ANALYZING THE INFLUENCE OF UTTERANCE FLUENCY ACROSS ASSESSED SPEAKING PROFILES IN THE ORAL ENGLISH PLACEMENT TEST AT THE UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

BY

ROSANA ALEJANDRA GOMEZ-CAYAPU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in Teaching of English as a Second Language
in the Graduate College of the
University of Illinois Urbana-Champaign, 2023

Urbana, Illinois

Advisers:

Teaching Assistant Professor Suzanne Franks
Associate Professor Xun Yan

ABSTRACT

With the aim of informing the assessment of fluency within the Oral English Placement Test, this study analyzes the correlation between utterance fluency measures and proficiency scores from one of the test tasks. Furthermore, it examines how speed and breakdown fluency variables distinguish test takers across a five-profile scale. To address both inquiries, sixty student recordings were used for automated measurements and thirty for those that needed manual coding.

The findings from a Spearman's rank correlation analysis and multiple one-way Analyses of Variance revealed measurements of speed fluency (i.e. speech rate, articulation rate, and mean length of run) and certain measures of breakdown fluency (i.e. frequency of silent and mid-clause pauses) correlate with proficiency scores. Additionally, utterance fluency measures overall tend to discriminate between profiles at the far ends of the scale but not necessarily within adjacent profiles, except for articulation rate. These results further support the test validity and inform the design of rating descriptors, as well as teaching practices to develop oral proficiency at different stages.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Attaining fluent speech is a paramount goal for most L2 speakers. Being perceived as fluent can lead to better academic and job opportunities as well as a smoother insertion into a new culture. Consequently, fluency has traditionally been a major area of focus across different domains. On the one hand, it concerns educators wanting to support their students in developing their speaking skills, and on the other, it is of great interest for researchers and administrators in language testing seeking to provide examinees with an accurate depiction of their proficiency. To achieve both endeavors, it is crucial to start by defining what fluency is. One of the approaches in second language assessment to address this question has been determining which measurable criteria, such as aspects of speed or pausing, can better reflect overall proficiency.

The present study will follow such a line of research by examining the correlation between fluency measures and proficiency scores given by oral EPT raters, as well as analyzing whether fluency measures can distinguish test takers across different profiles of performance. Thus, this study aims to contribute to the improvement of descriptors in the rating scale and provide further support for the validity of the test. Regarding teaching, it is expected to inform instructors about the fluency aspects that contribute to effective communication and the fluency components that may challenge students.

CHAPTER 2: LITERATURE REVIEW

DEFINING FLUENCY

One of the foundational definitions of L2 fluency was proposed by Lennon (1990). He classified fluency into two categories: Broad and narrow. In the broad sense, fluency is defined, in lay terms, as advanced oral competence. It is the type of term people would use in a job application form to indicate they are proficient in a language. In contrast, in the narrow sense, fluency is understood as one of the dimensions of oral proficiency along with others such as lexico-grammar accuracy and complexity, giving learners a more comprehensive view of their speaking proficiency. Earlier definitions of fluency sided with either the listener's view or the speaker's view by focusing on ease of perception or production, respectively (de Jong, 2018).

Addressing both perspectives and understanding fluency as a complex phenomenon, Segalowitz (2010) categorized fluency into 3 constructs: First, cognitive fluency, which describes the cognitive processes the speaker undergoes to produce a speech that resembles a native one on its smoothness and automaticity. The second category, utterance fluency, addresses the resulting utterances that can be analyzed into objective measures of fluent speech such as speed, the length, and frequency of certain types of pauses, and repair techniques. Lastly, the author adds perceived fluency to the triad referring to the interlocutor's perception of the speaker's cognitive fluency based on their performance.

MEASURING FLUENCY

Advocating for a "finer-grained" approach to fluency analysis, Skehan (2003) divides utterance fluency into three domains: Breakdown fluency "indexed by pausing"; repair fluency "indexed by measures such as reformulation, repetition, false starts, and replacements"; and speed fluency "with measures such as syllables per minute" (p. 513). Several studies on fluency have focused on the former objective measurements as a means to gain more insights into the oral proficiency of L2 speakers. Given the numerous possibilities, choosing adequate measurements is a common challenge for researchers. The following overview presents a rationale for the ones selected in this study.

Speed fluency

According to Tavakoli and Wright (2020), research in the field of SLA has contributed to a deeper understanding of the speech fluency domain. One of the significant changes was gaining a comprehensive outlook on the information the measurement can provide about speech production including the articulatory and monitoring stages as well as compensation strategies employed by speakers (e.g. vowel lengthening). Likewise, it has been revealed that although speed improves along with proficiency, at some point speed reaches a plateau, usually around the B2 and C1 levels as defined by the CEFR (2001). Therefore, it does not constitute a crucial variable when discriminating across intermediate to advanced levels. Additionally, the authors also suggest incorporating both pure (e.g. articulation rate) and composite measures (e.g. mean length of run) to control for speed with and without pauses. This distinction is important as pausing behavior is only one among many variables leading to a slow speed. In terms of data analysis, an important change was brought out by the introduction of digital tools to facilitate the treatment of larger amounts of speech samples (de Jong, 2018). Among all programs, Praat (Boersma & Weenink, 2023), one of the software used in this study, stands out as a popular choice offering features like annotating files, using scripts, and exporting data to other programs.

Breakdown fluency

This domain has also undergone changes in the way it is operationalized. When carrying out research in the domain of L2 fluency it is important to agree upon the definition of pause in terms of duration. With the aim of standardizing this measurement for comparability purposes across fluency studies, de Jong and Bosker (2013) set out to determine an optimal pause threshold. The authors concluded pauses longer than 250ms are noticeable enough to affect the listener's perception. Since then, other researchers have followed the same threshold validating its reliability (Yan et al. 2020; Hunter 2017; Kahng 2020; Suzuki & Kormos, 2023).

In addition to pause length, research in L2 fluency also focuses on pause frequency. Although some studies support the use of pause length as a predictor of proficiency (Wang, 2014), most studies agree on pause frequency as a more reliable predictor of proficiency, especially when researching cognitive fluency (Bosker et al. 2014; Prefontaine & Kormos 2016). As discussed by Hunter (2017) this is because a higher frequency of pauses will increase the chances that pauses occur in the middle of a clause, which are more salient compared to pauses at the end of the clause regardless of their length.

Consequently, pause location has also been acknowledged by contemporary research as a key criterion for discriminating between fluent and disfluent speakers. Given novice L2 learners are still developing their lexico-grammatical repertoire, they will tend to pause in the middle of clauses to deal with gaps in their linguistic knowledge while formulating their utterances. In contrast, it has been proven that L1 speakers can sustain speech in complete clauses. Therefore, pausing will occur at the end of idea units in response to conceptualization processes (e.g. planning upcoming utterances). (de Jong, 2016).

Another important criterion within breakdown fluency is considering silent versus filled pauses, also referred to as pause character. Both types are usually associated with disfluency instances as they signal an effort on behalf of the speaker to formulate their message. However, some scholars challenge this perspective arguing that filled pauses also have a communicative purpose (Kosmala & Crible, 2021; Tottie, 2016). They contribute to discourse organization, add emphasis, and may assist in turn-taking. These functions will vary depending on factors such as culture, context, register, and gender. In this regard, filled pauses have a dual quality and therefore cannot be categorized only as disfluency instances. In a single speech sample, filled pauses can serve both functions. A more accurate measurement of fluency will then control for these variables instead of solely relying on frequency and length.

Similarly, Hunter (2017) draws attention to the issue of pause categorization when silent and filled pauses occur in a sequence. In this scenario, there is a lack of consensus among researchers on how to count these occurrences. Considering stem from planning processes, the author claims they should be grouped into the same "hesitation cluster" (p. 169).

Repair fluency

Repair fluency refers to the speakers' "self-monitoring processes of speech production" (Tavakoli and Wright, 2020, p.52) with the aim of improving their intended message. Commonly used categories include false starts, hesitations, repetitions, reformulations, and self-corrections. The average for each is then calculated by dividing the number of occurrences into 60 (standing for seconds). Although there is fair agreement on the reliability of speed and breakdown fluency measurements, the authors state measuring repair fluency can be more complex. First, some categories overlap as in the case of hesitations (pre…sentation) and repetitions (pre…sentation, presentation) leading to variables that are not independent. In addition, categorization may vary across studies. While some will define verbatim repetitions as part of repair fluency (New Directions East Asia, 2023) others will find it more suitable within breakdown fluency (Dörnyei & Kormos, 1998; Witton–Davies, 2014). Lastly, some repair strategies such as repetitions may not necessarily signal disfluency, but rather personal styles that may be carried from the speakers' L1 (DeJong et al. 2015) or even the participants' perceptions of the value of self-corrections (Kahng, 2014).

It may be the case that associated with these limitations, several studies have found a low correlation between proficiency scores and repair phenomena. For a better approach to measuring repair fluency Hunter (2017) recommends looking into different types of error repairs (i.e. lexico-grammatical repairs/linguistic form repairs versus content appropriateness repairs) as well as distinguishing between overt versus covert errors, that is, errors that can be identified versus those that cannot. For instance, in the case of covert errors, they can be grammatically correct but do not match the examinee's intended meaning. Studies looking into the latter type of resort to stimulated recalls to identify the origin of the repair (Khang, 2014).

With no access to retrospective comments from test takers and a lack of emphasis on lexico-grammatical analysis, the present study has opted to exclude repair fluency.

FLUENCY IN LANGUAGE TESTING

Scale Descriptors

Within L2 oral assessment, fluency has always been portrayed as an important criterion of language proficiency. Early documented tests such as The Cambridge Certificate of Proficiency in English for English teacher trainees in 1913 and The Foreign Service Institute (FSI) oral interview for military recruitment in 1956 already included fluency as an item in their rubrics along with other criteria such as accent, comprehension, grammar, and vocabulary. Nonetheless, fluency was not explicitly defined or assigned an independent category in either of the tests. In the Cambridge Certificate of Proficiency, the levels of proficiency ranged from 1.0 to 5.0 and the major speaking categories were divided into grammatical resource, lexical resource, discourse management, pronunciation, and interactive communication. The word fluency is not employed, but there are some references to fluency measurements such as "appropriate length of contributions to develop the discourse" in discourse management and "participate in the development of the interaction without undue hesitation" (Weir and Milanovic, 2003, p. 470) under interactive communication. Similarly, in the FSI oral interview, fluency was included as follows: "There is a slight increase in utterance length, but frequent long pauses and repetitions of interlocutor's words still occur" (Novice-High), "Shows some spontaneity in language production but fluency is very uneven" (Intermediate-High), and "Often shows remarkable fluency and ease of speech" (Advanced Plus). (Lowe, 1983, p. 240). No definition was provided as to what fluency exactly meant. Levels for this test ranged from zero to five.

