheated" and expressed concern that her son might fall behind his peers academically at school due to his lack of English skills.

**Goals for Early English Education.** Another common belief among participants was that learning English from an early age will help their children thrive academically and eventually lead to greater success in a variety of ways in their lives in the future. Participants overwhelmingly perceived early English education as essential in Korean society and viewed good English language skills as the key to academic success, university success, and career success. On each item asking about the relationship between early English education and future success, 29 participants (96.7%) answered that it would help them achieve academic excellence and get good grades in primary and secondary school, 23 participants (76.7%) responded that it would help them get a good score on the college entrance exam, and 25 participants (83.3%) marked that it would help them find a job they want in the future.

Follow-up interviews showed that these responses of the participants come from their past experiences in college and work. Participants shared their opinions along with personal stories related to English and how they felt that poor English skills could be "a hindrance" to their studies and careers. Participant D, who attended college in the mid-2000s, when many Korean universities had begun to establish and implement English-medium instruction (EMI) policies,[13] saw English skills as "critical linguistic tools" to survive in the college.

> When I was in college, half of my lectures were taught in English, and I assume that most courses will be in English by the time my child goes to college. I hope that my child will be able to speak English fluently and confidently so that he can keep up with English classes in the future. (Participant D)

Even after graduating from college, participants continued to spend a significant amount of time studying English to prepare for the Test of English in International Communication (TOEIC) and strived to improve their English proficiency to obtain competitive occupations such as white-collar professions.

[13] For example, Korea University was one of the first universities to offer EMI in 1999 (Kang, 2018), mandating that all departments offer at least one course in English (Jon et al., 2020); Korea Advanced Institute of Science and Technology (KAIST), one of Korea's most prestigious universities, started incorporating EMI in 2003 and began offering all undergraduate courses in English. This trend was spurred by the media-initiated university rankings, such as 'QS World University Rankings,' and the globalization policy of Korean universities (Cho, 2012).

> From my work experience, I learned that lack of English language skills often becomes a hindrance and hampers one from gaining a higher position in workplace. English language skills are absolutely necessary in career advancement, especially in this era of globalization. (Participant E)

### *Factors That Influence Parents' Decisions in Selecting Kindergartens*

Participants chose private English kindergarten education to provide opportunities for their children to learn English as naturally and unconsciously as their mother tongue and to set a path toward success in primary school and later in life. Participant F said, "while parents can facilitate all the activities that toddlers need to learn and do at home, learning English is a specialty that can only be done in an English kindergarten." The second section of the questionnaire examined specific factors that affect parents' decisions in choosing English kindergartens, such as the learning type and qualifications of English teachers.

**Learning Type: Play-based vs. Academic-oriented.** Since private English kindergartens in Korea are largely divided into two types,[14] the parents selected English kindergartens in consideration of what kind of learning they would like to provide to their children. To explore participants' opinions about the two types of English kindergarten, participants were asked to rate the effectiveness of the two learning styles in the second section of the questionnaire.

As shown in Figure 2, the highest response to the effectiveness of academic-oriented learning style was 'moderately agree' with 10 (33.3%) participants, followed by 'slightly agree' with nine (30%) participants. From a broader view, 20 participants (66.7%) responded positively, and 10 participants (33.3%) responded negatively, showing that most participants perceived academic-oriented learning as an effective way to learn English.

As to the effectiveness of play-based learning style in Figure 3, in contrast, the total number of participants who selected positive items was 16 (53.5%), slightly higher than the number of participants who selected negative items, 14 (46.7%). However, 'slightly disagree' was chosen by the highest number of participants with 11 (36.7%), and 'slightly agree' was the next with nine participants (30%), representing contradictory views of parents' perceptions and experiences.

[14] The curriculum of the mixed-type program is more closely related to that of an academic-oriented program and thus often classified as an academic-oriented English kindergarten.

**Figure 2**

*Academic-oriented Learning is the Most Effective Way to Learn English in Korea.*

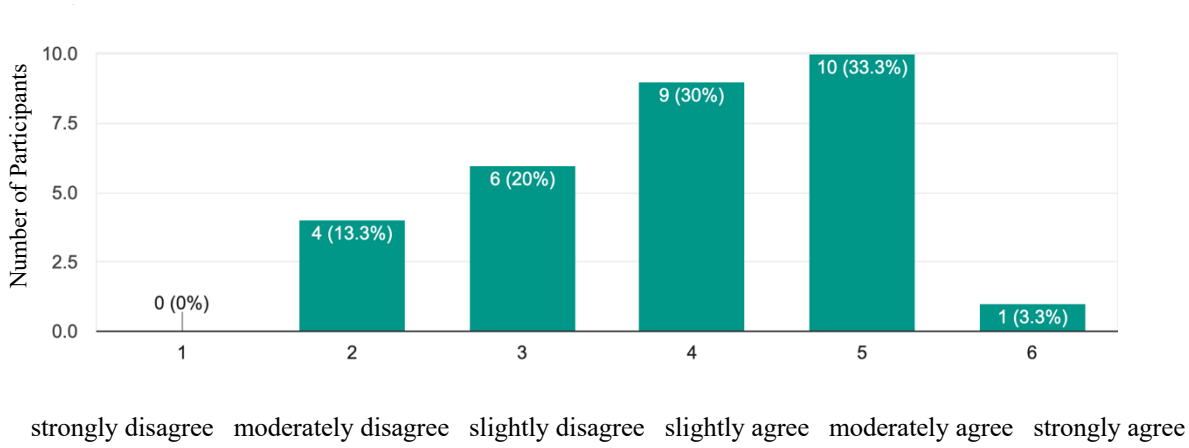
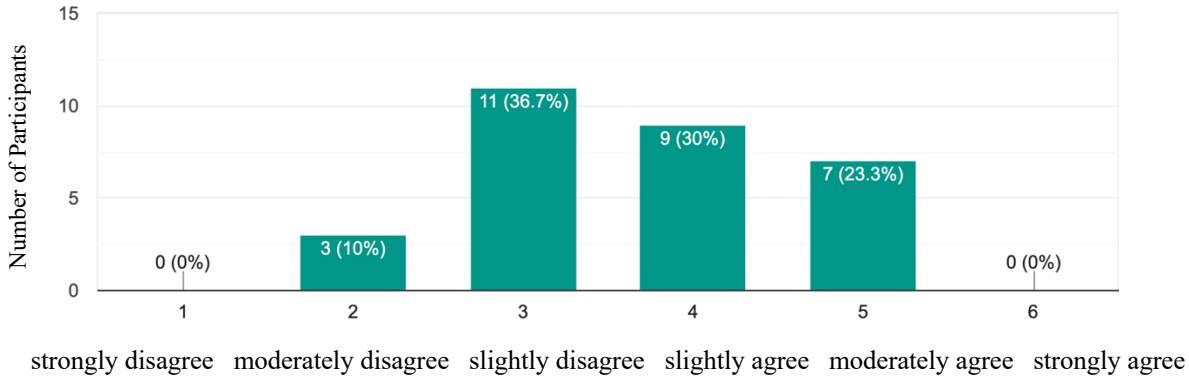