Since then, the approach to L2 fluency representation in rating scales has undergone numerous changes. Taking large international tests as a basis for analysis, Tavakoli and Wright (2020) summarize some of the successes and areas for revision as informed by current research in second language acquisition and language testing.

The authors start by highlighting the tests have a comprehensive definition of the speaking construct as they incorporate the three components of utterance fluency (speed, breakdown, and repair) as well as cognitive fluency when accounting for the origin of disfluencies as shown in the Aptis test "may be hesitant when searching for patterns of expression". Perceived fluency is also included, taking into consideration the listener's degree of effort to comprehend the message. For example, TOEFL iBT states "listener effort is needed because of… choppy rhythm/pace" (Tavakoli & Wright, 2020, p. 107). Lastly, when writing rating descriptors the authors highlight the importance of bringing raters' attention to characteristics of fluency that may co-occur instead of artificially presenting them as excluding traits. For instance, the IELTS test includes: “usually maintains flow of speech but uses repetition, self-correction or hesitation” (Level 5: Modest).

Following their description of strengths, Tavakoli and Wright (2020) present some crucial considerations for developing more accurate and easy-to-interpret scales. First, they suggest avoiding ambiguous descriptors that may be misinterpreted across raters such as “some fluidity”, “generally clear”, and “noticeable hesitations”. They also warn about the treatment of breakdown and repair fluency as exclusively linked to disfluency. As discussed above, some of these behaviors may be associated with personal style (deJong et al. 2015). This will lead to validity issues as the test will be measuring a variable irrelevant to L2 proficiency. As a final note, contrary to recent findings in language testing some rating descriptors in scales assume a “linear correlation” between proficiency levels and fluency features (deJong, 2018). Some of the underlying assumptions are an improvement in speed fluency for higher levels and more instances of breakdown or repair fluency in lower levels. This perspective can mislead raters to give favorable scores based on a single criterion such as fast speed rather than weighing in other components such as lexico-grammatical accuracy, coherent speech, or content relevance.

Task Type

Besides scale design, another key consideration in L2 oral assessment is choosing the tasks that best elicit language production within the targeted context. The lecture task researched in this study falls under the category of integrated speaking tasks. In this type, test takers are asked to respond to a prompt based on outside materials such as a reading or a lecture (Dimova et al., 2020). This modality is popular among tests for academic purposes as it emulates authentic scenarios, partially ensures equal background information on the topic among test takers, and taps into other cognitive skills involved in language production such as organizing and manipulating information from source(s) (Nakatsuhara et al. 2021).

PREVIOUS STUDIES

Further contributing to the understanding of fluency within language testing, numerous studies have analyzed the correlation between measures of utterance fluency and proficiency scores across different levels of performance, inquiring about the insights each measure can provide about this assessment criterion.

One of the foundational studies following this approach was conducted by Iwashita et al. (2008). The authors set out to investigate the relationship between features of spoken language and the raters' scores given to two hundred test takers on a prototype of the TOEFL iBT test comprised of five tasks and five proficiency levels. Throughout different types of measurements, the study analyzed students' grammatical accuracy, lexical range, pronunciation and fluency. The results demonstrated pronunciation and fluency were the two features that influenced scores more strongly. Regarding fluency measures, the author demonstrated speech rate, number of unfilled pauses and total pause time showed both a strong correlation with proficiency scores, and potential to discriminate examinees across different levels. On the other hand, no significant differences were found for measures such as filled pauses, repairs, and mean length of run.

Transitioning to the context of local tests, Ginther et. al (2010) analyzed the performance of 150 test-takers in the OEPT (Oral English Proficiency Test) across internally set levels (3 to 6). Three measures of speed fluency were included as well as pause patterns such as frequency and length of total silent and filled pauses. The authors found strong to moderate correlations between proficiency scores and speed rate, articulation, rate, and mean length of run. They suggest the results for speed fluency may be indicative of biased scores for fast speakers regardless of other aspects of performance. On the other hand, the results obtained for mean length of run are attributed to the capacity of high-level speakers to concatenate longer and more complex ideas without interruptions. In contrast to these results, low correlations were found between OEPT scores and breakdown fluency. Regarding silent pauses, the authors recommended studying pausing patterns across a wider range of levels and including the students' L1 as a variable. For filled pauses, the results can be attributed to the low percentage of response time these pauses represented and their inability to distinguish examinees across levels.

Studying similar fluency measures, Bosker et al. (2013) analyzed the contributions of speed, pause, and repair to perceived fluency. In their study, eighty untrained raters evaluated the speaking performance of thirty-eight L2 Dutch speakers on eight computer-mediated tasks. Four experiments were carried out with different groups of raters. In experiment one the three components of utterance fluency were assessed. In experiments, two to four, raters were asked to focus on either speed, breakdown, or repair fluency. Although the results for speed fluency aligned with those found by Ginther et al. (2010), breakdown fluency measures, in contrast, highly correlated with proficiency scores. It is important to note these results corresponded to silent pauses only. Similar to Ginther's study, the effect of filled pauses was irrelevant. Lastly, the authors argue repair fluency did not strongly correlate with the scores given. Agreeing with Hunter (2017) Bosker et al. 2013 consider these results stem from methodological issues. Perhaps, repair fluency is better examined when controlling for the types of errors that are repaired instead of just relying on the frequency of cases.

Working with a larger data set, Tavakoli et al. (2023) investigated the construct of fluency within the Test of English for Educational Purposes (TEEP). Fifty-six samples from a monologic task were used to examine features of speed, breakdown and repair fluency across a four-level scale with levels compared to those ranging from independent to proficient user in IELTS (5.0, 5.5, 6.5, 7.5). According to the authors, the findings replicate those from similar studies asserting speed measurements distinguish between high and low proficiency learners; silent pauses differentiate learners at adjacent levels for proficient learners but not independent learners while repair measurements and filled pauses did not differentiate across any levels. Similar to Iwashita et al. (2008) and Ginther et al. (2010) the results contributed to modify the rater descriptors in the scales in order to achieve a more accurate measurement of the construct.

Acknowledging the pressing need to keep contributing to the definition of the construct of fluency for both second language testing and teaching from a new context, this study proposes the following research questions:

RQ1: How do utterance fluency measures correlate with the five profiles in the oral English Placement Test scale?

RQ2: Could utterance fluency measures distinguish between examinees across the five profiles in the oral English placement test?

CHAPTER 3 METHODOLOGY

THE ENGLISH PLACEMENT TEST

The English Placement Test (hereafter, EPT) at the University of Illinois Urbana-Champaign assesses the English proficiency of undergraduate and graduate students as determined by campus requirements. Undergraduate students, the population in this study, are exempt from taking the test depending on their scores on either TOEFL iBT (Total of 103 and above; speaking sub-score of 23), IELTS (Total of 7.5 or above, speaking sub-score of 6.5), or the Duolingo Test (135 or above). For non-exempt students without additional proof of speaking proficiency, the oral EPT test is required.

The EPT assesses both writing and oral proficiency. In the writing test, students compose an argumentative essay based on the content provided in short reading passages and a video lecture. In the oral test, they complete four tasks based on the same topic. However, only the lecture task was analyzed using two different topics: Public versus private transportation, and books versus e-books. The test starts with three warm-up questions and then moves to the main items: a read-aloud, video lecture and summary question, a conversation response, and graph description. All the tasks are monologic, and the students have about ten minutes to complete the test. Responses are recorded on Moodle HQ (2023).

The test is assessed with a profile-based rating scale with three proficiency levels and five performance profiles as shown in Table 1.

Table 1: Profiles and Placement Levels

Level	Description	Placement
High	Strong in proficiency and fluency	Proficient (Exempt)

Table 1 (cont.)

Mid profile 1	Pronunciation issues	Advanced (ESL 110 recommended)
Mid profile 2	Delivery issues	
Low profile 1	Major pronunciation issues	Developing (ESL 110 strongly recommended)
Low profile 2	Fluency issues	

Passing scores will exempt students from taking ESL courses. Otherwise, to fulfill graduation requirements, they must enroll in either a two-course sequence (ESL 111-112) or an advanced-level course (ESL 115 or RHET 105) for composition. For the oral test, students are recommended (advanced) or strongly recommended (developing) to enroll in ESL 110 "English Pronunciation and Oral Fluency".

RATERS

The oral EPT raters are usually recruited from the pool of graduate assistants teaching at the ESL courses, most of whom are international students. Their certification process consists of three rounds of ratings and three discussion meetings. In the first meeting, raters review the benchmark and get acquainted with the rater scoring sheet. In the second and third meetings, the EPT coordinator and raters discuss their performance and go over misaligned samples. At each rating round, participants are asked to rate fifteen speech samples. In rounds one and two, they are provided with the number of speech samples per level in each of the four test tasks (e.g. one read-aloud section from each level, one lecture summary from each level, two conversation responses from each level, and one graph description). In order to increase complexity, in round three, although the number of speech samples will be the same for the read-aloud section and the lecture, for the graph description, the options will be open to five profiles, and there will be no level hints for the conversation responses. At the end of the process, raters will meet individually with the coordinator to discuss their performance and, if successful, receive their certification.

DATA CODING

The sixty speech files were de-identified and categorized in placement levels by the EPT coordinator after signing the EPT confidentiality agreement form and receiving IRB approval. A few files were enhanced by amplifying volume and reducing echoing using Audacity 3.2.2 (The Audacity Team, 2022). For automatic measurements, a Praat script (deJong et al., 2021) was used to obtain phonation time, speech rate, articulation rate, average syllable duration, number of syllables, and number of silent pauses set to 250 milliseconds. The results were tabulated and exported to a Google Sheets File.