**Figure 3**

*Play-based Learning is the Most Effective Way to Learn English in Korea.*




Comparing the most common response of the two items, it was found that the participants had clear perception that academic-oriented learning is the more effective way to learn English in an EFL environment. Yet, they were not heavily negative about the play-based approach. Many participants (86.7%) believed that English language learning is effective when it is integrated into everyday life, such as physical activity, art, and play (Question #16). An ambivalence in participants' perceptions between the two learning styles were found in follow-up interviews. Participant F, who recently moved to Hawai'i with her 5-year-old twins, strongly believed that learning English naturally is the best way to learn English. However, she said that if she had

lived in Korea, she would have enrolled her children in an academic-oriented English kindergarten.

The interviews also revealed different goals between the participants who chose the academic-type English kindergarten and of those who chose the play-type English kindergarten. For example, Participant A, who sent her two daughters to an academic-oriented English kindergarten when they were two, expressed a strong preference toward academic-oriented learning.

> I had never considered a play-based English kindergarten for my children because I think that a child's first learning experience determines his or her attitude toward school for years to come. Once they get used to play, it would be difficult to develop study habits later when they enter elementary school. (Participant A)

In contrast, the following quote is from Participant G, who enrolled her twin daughters in a play-based English kindergarten to learn English through a "joyful experience."

> When I learned English, I had to sit down and memorize vocabulary lists and learn how to conjugate verbs. I certainly did not want my children to go through that. Rather, I wanted them to have a fun joyful experience learning English and that is why I sent them to a play-based English kindergarten. (Participant G)

However, despite what she had hoped for, she said she regretted her decision after the program ended. She added that if she could choose again, she would send her children to an academic-oriented English kindergarten for the "better English language learning outcomes."

Participants C, D, and E, who enrolled their children in academic-oriented English kindergartens, shared similar opinions that despite paying the same tuition and spending the same amount of time in both types of kindergarten, students who learn English at the academic-oriented English kindergarten demonstrate better English language proficiency. These participants also had clear expectations and goals for their children to be able to "read, write, understand, and speak English fluently by the end of the program so that they can pass the exam to enter one of the best English language academies in Gangnam" (Participants C, D, & E).

**Professionalism: Program Quality and Teacher Expertise.** Parents' perspectives on criteria of a good quality of English kindergarten were shown as closely related to program quality and professionalism of teachers. The responses to the questionnaire showed that 27 participants (90%) indicated that the program quality is an important factor and 28 participants

(93.5%) expressed that the teacher expertise is an important factor that would influence their choice of English kindergarten. However, the interviews revealed an ambivalence between parents' perceptions of professionalisms and factors that actually affect their decision.

Participants defined a high-quality program of an English kindergarten as: whether it "utilizes imported English textbooks" (Participant D), whether it "provides activities necessary for children's growth and emotional and personality development" (Participant B), whether it "promotes appropriate learning to facilitate further learning" (Participant G), etc. However, it appears that there is no objective basis for measuring the overall quality of the English kindergartens and the programs. Rather, it was shown that participants gained evaluations of the kindergartens from other parents or at online forums called "Mom Cafes," which are online communities where mothers share parenting tips and updates with each other. In fact, half of the participants (50%) selected 'moderately agree' to the item regarding the importance of the reputation, nine participants (30%) chose 'slightly agree,' and four participants (13.3%) picked 'strongly agree.' The reason why the participants depended heavily on reputation was that there was no way to accurately identify the quality of the English kindergarten education program.

Another discrepancy related to factors affecting participants' choice of English kindergarten was found with respect to their perceptions of professionalism of English teachers. Participants defined highly qualified teachers with such objective qualifications as having "a good educational background" (Participant D), "a love for children" (Participant C), and "passion for teaching" (Participant E). They also showed a strong preference for teachers with a relevant degree to early childhood education or TESOL certification, who can understand the challenges young learners may experience while learning a second language. However, there was a more powerful belief that dominated participants' perceptions of a competent language teacher: 'native-speakerism,' the belief that native speakers of English make better English language teachers.

Despite the various factors presented to define the professionalism of teachers, 23 participants (76.6%) believed that learning English from a native English-speaking teacher was more effective than from a Korean English teacher and preferred a native speaker as their child's kindergarten teacher. Below is a quote from one of the participants who preferred foreign teachers to Korean English teachers.

> In my opinion, the presence of both the native teacher and Korean teacher is important especially in an EFL classroom environment. Native English teachers can teach students cultural aspects, accurate English expressions, good pronunciation and intonation, and Korean teachers can support children's emotional needs and explain cultural differences in more detail that foreign teachers cannot. Still, I think the reasonable ratio of the native English-speaking teacher to the Korean teacher for teaching English lessons is 8:2. (Participant C)

Though this participant expressed the need for both a native English teacher and a Korean English teacher in a kindergarten English classroom, she emphasized the role of a native English-speaking teacher as the primary teacher and the Korean teacher as an assistant teacher. In addition, Korean English teachers were recognized as serving as a bridge between children and foreign teachers, regardless of their professionalism and English proficiency.

### *Parents' Satisfaction After Completing the Programs*

The last section of the questionnaire explored whether participants' expectations for the English kindergarten program and their children's English learning outcomes were met after completing the programs. The responses showed that 28 participants (93.3%) were satisfied overall with the curriculum of the English kindergartens their children attended. Regarding their children's English learning outcomes, eight participants (26.7%) were highly satisfied, nine participants (30%) were moderately satisfied, eight participants (26.7%) were slightly satisfied, four participants (13.3%) were slightly unsatisfied, and one participant (3.3%) was moderately unsatisfied. Overall, participants showed high levels of satisfaction with their children's four English language skills: reading, writing, listening, and speaking, and were especially satisfied with the improvement of their English communication skills and confidence. In particular, participants often valued the ability to speak in English confidently and mentioned it during the interviews. For instance, Participant D, who was highly satisfied with her child's English learning outcome, said she was proud to see her son speaking English to her after returning home from English kindergarten and not feeling shy about making mistakes while speaking English. In contrast, Participant G said when she saw her child running away from a native speaker, she felt that her child lacked confidence in her English skills and thought it would have been better if she had sent her to an academic-oriented English kindergarten.