Lastly, the frequency of mid-clause pauses was manually calculated. The speech files were uploaded to Otter.ai (Liang & Fu, 2016), an online platform powered by artificial intelligence to automatically generate orthographic transcriptions. These were revised for word accuracy and then parsed into Analysis of Speech Units (ASUs) and then clauses. The results obtained from running the Praat Script (de Jong et al., 2021) with silent pauses were imported into ELAN 6.6. (Sloetjes & Wittenburg, 2008). There, the ASUs were marked again using the segmentation tool. This study chose Analysis of Speech Unit (ASU hereafter) given its suitability to analyze authentic speech over other units such as the T-unit understood as “a main clause plus any other clauses that are dependent on it, but it excludes non-clausal structures and sentence fragments” or the C-unit defined as “an utterance providing referential or pragmatic meaning, consisting of either a simple clause or an independent subclause, together with subordinate clauses associated with it” (Tavakoli & Wright, 2020). On the other hand, Foster et al. (2000) define an ASU unit as “a single speaker’s utterance consisting of an independent clause, or sub clausal unit, together with any subordinate clause(s) associated with either" (p.365). A clause is defined here as any unit containing a finite or non-finite verb along with another clause element (subject, object, complement, or adverbial). Choosing a grammatical unit such as this increases reliability as syntactic units are easier to analyze compared to other more ambiguous criteria such as intonation or semantics (Foster et al. 2000). Furthermore, clause division reflects the speakers' ability to plan speech, and including units other than clauses such as sub-clausal units or subordination facilitates the analysis of the speakers' ability to plan complex ideas, reflecting higher proficiency, and also accounts for authentic phenomena in speech such as abandonment, repetition, and false starts.

As a final step in the data coding process, the annotations from Otter.ai were copied and pasted into the annotation grid. The author listened to the speech file once more to mark the silent pauses occurring in the middle of clauses using the silent pause grid (250 ms or longer) from Praat as a baseline. The ELAN file was exported as Traditional Transcript Text to then count the frequency of mid-clause pauses, marked with an asterisk (*) using a word processor.

DATA ANALYSIS

First, students' scores at each profile were transformed into ordinal data (i.e. L2: 1, L1: 2, M2: 3, M1: 4, H:5) in order to carry out the statistical analysis. For Research Question 1 inquiring about the correlation between the utterance fluency measures and the profile scores, Spearman's rank-order correlation was run using IBM SPSS Statistics 27 (IBM, 2020). For Research Question 2, multiple one-way analyses of variance (ANOVA) were computed to determine whether utterances fluency measures could distinguish examinees across the five Oral EPT profile levels. Subsequently, a post hoc analysis using Bonferroni's test was run to identify the specific scoring differences across the scale. The same software was used for all the statistical analysis. Given the variable of frequency of mid-clause pauses had a smaller sample size, its ANOVA was computed separately.

UTTERANCE FLUENCY MEASURES

The following measures were chosen for the present study. Speech rate and articulation rate were automatically generated by the Praat script “Syllable Nuclei Version 3” (deJong et al., 2021). Mean length of run and calculations for breakdown fluency measures were made using Google Sheets.

Speed measurements

- Speech rate: total number of syllables divided by the total amount of phonation (including pauses)
- Articulation rate: Total number of syllables divided by the total amount of phonation time (excluding pauses)
- Mean length of run: The mean number of syllables between two pauses. The measurement was calculated by dividing the number of syllables by the number of pauses.

Breakdown fluency measures

- Frequency of silent pauses: Number of pauses (250 milliseconds or longer) divided by phonation time.
- Mean length of silent pauses: Pausing time divided by the number of silent pauses.
- Frequency of mid-clause pauses: Number of mid-clause pauses divided by the number of syllables.

CHAPTER 4: RESULTS AND DISCUSSION

This chapter reports on the results of the statistical analysis carried out in this study. The chapter is divided into two sections. In the first section, the results of a Spearman's rank correlation between utterance fluency measures and the Oral EPT proficiency scores are reported. The section opens with the descriptive statistics for both speed and breakdown measurements (Tables 2 & 3) followed by the correlations obtained and a discussion of the results. In the second section, the results of multiple analysis of variance (ANOVAs) and a post hoc test using Bonferroni's method are reported through line graphs and tables displaying the multiple comparisons among the profiles in the rating scale. Next, a discussion of the results is presented. The chapter concludes by summarizing the main findings.

Table 2: Descriptive statistics for speed fluency measures

	Level	Mean	Std. Deviation	Minimum	Maximum
Speech Rate	H	3.3	0.4	2.61	3.75
	M1	2.9	0.4	2.38	3.86
	M2	2.7	0.3	2.11	3.17
	L1	2.7	0.4	2.01	3.12
	L2	2.4	0.5	1.56	3.02
Articulation Rate	H	4.3	0.4	3.75	4.82
	M1	3.8	0.3	3.37	4.32
	M2	3.6	0.3	3.01	4.14
	L1	3.7	0.4	2.97	4.39
	L2	3.6	0.3	3.15	4.19

Table 2 (cont.)

Mean Length of Run	H	9.6	1.7	7.28	13.28
	M1	8.2	4.5	4.76	20.95
	M2	7.3	1.7	4.87	10.83
	L1	8.0	3.2	3.97	14.04
	L2	5.9	1.6	3.60	9.29

Table 3: Descriptive statistics for breakdown fluency measures

	Level	Mean	Std. Deviation	Minimum	Maximum
Frequency of Silent Pauses	H	.46	.07	.34	.59
	M1	.55	.20	.21	.89
	M2	.52	.13	.34	.72
	L1	.53	.19	.27	.92
	L2	.65	.16	.40	1.06
Mean Length of Silent Pauses	H	2.9	.34	2.34	3.59
	M1	2.8	1.0	1.66	5.43
	M2	2.7	.52	2.22	3.75
	L1	2.9	.82	1.82	4.51
	L2	2.5	.37	2.10	3.16

Table 3 (cont.)

Frequency of Silent Pauses at mid-clause position	H	.05	.011	.0359	.0621
	M1	.07	.033	.0250	.1077
	M2	.08	.020	.0516	.1041
	L1	.06	.039	.0189	.1245
	L2	.10	.036	.0615	.1579

SPEARMAN'S RANK CORRELATIONS

Spearman's rank correlation analysis was conducted to assess the relationship between proficiency scores and utterance fluency measures. There was a moderately positive correlation between speech rate [r (60) = .543, p = <.001], articulation rate [r(60) = .485, p = <.001], and mean length of run [r(60) = .432, p = <.001]. As for breakdown fluency measures, negative correlations were observed for frequency of silent pauses [r (60) = -.331, p =.010], and frequency of mid-clause silent pauses [r(30) = -.409, p =.025]. No significant correlation was found between mean length of silent pauses and proficiency scores.

The results indicate speed fluency measures serve as a good predictor of proficiency ratings. As suggested by de Jong (2016) this outcome is expected as speed is a salient feature in the rating scale. Furthermore, the connection between speed and performance is also supported in the literature in terms of the information this measure renders about language processing. For instance, measures such as mean length of run, beyond speed, showcase the learners' ability to produce longer and more complex sentences (Ginther et al., 2010). Similarly, it may display gains in automaticity through, for example, the acquisition of formulaic language (Khang, 2014).

In comparison to speech rate, variables of breakdown fluency were a weaker predictor of proficiency scores. Reasons for this outcome may be linked to the limited information the number of pauses alone offers. This explains why frequency of mid-clause pauses displayed a higher correlation. This pausing behavior is more salient as it signals processing difficulties that result in breakdowns in communication usually accompanied by perceivable occurrences such as reformulations, repetitions, hesitations, and self-corrections. On the other hand, silent pauses measured with a threshold of 250 milliseconds (or longer) in this study were not always easy to identify by ear while coding the data. In this regard, frequent pauses may have been disregarded among lower proficiency examinees if pauses were too short.

ANALYSES OF VARIANCE (ANOVAs)

Prior to running the analyses, a series of assumption tests was conducted. First, Pearson correlations were performed among the utterance fluency measures reporting moderate correlations for most of the variables except mean length of silent pauses versus mean length of run showing the hypothesis of absence of multicollinearity was justifiable. In contrast, when checking the multivariate normality through multiple linear regression, it was revealed that the Mahalanobis distance exceeded the critical value of 20.52 for a six-variable model (max.value=42.41), indicating the assumption of multivariate normality was not tenable. Given these results, multiple one-way analyses of variance (ANOVA) were conducted to examine whether utterance fluency measures (i.e. speech rate, articulation rate, mean length of run, frequency of silent pauses, mean length of silent pauses, and frequency of mid-clause pauses) could distinguish examines across the five profiles of the oral EPT (i.e. high (H), mid 1 (M1), mid 2 (M2), low 1 (L1), and low 2 (L2)).

The results revealed that there was a statistically significant difference for at least two profiles in speech rate (F (4, 55)=8.951, p <.001); in articulation rate (F (4, 55)=8.572, p <.001); in mean length of run (F (4, 55)=22.641, p=.030); in frequency of silent pauses (F (4, 55)=2.488, p=.054). No such difference was found in mean length of silent pauses (F (4, 55)=.743, p=.567) or frequency of mid-clause silent pauses (F (27, 2)=.815, p=.691).

Post-Hoc Test

Post hoc tests were conducted to identify differences across levels of proficiency using the Bonferroni method at a 0.5 alpha level. The findings for each of the utterance fluency measures are reported below.

Speech Rate

The post-hoc analysis shows the L2 level was different from the M1 level, and the H level was different from the L2, L1, and M2 levels demonstrating no significant differences across adjacent levels (See table 4). It is worth noting there is a trend displaying speech rate increasing across levels, except for the L1 level which ranks higher than the following M2 level (See figure 1).

Figure 1: Speech rate across proficiency profiles.