More importantly, the results revealed that parents who enrolled their children in an academic-oriented English kindergarten had greater satisfaction with their children's English skills at the end of the program than parents who enrolled their children in the play-based English kindergarten. Of the 20 participants in the academic-oriented group, 17 participants selected positive responses, showing that 85% of parents were satisfied with their child's English language learning outcomes. Moreover, 13 out of the 20 participants (65%) selected 'moderately agree' or 'strongly agree,' indicating a particularly high level of satisfaction. Of the five participants in the mixed group, four participants (80%) said 'moderately agree' or 'strongly agree,' and one (20%) participant said 'slightly agree.' In contrast, none of the participants in the play-based group chose either 'moderately agree' or 'strongly agree'; three participants (60%) responded, 'slightly agree,' and the remaining two reported dissatisfactions.

## DISCUSSION

This study revealed Korean parents' beliefs that early English education is beneficial for children's future academic and career success. They also perceived that academic-oriented learning is more effective in children's English language learning than play-based learning. Most of the parents who chose the academic-oriented English kindergartens for their children's English education were highly satisfied with the English language learning outcomes. In line with Korean parents' strong demands for early English education, studies have shown the benefits of learning a second language at an early age. Unfortunately, however, in Korea, English education for young children is excluded from the national curriculum in order to normalize public education and to prohibit teaching ahead of the regular public curriculum. Since the current education system fails to meet the demands of parents, more parents want to send their children to private English kindergartens.

The current trend of private English kindergartens and parents' high preference for academic-oriented programs reflect Korea's cultural values in education and competitive college-entrance-oriented education system. Although the results showed that parents who chose academic-oriented English kindergartens were highly satisfied with their children's English language learning outcomes, it is difficult to evaluate the effectiveness of education based on parental satisfaction alone. The overall and long-term effects of intensive foreign language education

have not been well studied, and more importantly, it is necessary to reconsider whether it is appropriate to provide English education to young children and to examine whether they enjoy learning English and whether early English education is linked to long-term learning outcomes. Thus, more studies on both sides are needed to determine the effectiveness of academic-oriented learning and play-based learning in various aspects.

The findings imply that despite parental demands for early English education, the national kindergarten curriculum does not support English language education, and thus parents choose to enroll their children in private English kindergartens. English kindergarten education is attractive to many Korean parents as it aligns with their beliefs about what supports language acquisition and concerns about academic and professional success in Korean society. Especially, English proficiency is directly related to students' academic success and college admissions in Korea, and thus it is understandable that parents have high demands and aspirations for early English education. However, it is necessary to reflect on whether English kindergarten education is academically appropriate for young children. Because private English kindergartens are owned and operated by individuals or companies, they tend to prioritize profit over the needs of parents (academic excellence, good grades, etc.).

In addition, most private English kindergartens are concentrated in the affluent area of Gangnam District. They tend to be more expensive than general kindergartens, making them inaccessible to children from families with low socioeconomic status. If only children of economically and culturally – with study abroad experience – wealthy parents continue to receive an early English education, the educational gap will widen. In order to prevent educational disparities, it is necessary to provide equal opportunities for all children to learn English in an era of globalization. Based on this implication, this study suggests that English language education needs to be integrated into public education, either by including English kindergartens as a type of general kindergarten and reducing tuition or by incorporating English education into the national kindergarten curriculum.

### *Limitations and Future Directions*

The findings of this study must be seen in light of some limitations. First, this study involved a small sample size (n=30). Moreover, there were too few participants in the play-based and the mixed-type groups. As different sample sizes may limit the generalizability of comparisons

between groups, data from a greater number of participants is suggested for future research. Second, although participants in the academic-oriented group indicated higher levels of satisfaction regarding their children's English language learning outcomes, more research is needed to examine the long-term effects of early English education on young learners in an EFL environment (i.e., whether parents' perceptions match actual gains in language development). Although some studies have investigated long-term benefits of an early start in instructed learners (Larson-Hall, 2008; Muñoz, 2011), the number is very small and show mixed findings. Thus, longitudinal studies that include a measurement of proficiency may offer a more complete picture about the effectiveness of early English education on children's later English and academic and career success. Third, there is a lack of prior studies on early English education in the EFL context. Many second language acquisition studies to date have focused on the acquisition of children or adults learning English in a bilingual setting (Montrul, 2005), but there is little research on young learners' English language acquisition an EFL context. The field could benefit from more research examining the dynamic process of children's second language learning and analyzing the effectiveness and weakness of language teaching approaches and methods used for young second language learners of English. This type of study will provide evidence in support of parents' perceptions of early English education and the effectiveness of academic-oriented learning and play-based learning. Additionally, it is suggested that future studies probe English language learning focusing on children's perspectives by exploring their experiences of learning English, such as whether children enjoy the learning process and how various contextual factors affect their English learning in an EFL environment.

## REFERENCES

Cho, D. (2012). English-medium instruction in the university context of Korea: Tradeoff between teaching outcomes and media-initiated university ranking. *The Journal of AsiaTEFL*, 9(4), 135-163.

Choi, J. (2020, August 10). OECD forecasts S. Korea's GDP ranking to take 9th. *The Korea Herald*. http://www.koreaherald.com/view.php?ud=20200810000670#:~:text=OECD%20forecasts%20S.,GDP%20ranking%20to%20take%209th

Hwang, H. (2004). Chogiyŏngŏgyoyuk i yua ŭi ijung ŏnŏbaltal e mich'inŭn yŏngyang [The influence of early English education on children's bilingual language development]. *Han'guksaengwalgwahakhoeji*, 13(4), 497-506.