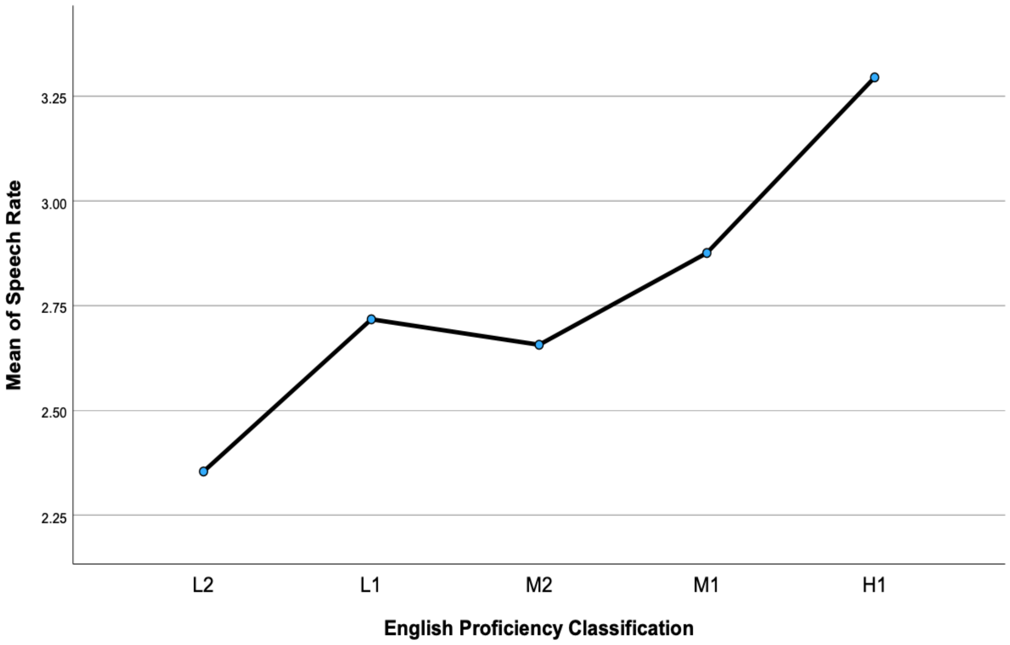


Table 4: Comparison of speech rate and English proficiency classification.

(I)Level	(J) Level	Mean I-J (Std. Error)	95% Confidence Interval	
			Lower Bound	Upper Bound
L2	M1	0.28 (.162)	-.9844	-.0339
	H	0.00 (1.62)	-1.4161	-.4656
L1	H	-5.77 (.162)	-1.0527	-.1023
M2	H	-.638 (.162)	-1.1136	-.1631
M1	L2	.509(.162)	.0339	.9844
H	L2	.940 (.162)	.4656	1.4161
	L1	.577 (.162)	.1023	1.0527
	M2	.683 (.162)	.1631	1.1136

Articulation Rate

For articulation rate the post hoc analysis reveals differences for the H level and the L2, L1, M1, and M2 levels with a statistically significant difference for the M1 level (See table 5). Similarly to speech rate, the L1 variable behaves differently as it performs higher than the level immediately above it (M2) (See figure 2).

Figure 2: Articulation rate across proficiency profiles.

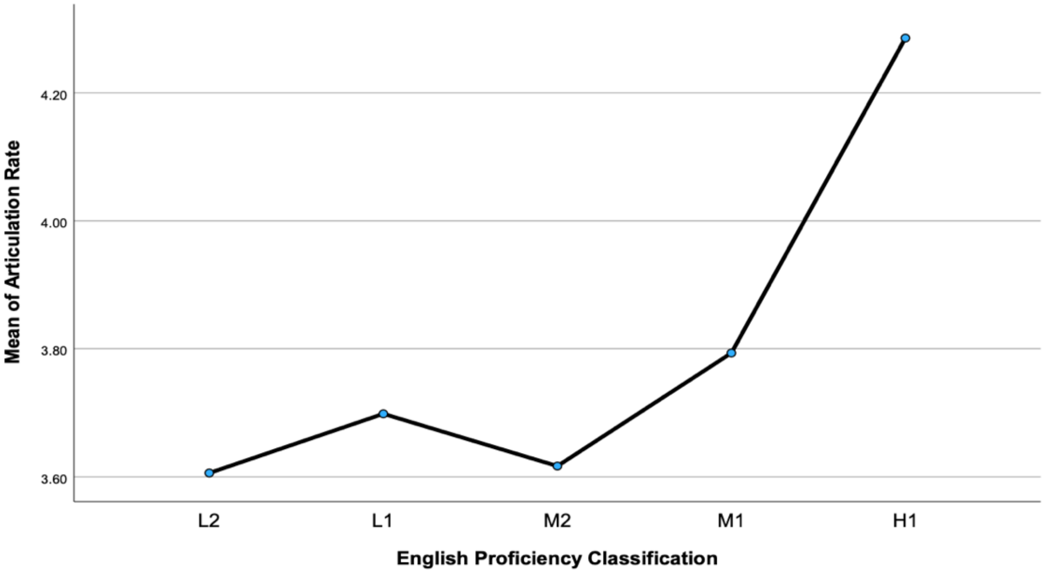


Table 5: Comparison of articulation rate and English proficiency classification.

(I)Level	(J) Level	Mean I-J (Std. Error)	95% Confidence Interval	
			Lower Bound	Upper Bound
L2	H	-.680 (.136)	-1.0781	-.2819

Table 5 (Cont.)

L1	H	-.587(.136)	-.9856	-.1894
M2	H	-.669(.136)	-1.0673	-.2710
M1	H	-.492(.136)	-.8906	-.0944
H	L2	.680(.136)	.2819	1.0781
	L1	.587(.136)	.1894	.9856
	M1	.669(.136)	.2710	1.0673
	M2	.492 (.136)	.0944	.8906

Mean length of Run

In contrast to the previous analysis, mean length of run only differentiated between levels at the extremes of the scale (L2 vs H) (See table 6). Consistent with the results for speech and articulation rate, the L1 level performed significantly higher than the M2 level, and nearly as high as the M1 level (See figure 3).

Figure 3: Mean length of run across proficiency profiles.

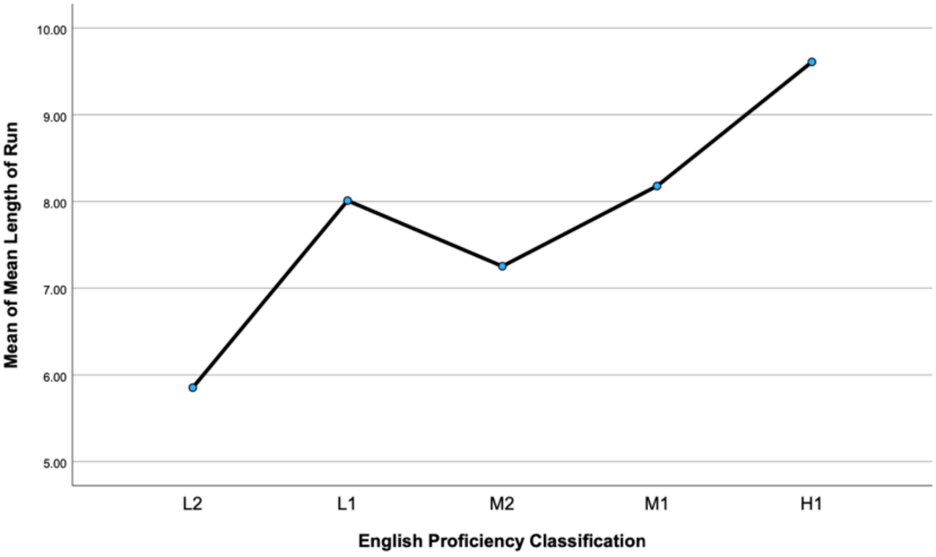


Table 6: Comparison of mean length of run and English proficiency classification.

(I)Level	(J) Level	Mean I-J (Std. Error)	95% Confidence Interval	
			Lower Bound	Upper Bound
L2	H	-3.75 (1.14)	-7.0966	-.4151
H	L2	3.75 (1.14)	.4151	7.0966

Frequency of Silent Pauses

The post hoc analysis for frequency of silent pauses demonstrated level differences at the extremes of the scale (L2 and H) (See table 7). A similar number of pauses were found for the L1 (M=.5250) and M2 levels (M=.5242), with a slight increase for the M1 level (M=.5475) (See figure 4).

Figure 4: Frequency of silent pauses across proficiency profiles.