Jeon, M. (2012). English immersion and educational inequality in South Korea. *Journal of Multilingual and Multicultural Development*, 33(4), 395-408. https://doi.org/10.1080/01434632.2012.661438

Jon, J., Cho, Y., & Byun, K. (2020). Internationalization by English-medium instruction? Professors' decoupling behaviors to EMI policy in Korean higher education. *Korean Educational Development Institute Journal of Educational Policy*, 17(2), 297-318

Jung, M. (2019, March 15). Korea lifts English education ban for first, second graders. *The Korea Times*. https://www.koreatimes.co.kr/www/nation/2019/03/181_265365.html

Kang, K. (2018). English-medium instruction policies in South Korean higher education. *Working Papers in Educational Linguistics*, 33(1), 31-52. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1304&context=wpel

Koo, H., & Lee, Y. (2006). 'Wild geese fathers' and a globalised family strategy for education in Korea. *International Development Planning Review*, 28(4), 533-553.

Korea Institute of Child Care and Education. (2019). *Understanding the 2019 revised Nuri curriculum*. https://kicce.re.kr/board/boardFile/download/89/32581/10887.do

Korean Educational Statistics Services. (2021). *Educational statistics*. https://kess.kedi.re.kr/index

Korean Ministry of Education. *Education system*. http://english.moe.go.kr/sub/info.do?m=020101&s=english

Kwon, S., Lee, M., & Shin, D. (2017). Educational assessment in the Republic of Korea: Lights and shadows of high-stake exam-based education system. *Assessment in Education: Principles, Policy & Practice*, 24(1), 60-77. https://doi.org/10.1080/0969594X.2015.1074540

Kwon, Y. (2002). Western influences in Korean preschool education. *International Education Journal*, 3(3), 153-164.

Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. *Second Language Research*, 24(1), 35-63. https://doi.org/10.1177%2F0267658307082981

Lee, M., Kim, H., & Han, M. (2020). Language ideologies of Korean mothers with preschool-aged children: Comparison, money, and early childhood English education. *Journal of Multilingual and Multicultural Development*, 42(7), 637-649. https://doi.org/10.1080/01434632.2020.1713137

Lim, C., & Torr, J. (2008). Teaching literacy in English language in Singaporean preschools: Exploring teachers' beliefs about what works best. *Contemporary Issues in Early Childhood*, 9(2), 95-106. https://doi.org/10.2304%2Fciec.2008.9.2.95

Ma, S. (2003). Yua ŭi sangwiŏnŏnŭngnyŏk kwa mogugŏ mit oegugŏ ŏhwiryŏkkan ŭi kwan'gye [The relationship of preschoolers' metalinguistic awareness to their first language and foreign language vocabulary]. *Yuagyoyukyŏn'gu*, 23(2), 267-289.

Montrul, S. (2005). Second language acquisition and first language loss in adult early bilinguals: Exploring some differences and similarities. *Second Language Research*, 21(3), 199-249. https://doi.org/10.1191%2F0267658305sr247oa

Muñoz, C. (2011). Is input more significant than starting age in foreign language acquisition? *International Review of Applied Linguistics*, 49(1), 13-33.

Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics*, 35(4), 463-482. https://doi.org/10.1093/applin/amu024

National Association for the Education of Young Children. (1986). NAEYC position statements on developmentally appropriate practice in early childhood programs serving children from birth through age 8. *Young Children*, 41(6), 3-19.

Organisation for Economic Co-operation and Development. (2018). *Towards Better Social and Employment Security in Korea*, Connecting People with Jobs, OECD Publishing, Paris, https://doi.org/10.1787/9789264288256-en

Park, J. (2009). 'English fever' in South Korea: its history and symptoms. *English Today*, 25(1), 50-57. https://doi.org/10.1017/S026607840900008X

Paterson, A. (2020). The play paradox: a systematic literature review of play-based pedagogy applied in the classroom. *Educational & Child Psychology*, 37(4), 96-114.

Piaget, J. (1962). The stages of the intellectual development of the child. *Bulletin of the Menninger Clinic*, 26(3), 120-128.

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and methods in language teaching*. USA: Cambridge University Press.

Seo, H., Choi, M., Jwa, S., & Chun, H. (2003). Ch'wihakchŏnhu yŏngŏgyoyuk kwa ch'odŭngakkyo 3hangnyŏn ŭi yŏngŏyuch'angsŏng ŭi kwan'gyeyŏn'gu [A study on the relationship between pre- and post-school English education and English fluency in the third grade of elementary school]. *Mirae Yuagyoyukhakhoeji*, 10(4), 299-319.

Seth, M. J. (2002). *Education fever: society, politics, and the pursuit of schooling in South Korea*. University of Hawaii Press.

Sharpe, P. (2002). School days in Singapore: Young children's experiences and opportunities during a typical school day. *Childhood Education*, 79(1), 9-14. https://doi.org/10.1080/00094056.2002.10522757

Shim, S., & Herwig, J. E. (1997). Korean teachers' beliefs and teaching practices in Korean early childhood education. *Early Child Development and Care*, 132(1), 45-55. https://doi.org/10.1080/0300443970132010004

Shin, E. (2002). Chogi kyoyuk paltal kwa paltalbyŏngnijŏk munje: Han'guk chogi kyoyuk ŭi hyŏnhwang kwa kwaje [Early education and developmental pathological problems: current status and tasks of early education in Korea]. *Han'guk adonghakhoe ch'ugye haksultaehoe*, 29-42.

Statistics Korea. (2021). *Total expenditure on private education in South Korea in 2019, by subject*. https://www.statista.com/statistics/1042882/south-korea-total-spending-for-private-education-by-subject/

The World Without Concerns for Private Education. (2020, October 20). 140 million won for 8-year tuition, the serious polarized education from the st art. *OhmyNews*. https://news.v.daum.net/v/20201020160601019?x_trkm=t

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Woo, N., Seo, Y., & Kang, Y. (2002). Yŏngyu e taehan chogiyŏngŏgyoyuk ŭi chŏkchŏlsŏng e kwanhan yŏn'gu [A study on the adequacy of early English education for young children]. *Kyoyuginjŏkchawŏnpu chŏngch'aegyŏn'gu kwaje*, 2002-16.

Yi, Y., & Yang, S. (2009). Chanyŏ rŭl yŏngŏyuch'iwŏn e ponaenŭn ŏmŏnidŭl ŭi kyŏngŏm e taehan yŏn'gu [A qualitative study on the experience of mothers sending their children to English kindergarten]. *Han'guksaengwalgwahakhoeji*, 18(5), 985-994.