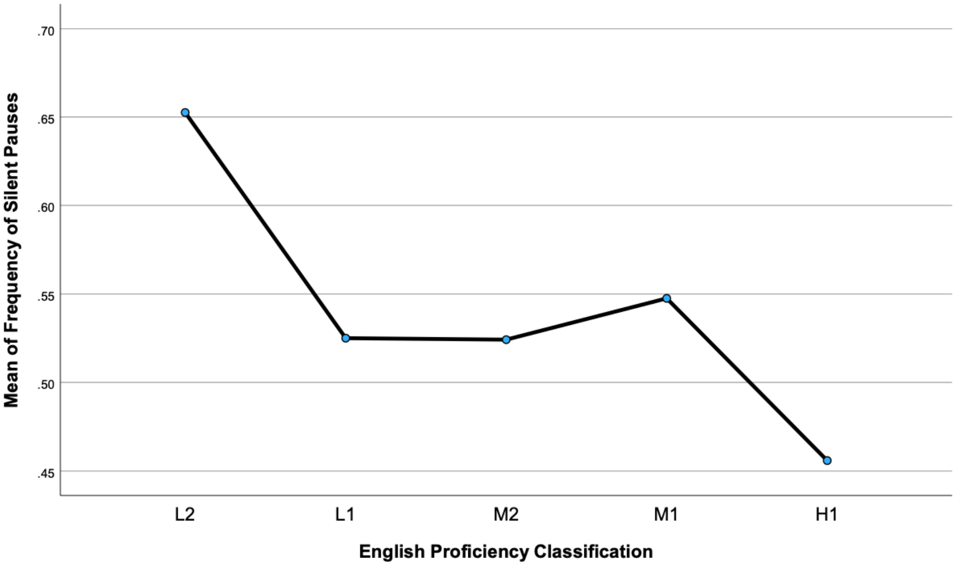


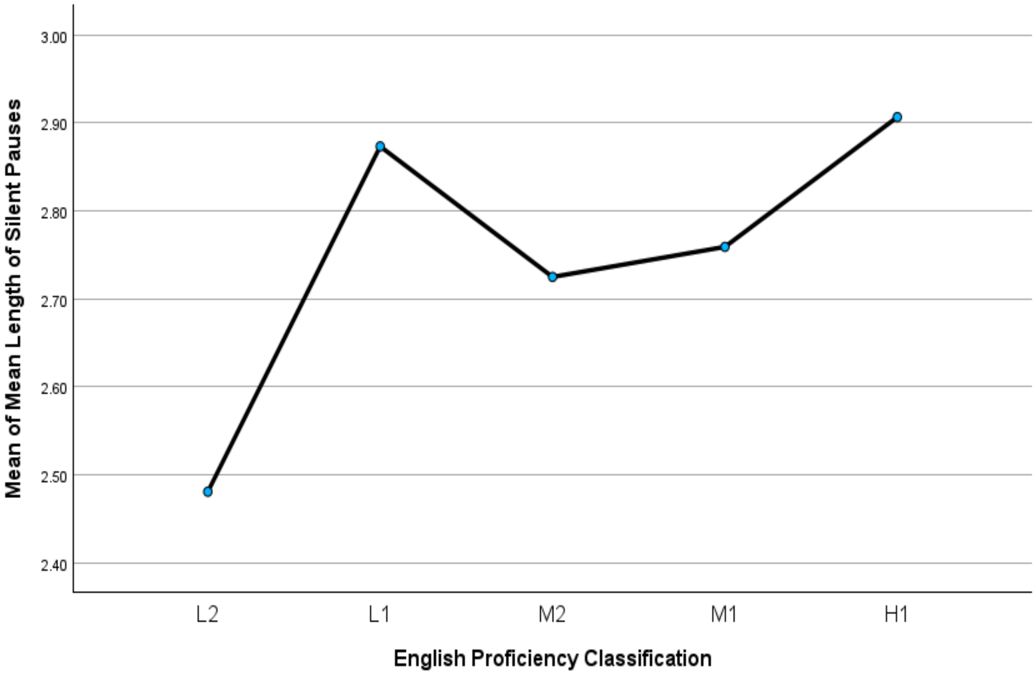
Table 7: Comparison of frequency of silent pauses and English proficiency classification.

(I)Level	(J) Level	Mean I-J (Std. Error)	95% Confidence Interval	
			Lower Bound	Upper Bound
L2	H	.196 (0.63)	.0100	.3833
H	L2	-.196 (0.63)	-.3833	-.0100

Mean Length of Silent Pauses

No statistically significant differences were found for the means of the mean length of silent pauses among profiles. However, a decreasing tendency of the means was observed. L2 (M=2.5), L1 (M= 2.9), M2 (M=2.7), M1 (M=2.8), H (M=2.9) (See figure 5).

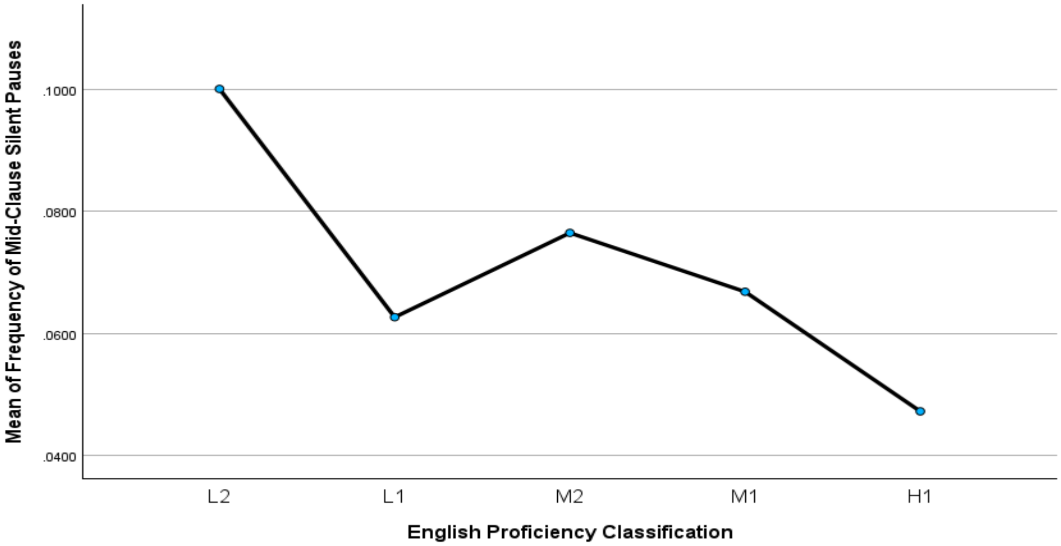
Figure 5: Mean length of silent pauses across proficiency profiles.



Frequency of mid-clause pauses

No statistically significant differences were found for the means of the mean length of silent pauses among profiles. However, a decreasing tendency of the means was observed. L2 (M=.10), L1 (M= .06), M2 (M=.08), M1 (M=.07), H (M=.05) (See figure 6).

Figure 6: Mean of frequency of mid-clause silent pauses across proficiency profiles.



DISCUSSION

Speed Fluency Measures

Overall the findings demonstrate that speed fluency measures (speech rate, articulation rate, and mean length of run) can distinguish examinees at different proficiency levels. While articulation rate can set apart H level students from all of the other levels, speech rate can discriminate examinees between the L2, L1, M2, and the H level, and between the L2 and M1 level. On the other hand, mean length of run could only differentiate between levels at the opposite ends of the scale (L2 vs. H). With the exception of articulation rate (H vs. M1 level), speed variables did not distinguish between adjacent levels. This trend suggests that, when faced with borderline scores, raters resort to other criteria on the scale (e.g. lexical variety, grammatical complexity) to decide. This is a positive outcome considering fast speech alone is not a reliable measurement for fluency (Ginther et al., 2010; Kormos & Denes, 2004). As proven in other studies, speed could even be a trait of personal style (De Jong et al., 2015; Segalowitz, 2016).

Furthermore, it is important to highlight that although speed measurements increase along with the proficiency level, this distribution is consistently disrupted at the L1 level. Admittedly, a linear relationship cannot always be assumed between fluency measures and proficiency scores. However, considering speed measurements tend to display a linear pattern, it is worth analyzing the behavior for the L1 variable in light of the oral EPT scale. See Appendix A for full descriptions that appear in Versions 2.0 of the EPT Oral Rating Rubric and Appendix B for Version 3.0.

The L1 level is reserved for examinees with “major pronunciation issues” while the adjacent M2 level considers “delivery issues: fluency and intonation”. This explains the scoring, demonstrating raters are consistent in following the rubric. Nonetheless, this categorization can be problematic because it does not consider participants with issues in both areas. Moreover, it seems the descriptors in the scale portray fluency in more advanced terms for the L1 profile level compared to the M2 profile. This difference could be considered inconsistent with level categorization as low profiles are labeled as “developing” and are required (grad)/ or strongly recommended (undergrad) for ESL courses. On the other hand, mid-profile scores are considered “advanced”, and thus undergraduate and graduate students receive only a recommendation to take ESL 110 or ESL 510 (i.e., they are not required to take any oral courses). In this sense, fluency could have been represented as a less prominent factor compared to other criteria in the scale.

Addressing both issues, the new oral EPT scale, version 3.0, replaced the L1 profile with a Mid-low profile to place students with “Mixed Issues”. With this change, the distinction between neighboring levels based on fluency is more transparent. For instance, while the Mid-low profile points out “Noticeable effort in lexical retrieval but attempts to speak at a coherent rate” the Mid profile 2 reads “Some effort in maintaining speech rate and pragmatically appropriated rhythm”. It is interesting to note that while the criteria of “Noticeable effort in lexical retrieval” was present in both L1 and M2 profiles in the old scale (version 2.0), in the new scale (version 3.0) it is exclusive of the Mid-low (developing) profile. This change is important because it creates a criterion that distinguishes between adjacent profiles, aligning with the literature stating that lexical retrieval is a clear indicator of low performance as it shows a lack of linguistic command, and consequently, automaticity (Hilton, 2008).

Breakdown Fluency Measures

The analysis for breakdown fluency measures shows only frequency of silent pauses could distinguish between test takers at the L2 and H1 level, indicating students with lower proficiency tend to pause more compared to advanced learners. In line with previous studies, despite distinguishing at the extremes of the scale, a linear relationship between the frequency of silent pauses and proficiency scores was not observed. The same result applied to mean length of silent pauses. As suggested by Park (2016) this can be explained by considering pauses comprised different variables such as rate, length, and location. For instance, it may be possible that speakers at higher proficiency levels pause longer as they are planning for content while debutant learners may produce shorter but more recurrent pauses, sometimes followed by self-corrections. This was a common pattern observed in the manual coding of the data and has also been discussed in the literature (Götz, 2013). It can then be concluded that the focus should not solely be on the number of pauses. Looking at other criteria, such as location, can provide valuable information on the underlying reasons for pausing behavior indicating whether it stems from linguistic limitations or formulation of ideas.

In the spirit of providing a comprehensive view of breakdown fluency, this study analyzed mid-clause pausing in addition to the other automatically measured variables. However, in contrast to other L2 fluency studies (De Jong, 2016; Tavakoli, 2011; Khang, 2014), frequency of silent pauses at mid-clause location only discriminated between the L2 and H level with an alpha of 0.05, suggesting results are not conclusive. A possible explanation is that, given the time commitment to annotate each of the files, only 50% of data was analyzed (30 speech samples). This factor may have lowered the statistical significance. Nonetheless, the decreasing tendency observed in the descriptive statistics coincides with the hypothesis of an increasing number of mid-clause pauses among examinees with lower proficiency.

The ideas presented above are in line with the design of both oral EPT scales (version 2..0 and version 3.0) for breakdown fluency. For instance, version 2.0, which was used to rate the speech files in this study, stressed that for the L1 level (issues with pronunciation but not fluency) long pauses were acceptable as long as they primarily occurred at expected places ("sentence/clause boundaries"). As for the new scale, the new profile (Mid-low, mixed issues) changes the term "pauses" for "reformulations", identifying a specific type of repair strategy, usually following pauses, that tend to occur more frequently among students with developing proficiency.

Similar to speech fluency, for breakdown fluency the new scale provides a contrasting criterion for adjacent levels. While Mid-low profile examinees will pause "*within* idea units or at unexpected spots", Mid 2 (M2) profile students will do so "largely *in between* major syntactic chunks or idea units" (i.e. words, phrases, or sentences expressed coherently).

In summary, 60 speech samples were analyzed automatically and 30 were analyzed manually. The results show speech rate, articulation rate and mean length of run were moderately correlated with the five proficiency profiles on the rating scale. Frequency of silent pauses and frequency of mid-clause silent pauses were negatively correlated, and no significant correlation was found for mean length of silent pauses. Regarding differences across the profiles, none of the measures except for articulation rate could distinguish between adjacent profiles. The results of this study support the changes made from version 2.0 to version 3.0 in terms of descriptors that can more transparently differentiate examinees across the rating scale. For instance, while the variable "Noticeable effort in lexical retrieval" was present in both the L1 and M2 adjacent profiles in version 2.0, in version 3.0 was solely reserved for the new Mid-Low profile. The next chapter will report on the implications and limitations of this study concluding with some directions for future research.

CHAPTER 5: IMPLICATIONS, LIMITATIONS, AND FURTHER RESEARCH

IMPLICATIONS OF THE STUDY

Regarding language testing, the findings in the present study provide further support for the changes in the new oral EPT rating scale version 3.0. The new rating scale includes more precise descriptors to differentiate between examinees of different speaking profiles compared to the version 2.0. For example, to capture developing proficiency, the criteria "struggle to retrieve words to express their ideas/opinions" and "noticeable effort in lexical retrieval" were reserved for the low profile and the mid-low profile, respectively. In contrast, in line with the literature asserting uttering ideas faster reflects an advanced command of the language as it translates into easy access to linguistic resources, an ability to create larger chunks of speech, and acquiring a good array of formulaic language (Hunter, 2017; Tavakoli and Wright, 2020), "speech rate" was exclusively allocated to mid-profiles and above (M1, M2, H). As for breakdown fluency, both scales provide a comprehensive portrayal of pausing, incorporating variables of frequency, duration, and location. As revealed by the statistical analysis in the present study, pause frequency distinguished solely between levels at the extremes of the scale. Consequently, this criterion is only found at the lowest profile. On the other hand, duration "long pauses" was solely found in the mid profiles whereas distribution, that is, pauses "in between" versus "at sentence/clause boundaries", was an aspect of breakdown fluency present across the scale.

Beyond the oral EPT context, the results in this study can also help other local oral tests in writing detailed and accurate descriptors for their rating scales.

Also of key importance are the insights this research provides for administrators and instructors of the English Pronunciation and Oral Fluency courses regarding their approach to designing and sequencing lessons based on the criteria that yielded a better perception of fluency and set proficiency levels apart. All of this with the caveat that these conclusions are based on raters' judgments and performance in monologic tasks. Consequently, certain criteria may still be misinterpreted as flaws, instead of choices of personal style stemming from the speaker's own perception of effective communication built through their interactional experiences across different contexts.

LIMITATIONS

In terms of data analysis, the manual coding of the variable of frequency of mid-clause-pauses was not verified by a second member in order to control for the reliability of the annotations. Furthermore, the data analysis of this variable yielded results that were not as reliable as the ones observed in other research given the small sample of responses annotated. Lastly, considering the mixed-results in fluency research regarding the contributions of repair measurements, this study could have benefitted from analyzing the use of repair strategies among examinees at all levels to confirm whether its relevancy is limited to low proficiency. So far, only the low and the mid-low profiles allude to this variable. This assumption is worth exploring, considering some disfluencies are natural too in effective communication.

FUTURE RESEARCH

While the present study contributes to supporting the validity of the test, the sole analysis of utterance fluency measures based on raters' perception still falls short when dealing with a complex construct such as fluency. It is advised to complement these results by studying the relationship between content and language processing (i.e. cognitive fluency), utterance fluency, and speaking performance. This approach will shed light on further aspects that are deemed important in effective communication such as lexical and grammatical knowledge, the speed with which examinees can access these linguistic resources, as well as how they may be reflected in utterance fluency measures. This additional perspective can, for instance, contribute to the distinction between examinees at adjacent levels.

Furthermore, it will be valuable to draw on qualitative methods to explore raters' interpretation of the scale through their comments and analyze potential biases based on traits such as L1 background, and teaching or rating experience, particularly now with the introduction of a new scale. Likewise, retrospective comments from test takers about their performance can provide researchers with information on the reasons they had for pausing or using certain repair strategies (Khang, 2020).

In the same vein, some have suggested looking into research in other fields in order to get a more comprehensive view of the construct of fluency. As explained by de Jong (2018), studies from the field of psycholinguistics regarding fluency in L1 lend support to occurrences of pauses and hesitations in upper levels of proficiency given this behavior is also observed among L1 speakers. On the other hand, findings from the field of discourse analysis on the use of repair strategies portrayed them as interaction strategies rather than deficiencies, inviting researchers to rethink the purpose of this measure in L2 speech to avoid unfair penalizations in scoring.

CHAPTER 6: CONCLUSION

This study has contributed to the growing literature in language testing concerned with defining the construct of fluency and its influence on the assessment of language proficiency. Through the statistical analysis of examinees' performance in one of the oral EPT tasks, it was determined, as suggested in the literature, that speed measurements correlated more strongly with raters' scoring compared to breakdown fluency measurements. Furthermore, a trend was observed in which the L1 profile consistently ranked higher than its subsequent higher proficiency profile (M2 profile). These results validate the administrators' decision to substitute the L1 profile for a Mid-low profile in the new scale to provide a more accurate description of adjacent levels in which fluency can be a discriminating factor in course recommendation. Lastly, it was found that, with the exception of speech rate, utterance fluency measures could only distinguish examinees at the extremes of the scale. This favorably suggests raters resort to other criteria to make their judgments.

Considering this study only examined utterance fluency measures, future research should look into other aspects of fluency. For instance, criteria tapping into learners' language processing mechanisms through the analysis of lexico-grammatical choices and speed of retrieval. Likewise, future studies can benefit from including the perspective of raters to unveil potential factors influencing their scoring, and on that of examinees to gain further insights on their performance. When examining a multidimensional construct such as fluency, an approach drawing on diverse angles is called for.

REFERENCES

Boersma, P., & Weenink, D. (2023). Praat: doing phonetics by computer [Computer program]. Version 6.4. http://www.praat.org/

Bosker, H. R., Pinget, A., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*(2), 159–175. doi:10.1177/0265532212455394

Bosker, H. R., Quené, H., Sanders, T., & Jong, N. H. (2014). The perception of fluency in native and non-native speech. *Language Learning, 64*(3), 579–614. doi:10.1111/lang.2014.64.issue-3

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

Dörnyei, Z., & Kormos, J. (1998). Problem-solving mechanisms in L2 communication: A psycholinguistic perspective. *Studies in Second Language Acquisition, 20*(3), 349–385. https://doi.org/10.1017/S0272263198003039

Dimova, S., Yan, X., & Ginther, A. (2020). *Local language testing: Design, Implementation, and Development.* Routledge.

de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics, 36*(2), 223–243. doi:10.1017/S0142716413000210

de Jong, N. H. (2016). Fluency in second language assessment. In D. Tsagari & J. Banerjee (Eds.), Handbook of second language assessment (pp. 203–218). Boston/Berlin, Massachusetts/Germany: Mouton de Gruyter.

de Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly, 15*(3), 237–254. https://doi.org/10.1080/15434303.2018.1477780

de Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. Paper presented at the 6th Workshop on Disfluency in Spontaneous Speech, Stockholm.

de Jong, N.H., Pacilly, J., & Heeren, W. (2021). Praat scripts to measure speed fluency and breakdown fluency in speech automatically, *Assessment in Education: Principles, Policy & Practice*, 28:4, 456-476, DOI: 10.1080/0969594X.2021.1951162

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379–399. https://doi.org/10.1177/0265532210364407

Götz, S. (2013). Fluency in native and non-native English speech. In *Studies in corpus linguistics*. https://doi.org/10.1075/scl.53

Hilton, H. E. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal*, *36*(2), 153–166. https://doi.org/10.1080/09571730802389983

Hunter, A.-M. (2017). Fluency development in the ESL classroom: The impact of immediate task repetition and procedural repetition on learners' oral fluency [Unpublised doctoral dissertation]. University of Surrey, Guildford.

IBM Corp.(2020). *IBM SPSS Statistics for Windows, Version 27.0 I* [Computer Software]. *Armonk*, NY: IBM Corp

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: temporal measures and stimulated recall. *Language Learning*, *64*(4), 809–854. https://doi.org/10.1111/lang.12084

Kahng, J. (2020). Explaining second language utterance fluency: Contribution of cognitive fluency and first language utterance fluency. *Applied Psycholinguistics*, *41*(2), 457–480. https://doi.org/10.1017/s0142716420000065

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164. https://doi.org/10.1016/j.system.2004.01.001

Kosmala, L., & Crible, L. (2021). The dual status of filled pauses: Evidence from genre, proficiency and co-occurrence. *Language and Speech*, *65*(1), 216–239. https://doi.org/10.1177/00238309211010862

Lennon, P. (1990). Investigating fluency in EFL: a quantitative approach. *Language Learning*, *40*(3), 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669

Lowe, P. (1983). The ILR Oral Interview: Origins, Applications, Pitfalls, and Implications. *Die Unterrichtspraxis / Teaching German*, *16*(2), 230–244. https://doi.org/10.2307/3530138

Moodle HQ. (2023). Moodle Version 4.3 [Learning Management System]. https://moodle.org/

Nakatsuhara, F., Nahal, K., Chihiro, I. (2021). Assessing Speaking. In Fulcher, G., & Harding, L (Eds.), *The Routledge handbook of language testing* (pp. 209-222). Routledge.