Yoon, Y. (2008). Yŏngŏmorip p'ŭrogŭraem, yŏngŏnŭngnyŏk kwa munhwajŏngch'esŏng e ŏttŏn yŏngyang ŭl chulkka? [What effect does English immersion program have on English proficiency and cultural identity?]. Han'gukhaksulchŏngbo.

**APPENDIX A**

*Survey Questionnaire*

**Background Questions**

**If you have more than one child who attended a private English kindergarten, please fill out this form based on the experience of the younger or youngest child.*

**If your child attended more than one English kindergarten, please fill out this form based on the kindergarten attended the longer or longest.*

1. Please indicate the type of English kindergarten your child attended.

   ☐ Academic-oriented ☐ Play-based ☐ Mixed-type

2. Please indicate the location of the English kindergarten your child attended.
3. Please indicate the age at which your child first entered English kindergarten.
4. Please indicate the length of time your child attended English kindergarten.

   (**If your child attended more than one English kindergarten, please indicate each period.)*

**Research Objective 1: Perceptions**

5. English proficiency is necessary in the era of globalization.

   ☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

6. Early English education is very important in English education.

   ☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

7. Early English education will be helpful for academic excellence and getting good grades in primary and secondary school.

   ☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

8. Early English education will be helpful for getting a good score on the college entrance exam in the future.

   ☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

9. My child's English skills will be helpful for finding the job s/he wants in the future.

   ☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

10. Learning English is easier when a child starts it at an early age.

    ☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

**Research Objective 2: Factors**

11. I sent my child to English kindergarten because it provides professional English education.

    ☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

12. I sent my child to English kindergarten because it provides an excellent environment and care system for young children.

    ☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

13. Academic-oriented learning is the most effective way to learn English in Korea.

    ☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

14. In English language classrooms for young learners, it is important to focus on the four skills: reading, writing, listening, and speaking.

    ☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

15. Play-based learning is the most effective way to learn English in Korea.

    ☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

16. English language learning is effective when it is integrated into everyday life, such as physical activity, art, and play.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

17. It is more effective to learn English from native English-speaking teachers than Korean teachers fluent in English.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

18. The professional expertise of the English education program is an important consideration when choosing my child's English kindergarten.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

19. The professionalism of the teachers is an important consideration when choosing my child's English kindergarten.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

20. The number of classes taught by native speakers of English is an important consideration when choosing my child's English kindergarten.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

21. The reputation of the kindergarten is an important consideration when choosing my child's English kindergarten.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

22. Small class size is an important consideration when choosing my child's English kindergarten.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

23. Tuition is an important consideration when choosing my child's English kindergarten.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

24. Inclusion of after-school classes is an important consideration when choosing my child's English kindergarten.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

25. Distance from home is an important consideration when choosing my child's English kindergarten.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

**Research Objective 3: Satisfaction**

26. I am satisfied overall with my choice of English kindergarten for my child.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

27. I am satisfied with the curriculum of the English kindergarten that my child attended.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

28. I am satisfied with my child's English language learning outcomes.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

29. My expectations for my child's English reading skills development were met.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

30. My expectations for my child's English writing skills development were met.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

31. My expectations for my child's English listening skills development were met.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

32. My expectations for my child's English speaking skills development were met.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

33. My expectations for my child's basic English communication skills were met.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

34. My child has become more confident in English.

☐ Strongly disagree ☐ Moderately disagree ☐ Slightly disagree ☐ Slightly agree ☐ Moderately agree ☐ Strongly agree

**Demographic Questions**

35. What is the gender of your child?

☐ Female ☐ Male

36. How old is your child?
37. How old is the mother / father?
38. What is the mother's / father's occupation?
39. What is the highest level of education that the mother / father has completed?

☐ High school graduate ☐ Bachelor's degree ☐ Master's degree and/or doctorate degree

40. Does the mother / father have study abroad experience?

☐ Yes ☐ No

41. What is the mother's / father's English proficiency level?

☐ Good ☐ Average ☐ Poor

42. How much was the monthly tuition fee for the English kindergarten that your child attended?

☐ 1 million won or less ☐ 1 – 1.5 million won ☐ 1.5 – 2 million won ☐ 2 – 2.5 million won ☐ 2.5- 3 million won

43. Are you willing to participate in a follow-up interview?

☐ Yes ☐ No

44. If you are, please leave your phone number or email address where you can be reached.

Survey Questionnaire (Korean-translated Version)

**Background Questions**

*만약 영어유치원에 다닌 자녀가 두 명 이상인 경우, 가장 어린 자녀의 경험을 기준으로 응답해 주십시오.

*만약 자녀가 2개 이상의 영어유치원에 다닌 경우, 가장 오래 다닌 영어유치원의 경험을 중심으로 응답해 주십시오.

1. 자녀의 영어유치원은 어떤 종류입니까?

   ☐ 학습식 ☐ 놀이식 ☐ 혼합식

2. 자녀의 영어유치원이 소재한 지역은 어디입니까?
3. 자녀는 몇 살에 영어유치원에 처음 등록 했습니까?
4. 자녀가 영어유치원에 다닌 총 기간을 기입해 주십시오.

   (*만약 자녀가 2개 이상의 영어유치원에 다닌 경우, 각각의 기간을 기입해 주시기 바랍니다.)

**Research Objective 1: Perceptions**

5. 영어 능력은 세계화 시대를 살아가는데 필수적이다.

   ☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

6. 영어교육에서 조기교육은 매우 중요하다.

   ☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

7. 유아 영어학습은 향후 학업과 성적에 도움이 될 것이다.

   ☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

8. 유아 영어학습은 향후 대학 입학시험에서 좋은 성적을 거두는데 도움이 될 것이다.

   ☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

9. 자녀의 영어 실력은 미래에 원하는 직업을 갖는데 도움이 될 것이다.

   ☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

10. 영어교육은 어린 나이에 시작할수록 효과적이다.

    ☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

**Research Objective 2: Factors**

11. 내가 자녀를 영어유치원에 보낸 이유는 전문적 영어교육에 최우선을 두었기 때문이다.

    ☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

12. 내가 자녀를 영어유치원에 보낸 이유는 영어유치원의 교육 환경과 돌봄 시스템이 우수하기 때문이다.

    ☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

13. 한국의 유아 영어교육 환경에서는 학습식 영어교육이 가장 효과적이다.

    ☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

14. 유아 영어학습에서 읽기, 쓰기, 듣기, 말하기 등 영어 능력 성취에 초점을 맞추는 것이 중요하다.

    ☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

15. 한국의 유아 영어교육 환경에서는 놀이식 영어교육이 가장 효과적이다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

16. 유아 영어학습은 신체활동이나, 미술, 놀이 등 일상생활 속에서 배우는 것이 효과적이다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

17. 한국에서 유아 영어는 이중언어 교사(한국인 영어교사)보다 원어민(영어권 출신) 교사에게 배우는 것이 더 효과적이다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

18. 나는 자녀의 영어유치원 선택 시 영어교육 프로그램의 전문성이 중요한 고려 사항이다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

19. 나는 자녀의 영어유치원 선택 시 교사들의 전문성이 중요한 고려 사항이다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

20. 나는 자녀의 영어유치원 선택 시 원어민 교사의 수업 비중이 중요한 고려 사항이다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

21. 나는 자녀의 영어유치원 선택 시 유치원에 대한 주위 평판이 중요한 고려 사항이다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

22. 나는 자녀의 영어유치원 선택 시 소규모 학급 운영이 중요한 고려 사항이다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

23. 나는 자녀의 영어유치원 선택 시 수업료가 중요한 고려 사항이다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

24. 나는 자녀의 영어유치원 선택 시 방과후 수업 유무가 중요한 고려 사항이다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

25. 나는 자녀의 영어유치원 선택 시 등/하원에 걸리는 시간이 중요한 고려 사항이다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

**Research Objective 3: Satisfaction**

26. 나는 자녀를 위한 영어유치원 선택에 전반적으로 만족한다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

27. 나는 자녀의 영어유치원 교육과정(커리큘럼)에 만족한다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

28. 나는 자녀의 영어학습 결과에 만족한다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

29. 자녀의 영어 읽기 (reading) 능력 발달에 대한 나의 기대가 충족되었다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

30. 자녀의 영어 쓰기 (writing) 능력 발달에 대한 나의 기대가 충족되었다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

31. 자녀의 영어 듣기 (listening) 능력 발달에 대한 나의 기대가 충족되었다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

32. 자녀의 영어 말하기 (speaking) 능력 발달에 대한 나의 기대가 충족되었다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

33. 자녀의 기본적인 영어 의사 소통 (communication) 능력에 대한 나의 기대가 충족되었다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

34. 내 자녀는 영어유치원에 다닌 후 영어에 대한 자신감을 갖게 되었다.

☐ 전혀 그렇지 않다 ☐ 그렇지 않다 ☐ 별로 그렇지 않다 ☐ 어느정도 그렇다 ☐ 그렇다 ☐ 매우 그렇다

**Demographic Questions**

35. 자녀의 성별은 무엇입니까?

☐ 여자 ☐ 남자

36. 자녀의 현재 나이는 만 몇 세입니까?

37. 어머니 / 아버지의 현재 나이는 만 몇 세입니까?

38. 어머니 / 아버지의 현재 직업은 무엇입니까?

39. 어머니 / 아버지의 최종 학력은 무엇입니까?

☐ 고등학교 졸업 ☐ 대학교 졸업 ☐ 대학원 졸업

40. 어머니 / 아버지는 유학 경험이 있으십니까?

☐ 예 ☐ 아니오

41. 어머니 / 아버지의 영어 능력은 다음 중 어디에 해당된다고 생각하십니까?

☐ 상 ☐ 중 ☐ 하

42. 자녀의 영어유치원 월 원비는 얼마입니까?

☐ 100만원 미만 ☐ 100–150만원 미만 ☐ 150-200만원 미만 ☐ 200-250만원 미만 ☐ 250-300만원 미만

43. 귀하는 만약 연구자가 추가 질문을 위해 인터뷰를 요청하는 경우 응해 주시겠습니까?

☐ 예 ☐ 아니오

44. 연락 가능한 전화번호 혹은 이메일 주소

**APPENDIX B**

*Semi-structured Interview*

**Open-ended Questions**

1. Why did you choose to enroll your child in a private English kindergarten and not general Korean kindergarten?
2. Why did you choose to enroll your child in an academic-oriented program / play-based program?
3. Why do you think that English proficiency is necessary in the era of globalization?
4. Why do you think that it is important to start learning English at an early age?
5. How would you define the professional expertise of the English education program?
6. How would you define the professionalism of the teachers?
7. Why do you think that it is more / less effective to learn English from native English-speaking teachers than Korean teachers fluent in English?
8. What skills did you expect your child to have mastered by the end of the kindergarten year?
9. Do you feel that expectations were met? If not, please explain in what way your expectations were not met.
10. If you could choose again, would you like to enroll your child in the same type of English kindergarten? Why or why not?

Semi-structured Interview (Korean-translated Version)

**Open-ended Questions**

1. 귀하가 일반유치원이 아닌 영어유치원을 선택한 이유는 무엇인가요?
2. 귀하가 자녀를 학습식 / 놀이식 영어유치원에 보낸 이유는 무엇인가요?
3. 세계화 시대에서 영어 능력이 필수라고 하셨는데 개인적으로 그런 생각을 하시게 된 배경이나 계기가 있으신가요?
4. 영어 조기교육의 중요성에 대한 생각을 갖게 된 계기는 무엇인가요?

5. 영어유치원 프로그램의 전문성이라고 할 때, 귀하는 구체적으로 어떤 요소를 생각하시나요?
6. 영어유치원 교사들의 전문성을 고려할 때, 귀하는 구체적으로 어떤 요소를 생각하시나요? (예, 관련 전공, 자격증 취득 등)
7. 유아 영어교육에서 원어민 교사 / 한국인교사에게 배우는 것이 더 효과적인 이유는 무엇인가요?
8. 영어유치원 교육을 통해서 어느 정도의 영어능력 성취를 기대하셨나요?
9. 선택 및 학습 결과에 대한 만족도가 기대치에 충족했나요? 만족스럽지 못한 이유는 무엇인가요?
10. 다시 선택을 한다면 같은 유형의 영어유치원에 자녀를 등록시키고 싶으신가요? 그 이유는 무엇인가요?