New Directions East Asia. (2023, March 5). *Investigating repair fluency across tasks and levels of proficiency in the APTIS Teens Speaking test* [Video]. YouTube. https://www.youtube.com/watch?v=axUk9KDMelg

Liang, S., & Fu, Y. (2016). Otter.ai. [Online Software].https://otter.ai/

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, *33*(1), 53–73. https://doi.org/10.1177/0265532215579530

Tavakoli, P. (2011). Pausing patterns: differences between L2 learners and native speakers. *ELT Journal*, *65*(1), 71–79. https://doi.org/10.1093/elt/ccq020

Tavakoli, P., & Wright, C. (2020). *Second language speech fluency*. Cambridge University Press.

The Audacity Team. (2022). *Audacity* (Version 3.2.2). [Computer Software]. Audio Team Website. https://www.audacityteam.org/download/

Tottie G. (2016). Planning what to say: Uh and um among the pragmatic markers. In Kaltenböck G., Keizer E., Lohmann A. (Ed.) *Outside the clause. Form and function of extra-clausal constituents* (pp. 97–122). John Benjamins.

Segalowitz, N. (2010). *Cognitive bases of second language fluency.* New York, NY: Routledge.

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching*, *54*(2). https://doi.org/10.1515/iral-2016-9991

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36*(1), 1–14.
https://doi.org/10.1017/S026144480200188X

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category - ELAN and ISO DCR. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Suzuki, S., & Kormos, J. (2023). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*, *45*(1), 38–64. https://doi.org/10.1017/s0272263121000899

Yan, X., Kim, H. R., & Kim, J. Y. (2020). Dimensionality of speech fluency: Examining the relationships among complexity, accuracy, and fluency (CAF) features of speaking performances on the Aptis test. *Language Testing*, *38*(4), 485–510. https://doi.org/10.1177/0265532220951508

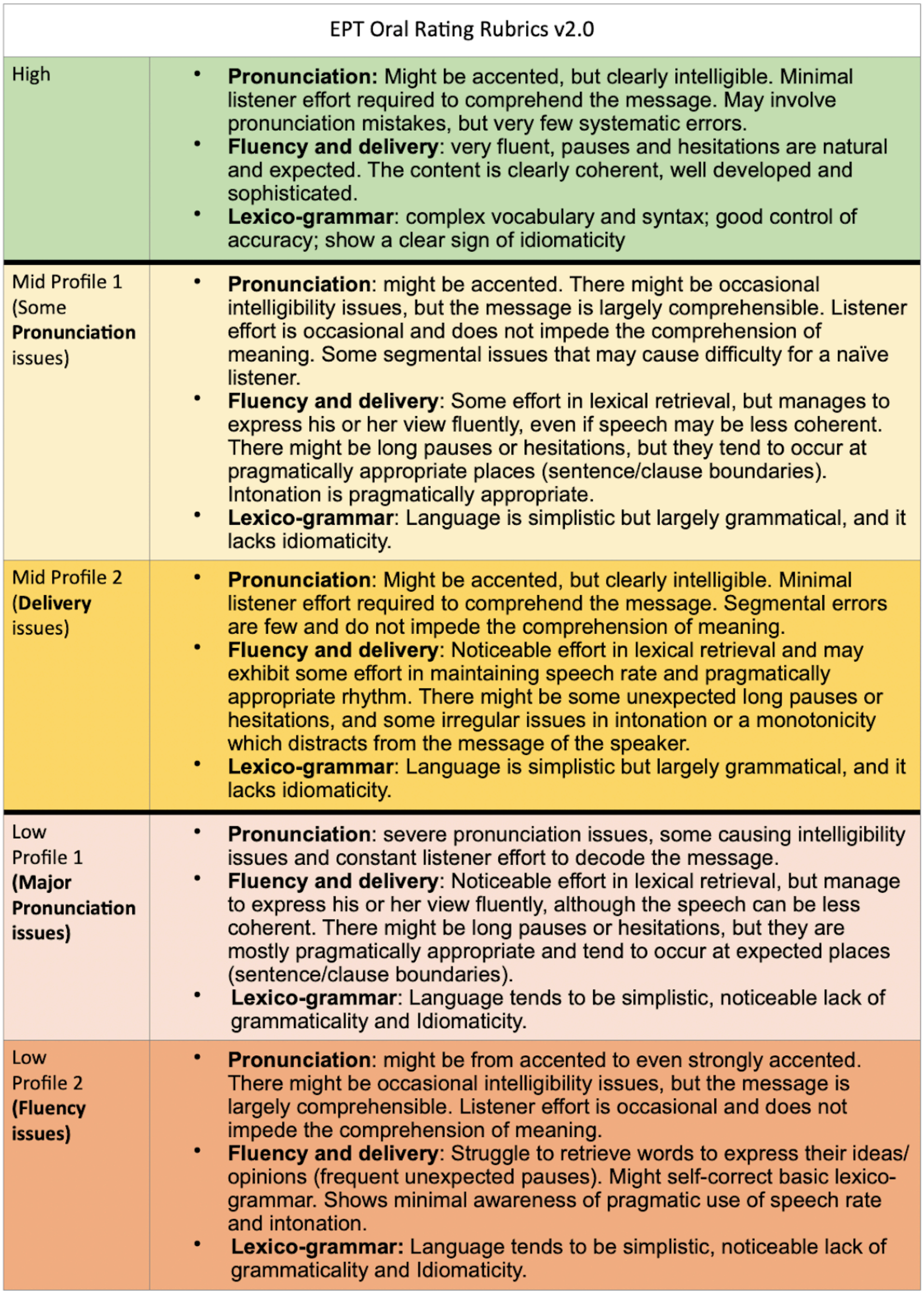
Wang, Z. (2014). Online time pressure manipulations: L2 speaking performance under five types of planning and repetition conditions. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 27–61). John Benjamins Publishing Company. https://doi.org/10.1075/tblt.5.02wan

Weir, C., & Milanovic, M., (2003). *Continuity and innovation: Revising the Cambridge Proficiency in English examination 1913-2002.* Cambridge University Press.

Witton-Davies, G. (2014). The study of fluency and its development in monologue and dialogue.

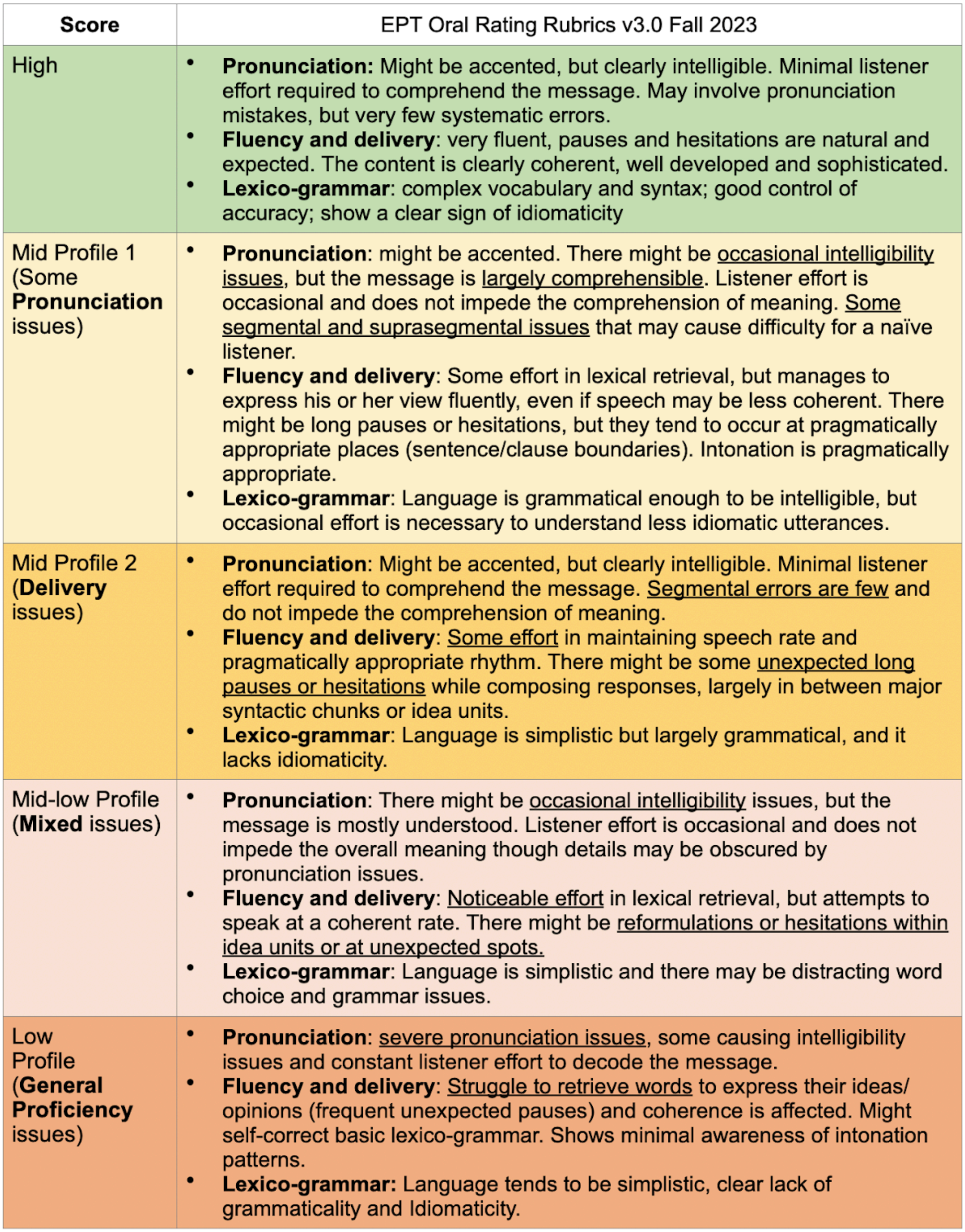
[Unpublished doctoral thesis]. University of Lancaster.

APPENDIX A: EPT ORAL RATING RUBRIC VERSION 2.0



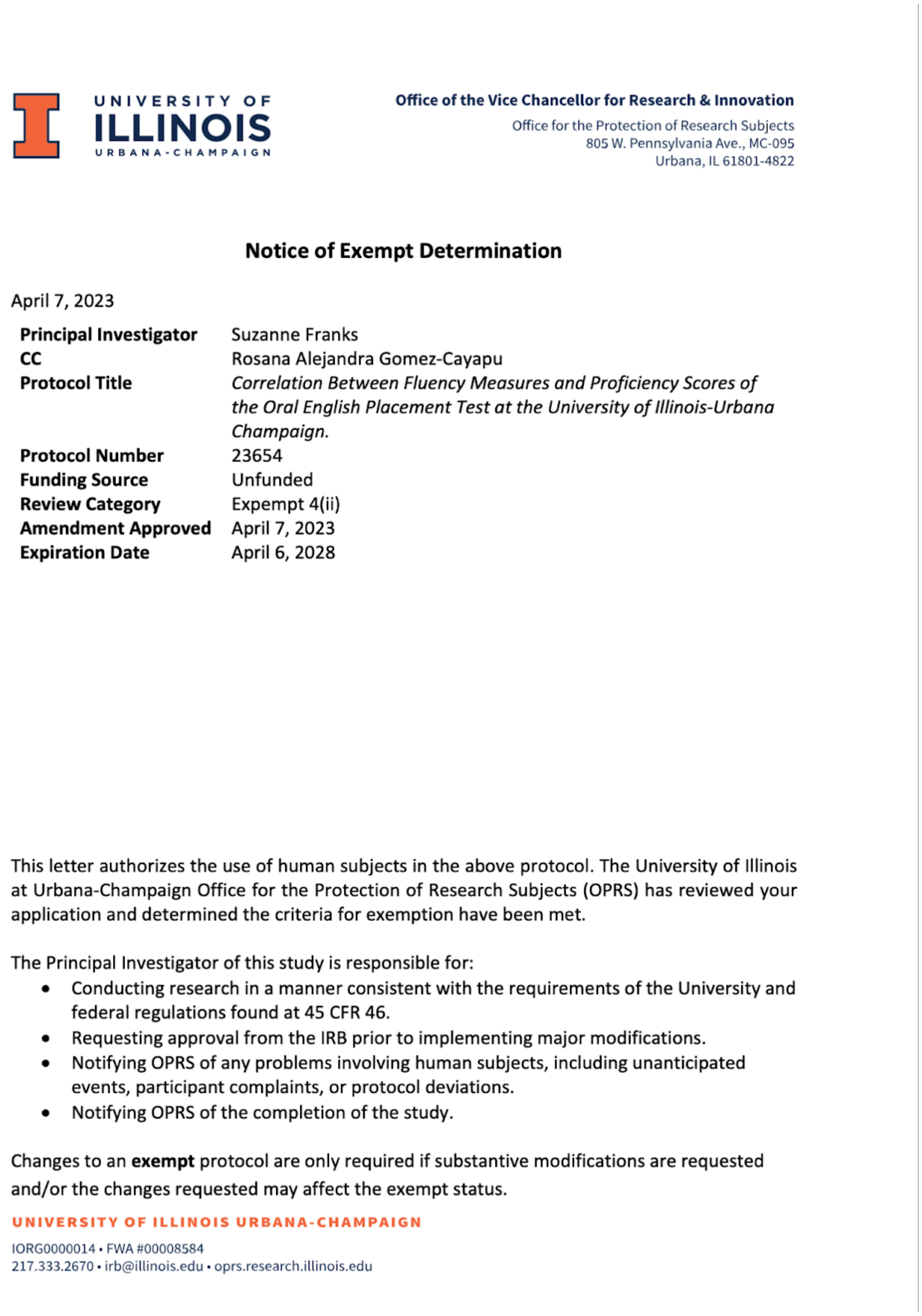
EPT Oral Rating Rubrics v2.0	
High	• **Pronunciation:** Might be accented, but clearly intelligible. Minimal listener effort required to comprehend the message. May involve pronunciation mistakes, but very few systematic errors. • **Fluency and delivery**: very fluent, pauses and hesitations are natural and expected. The content is clearly coherent, well developed and sophisticated. • **Lexico-grammar**: complex vocabulary and syntax; good control of accuracy; show a clear sign of idiomaticity
Mid Profile 1 (Some **Pronunciation** issues)	• **Pronunciation**: might be accented. There might be occasional intelligibility issues, but the message is largely comprehensible. Listener effort is occasional and does not impede the comprehension of meaning. Some segmental issues that may cause difficulty for a naïve listener. • **Fluency and delivery**: Some effort in lexical retrieval, but manages to express his or her view fluently, even if speech may be less coherent. There might be long pauses or hesitations, but they tend to occur at pragmatically appropriate places (sentence/clause boundaries). Intonation is pragmatically appropriate. • **Lexico-grammar**: Language is simplistic but largely grammatical, and it lacks idiomaticity.
Mid Profile 2 (**Delivery** issues)	• **Pronunciation**: Might be accented, but clearly intelligible. Minimal listener effort required to comprehend the message. Segmental errors are few and do not impede the comprehension of meaning. • **Fluency and delivery**: Noticeable effort in lexical retrieval and may exhibit some effort in maintaining speech rate and pragmatically appropriate rhythm. There might be some unexpected long pauses or hesitations, and some irregular issues in intonation or a monotonicity which distracts from the message of the speaker. • **Lexico-grammar**: Language is simplistic but largely grammatical, and it lacks idiomaticity.
Low Profile 1 **(Major Pronunciation issues)**	• **Pronunciation**: severe pronunciation issues, some causing intelligibility issues and constant listener effort to decode the message. • **Fluency and delivery**: Noticeable effort in lexical retrieval, but manage to express his or her view fluently, although the speech can be less coherent. There might be long pauses or hesitations, but they are mostly pragmatically appropriate and tend to occur at expected places (sentence/clause boundaries). • **Lexico-grammar**: Language tends to be simplistic, noticeable lack of grammaticality and Idiomaticity.
Low Profile 2 **(Fluency issues)**	• **Pronunciation**: might be from accented to even strongly accented. There might be occasional intelligibility issues, but the message is largely comprehensible. Listener effort is occasional and does not impede the comprehension of meaning. • **Fluency and delivery**: Struggle to retrieve words to express their ideas/opinions (frequent unexpected pauses). Might self-correct basic lexico-grammar. Shows minimal awareness of pragmatic use of speech rate and intonation. • **Lexico-grammar:** Language tends to be simplistic, noticeable lack of grammaticality and Idiomaticity.

APPENDIX B: EPT ORAL RATING RUBRIC VERSION 3.0



Score	EPT Oral Rating Rubrics v3.0 Fall 2023
High	• **Pronunciation:** Might be accented, but clearly intelligible. Minimal listener effort required to comprehend the message. May involve pronunciation mistakes, but very few systematic errors. • **Fluency and delivery**: very fluent, pauses and hesitations are natural and expected. The content is clearly coherent, well developed and sophisticated. • **Lexico-grammar**: complex vocabulary and syntax; good control of accuracy; show a clear sign of idiomaticity
Mid Profile 1 (Some **Pronunciation** issues)	• **Pronunciation**: might be accented. There might be occasional intelligibility issues, but the message is largely comprehensible. Listener effort is occasional and does not impede the comprehension of meaning. Some segmental and suprasegmental issues that may cause difficulty for a naïve listener. • **Fluency and delivery**: Some effort in lexical retrieval, but manages to express his or her view fluently, even if speech may be less coherent. There might be long pauses or hesitations, but they tend to occur at pragmatically appropriate places (sentence/clause boundaries). Intonation is pragmatically appropriate. • **Lexico-grammar**: Language is grammatical enough to be intelligible, but occasional effort is necessary to understand less idiomatic utterances.
Mid Profile 2 (**Delivery** issues)	• **Pronunciation**: Might be accented, but clearly intelligible. Minimal listener effort required to comprehend the message. Segmental errors are few and do not impede the comprehension of meaning. • **Fluency and delivery**: Some effort in maintaining speech rate and pragmatically appropriate rhythm. There might be some unexpected long pauses or hesitations while composing responses, largely in between major syntactic chunks or idea units. • **Lexico-grammar**: Language is simplistic but largely grammatical, and it lacks idiomaticity.
Mid-low Profile (**Mixed** issues)	• **Pronunciation**: There might be occasional intelligibility issues, but the message is mostly understood. Listener effort is occasional and does not impede the overall meaning though details may be obscured by pronunciation issues. • **Fluency and delivery**: Noticeable effort in lexical retrieval, but attempts to speak at a coherent rate. There might be reformulations or hesitations within idea units or at unexpected spots. • **Lexico-grammar**: Language is simplistic and there may be distracting word choice and grammar issues.
Low Profile (**General Proficiency** issues)	• **Pronunciation**: severe pronunciation issues, some causing intelligibility issues and constant listener effort to decode the message. • **Fluency and delivery**: Struggle to retrieve words to express their ideas/ opinions (frequent unexpected pauses) and coherence is affected. Might self-correct basic lexico-grammar. Shows minimal awareness of intonation patterns. • **Lexico-grammar:** Language tends to be simplistic, clear lack of grammaticality and Idiomaticity.

APPENDIX C: IRB APPROVAL



UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Office of the Vice Chancellor for Research & Innovation
Office for the Protection of Research Subjects
805 W. Pennsylvania Ave., MC-095
Urbana, IL 61801-4822

Notice of Exempt Determination

April 7, 2023

Principal Investigator	Suzanne Franks
CC	Rosana Alejandra Gomez-Cayapu
Protocol Title	*Correlation Between Fluency Measures and Proficiency Scores of the Oral English Placement Test at the University of Illinois-Urbana Champaign.*
Protocol Number	23654
Funding Source	Unfunded
Review Category	Expempt 4(ii)
Amendment Approved	April 7, 2023
Expiration Date	April 6, 2028

This letter authorizes the use of human subjects in the above protocol. The University of Illinois at Urbana-Champaign Office for the Protection of Research Subjects (OPRS) has reviewed your application and determined the criteria for exemption have been met.

The Principal Investigator of this study is responsible for:

- Conducting research in a manner consistent with the requirements of the University and federal regulations found at 45 CFR 46.
- Requesting approval from the IRB prior to implementing major modifications.
- Notifying OPRS of any problems involving human subjects, including unanticipated events, participant complaints, or protocol deviations.
- Notifying OPRS of the completion of the study.

Changes to an **exempt** protocol are only required if substantive modifications are requested and/or the changes requested may affect the exempt status.