# ANNOTATED SUMMARIES OF SECOND LANGUAGE STUDIES GRADUATE WORK AT THE UNIVERSITY OF HAWAIʻI AT MĀNOA, 2021-2022

**KRISTEN URADA**
**RICKEY LARKIN, JR**
***University of Hawaiʻi at Mānoa***

As progress is made in returning to normal after the COVID-19 pandemic in 2020, the pandemic's effects on research can still be observed in this issue's Masters-level and Advanced Graduate Certificate scholarly papers and Doctoral dissertations that were completed during the 2021 - 2022 academic school year. These scholarly papers and dissertations show the continued resilience of the graduate students in the Department of Second Language Studies at the University of Hawaiʻi at Mānoa as several projects took place during the pandemic. Furthermore, the researchers of these projects were resourceful in their efforts to collect data during the pandemic as many of these studies included participants from across the globe and have various language and social backgrounds. To read more about each project, all of the scholarly papers and dissertations that have been summarized below are available on ScholarSpace at University of Hawaiʻi at Mānoa (https://scholarspace.manoa.hawaii.edu).

| Graduation Term | Student Name | Degree | Title of Scholarly Paper (AGC, MA) or Dissertation (PhD) |
|---|---|---|---|
| Summer 2022 | Christensen, Cade | MA | *The Effects of a Classroom Environment on Mutual Visibility, Transparency, and Sharing on ESL Students' Writing*<br>This study explored the effects of mutual visibility, transparency, and sharing that resulted from student use of Google Docs during an L2 writing class. Data from six students were |

| | | | |
|---|---|---|---|
| | | | gathered in order to examine these effects, which included after-essay questionnaires, focus group interviews, and observational notes. Data were then analyzed via descriptive statistics and descriptive coding. Results showed universal benefits from the effects of Google Docs as a classroom writing tool. |
| Summer 2022 | Ishiyama, Hikaru | MA | *EFL College Students' Perception of English Writing Activities in High School and College*<br>This quantitative study analyzed Japanese English as a foreign language undergraduate students' motivation, self-efficacy, and apprehension toward writing in English. The analysis used a 5-point Likert scale targeting the three aforementioned constructs. Descriptive and inferential statistics were used during the analysis. Results indicated a continuation of apprehension and low self-efficacy from high school to college. These findings may inform changes to curriculum development. |
| Summer 2022 | Jennings, Annika | MA | *Adult Language Learners' Attitudes Towards Native VS. Non-Native Speakers as the Ideal Foreign Language Teacher* |
| Summer 2022 | Motomura, Kaoru | MA | *Development of Japanese Students' Critical Consciousness in a Multilingual/Multicultural Society* |
| Summer 2022 | Park, Leesa | MA | *An Examination of Critical Pedagogy in Hyukshin School EFL Classes in Korea* |
| Summer 2022 | Schweingruber, Melanie | MA | *Talking Sustainability: Shaping Environmental Narratives on Reddit*<br>As environmental concerns increasingly become a concern in the daily lives of people, online communities have formed to discuss alternative methods of living. This study examined one such community on Reddit, the /r/Sustainability community. Adopting a critical ecolinguistics frame, this observational case study explored how a single |

| | | | |
|---|---|---|---|
| | | | participant negotiates their shared online space within the community. A corpus was built using the participant's online posts and comments throughout 2019-2022. These posts were then (re)analyzed using descriptive coding and applying the concept of framing, narrative analysis, and affordances. Results described how narratives formed online over time and how they shifted due to environmental factors such as COVID-19. |
| Spring 2022 | Argueta, Jenny | MA | *Korean Dramas as a Site for Intercultural Representations* |
| Spring 2022 | Banov, Ivan | PhD | *Agentic Development in Offline Affinity Spaces: Reddit as a Place for Second Language Learning*<br>This ecolinguistic-based dissertation examined the use of Reddit as a second language (L2) learning space. Reddit, with its function as a popular social media site with its own sociocultural histories, is shown to afford and constrain the types of communication that occurs in the space. 21 L2 English speakers' communication and social media actions were examined in order to explore how these long-term interactions develop into a complex system that struggles to be analyzed using traditional constructs such as complexity or accuracy. Results of this study inform ecolinguistic research by addressing new concepts such as agentic development, which, when deployed, is shown to modify offline behavior of participants. Discussion on how to apply this research to learning is also discussed. |
| Spring 2022 | DeVore, Susanne | PhD | *Syntactic Complexity and its Development in Early Learners of Mandarin*<br>This two-part corpus-based study examined L2 Mandarin speaker's writing samples. The first part of the study used a modified version of the Tool for Automated Analysis of Syntactic Sophistication and Complexity to identify and tally indices of proficiency in the corpus. Results of this portion were analyzed using linear and polynomial multiple regression models in order to see which type of model fit the data best. Results of this portion of the study were then used to research the development of L2 proficiency. Due to various challenges in accounting for lexical, phrasal, and clausal level constructions, as well as how these constructions are embedded, Network Science was adopted to provide a holistic account of development. Results of the first part of the study showed usage-based |

| | | | |
|---|---|---|---|
| | | | indices were more suited for use as a predictor of proficiency in L2 Mandarin. The second portion demonstrated how Network Science can be used as a framework in L2 development research. |
| Spring 2022 | Kitada, Katsuhiko | MA | *The Use of English Loanwords in Question-Answer Sequences of Question Time in the Japanese Parliament*<br>The use of English loan words throughout institutional talk of the Japanese Parliament was examined in this study using a conversation analytic approach. 29 videos of "Question Time" during debates between political parties were analyzed for the use of such words. Line-by-line analysis of the opening and closing sequences of the conversations revealed that the use of such loan words related to the construction of political arguments, clarification of technical terms, and the identity of the speaker. Additionally, the use of these words in political speech was discussed. |
| Spring 2022 | Park, Heejin | MA | *Is CLP Possible for Korean Law Professionals to Develop Their Multicultural Competence? A Critical Study of Korean Lawyers' Views Towards Multiculturalism*<br>This critical needs analysis using ethnographic interviews reported on the education and training of 5 Korean lawyers and their views on multicultural interaction. The study explored how to develop cultural sensitivity through the practice of critical language pedagogy (CLP). Participant responses coincided with professional reasoning and values. The integration of CLP into English for Specific Purposes (ESP) coursework was recommended to foster multicultural sensitivity. |
| Fall 2021 | Abe, Carolyn | MA | *Insight Gained from the Observation of Two English as a Foreign Language Reading Courses* |
| Fall 2021 | Choo, Sun Young | MA | *Age and Korean Language Learners' Use of Mobile Applications*<br>As COVID-19 halted study abroad for many, learners resorted to online platforms to fulfill their language learning needs. This was made possible through mobile assisted language learning (MALL), where there are language resource apps (e.g., Google Translate) and language learning apps (e.g., Duolingo). The goal of this study was to investigate L2 |

| | | | |
|---|---|---|---|
| | | | Korean learners' use of MALL and to specifically look at Duolingo, which has been one of the most successful MALL apps on the market. Data collected from a survey showed that the greatest challenge for L2 Korean learners in using MALL was the lack of opportunities for interaction with native speakers. Looking more closely at the apps learners used, they largely used language resource apps, such as Papago, Naver Dictionary, and Google Translate and entertainment apps, such as YouTube and Netflix. Furthermore, L2 Korean learners over 60 years old used translation and dictionary apps more than younger learners in their 20s, though these two groups used Papago equally as much. Language learning apps, such as Duolingo, were also still relatively new for L2 Korean learners as participants in this study continued to use language resource apps more than Duolingo, especially among older participants. |
| Fall 2021 | Davis, Abigail | MA | *Effects of Immersion on JFL Learners' Intensive Listening Skills: Vowel Length Contrasts* |
| Fall 2021 | Jung, Hyeyoung | PhD | *Towards Critical Literacy in Korean High School EFL Classrooms: Narrative Inquiry into Teacher Emotions and the Use of Critical Materials*<br><br>This study examined teachers' emotions towards using critical literacy in Korean high school EFL classes. In using a narrative inquiry approach, the findings from this study showed that teachers felt positively towards teaching critical literacy. As the teachers began teaching critical literacy, they noticed that their interactions with the students became more meaningful through class discussions about real-world topics that extended beyond the classroom. This approach to language education in Korea is still relatively new as it breaks away from the traditional top-down structure of education in Korea. Rather, teachers who used a critical literacy approach had agency to make changes in education. |
| Fall 2021 | Kang, Jia | AGC | *Binary Categorization Practices in a Korean TV Show* |

| | | | |
|---|---|---|---|
| Fall 2021 | Katz, Shayna | AGC | *Rapport in the World Language Classroom: From Face-to-Face to Online in Times of Pandemic*<br><br>This study took place as the COVID-19 pandemic happened in 2020. The author collected students' perceptions about the transition from face-to-face (FtoF) to online synchronous classes. Using an open and closed-item questionnaire format, L2 Spanish students reported on their perceptions about FtoF versus online classes in terms of supporting the different types of relationships. This study found that students perceived FtoF classes had a greater impact on teacher-student and student-student relationships, due to the interaction and personal experiences FtoF classes offered. Students had a stronger preference for feedback in FtoF classes as they claimed it was more effective because it was easier for them to ask questions. While positive comments, corrective feedback, and personal thematic discourse were all found to have a positive impact on the students' L2 development, positive comments and corrective feedback were most helpful in the development of positive teacher-student relationships while personal thematic discourse was the most helpful for positive student-student relationships. |
| Fall 2021 | Kunimatsu, Hiroko | MA | *A Needs Analysis of English Teachers at a Public High School in Japan for a New English Curriculum* |
| Fall 2021 | Lim, Soo Jin | MA | *Parents' Perceptions and Experiences of Early English Education in South Korea: A Focus on English Kindergartens*<br>As many studies report on the effectiveness of early English education in South Korea, this study investigated the perceived need for and satisfaction of early English education from the parents' point of view. Currently, English is mandated in all public schools beginning in the third grade. However, some parents in Korea want their child to learn English as early as possible. Parents who participated in this study explained that they sent their child to a private English kindergarten school because they believe learning English at a young age is easier and becoming proficient in English is necessary to compete in a globalized society. The main goal of having their children become proficient in English as early as |

| | | | |
|---|---|---|---|
| | | | possible is to ensure their best chance of getting good grades in school, passing entrance exams, and securing a good job. Many of the parents in this study preferred academic-oriented kindergartens where students are taught all subjects in English. However, the quality of these educational institutions was based on reputation from parenting forums rather than measurable qualities. Parents also preferred teachers who had a background related to early childhood education, TESOL certification, and were native speakers of English. Parents who sent their children to these English kindergarten schools were highly satisfied as they observed their child had higher self-confidence in their English abilities. |
| Fall 2021 | Ling, Wenyi | PhD | *The Perception, Processing and Learning of Mandarin Lexical Tone by Second Language Speakers*<br><br>Learning Mandarin has been considered difficult for learners whose first language does not have tone. This dissertation sought to understand native English speakers' perspectives and challenges with tones as they learned Mandarin, specifically how they perceived, processed, and learned lexical tones. Through an identification task and discrimination task, the results showed that learners with higher proficiencies had a higher categorical perception of tone, meaning they were better able to form mental categories of the tones they heard. Findings from this dissertation also showed that L2 Mandarin learners process tone differently from native Mandarin speakers. This dissertation also demonstrated that cue-focus training was not an effective pedagogical approach to learn tonal languages. |
| Fall 2021 | Otto, Jeffrey | MA | *Homeless in Hawai'i: Developing Critical Materials for an Intensive English Program*<br><br>This scholarly paper took a critical language pedagogy approach to materials development for English as a second language. The goal of this study was to design materials and activities that were based on interviews with target language community members. More specifically, this study draws attention to how social issues and social justice topics, such as homelessness, can be brought into the classroom to challenge the stereotypes and perceptions of marginalized groups held by study abroad students. The materials were |

| | | | |
|---|---|---|---|
| | | | developed from an interview that the author conducted with a homeless person, who's narrative was different from the prevailing stereotype about homeless people. |
| Fall 2021 | Rock, Kristin | PhD | *Using Analytic Rubrics to Support Second Language Writing Development in Online Tasks*<br><br>This dissertation used a mixed methods approach in generating a rubric to assess academic blog posts. Generating the rubric began with 148 blog posts in which six raters placed each post into one of six levels based on merit. In evaluating the blog posts, the six raters provided qualitative comments, which were then used as the descriptors for each level on the rubric. At the end of the first phase, the rubric consisted of five categories with six levels. The goal of the second phase in this project was to revise and refine the rubric in which 163 blog posts were evaluated by six new raters. By the end of the second phase, the rubric had the same five categories, but the levels were reduced to four with more refined descriptors. In using the rubric, this study found that when students were provided with the rubric while they completed their writing assignment, they performed better than students who were not provided with the rubric. This study also delved into the learners' longitudinal development, specifically looking at learners' linguistic and rhetorical move development. |