A Critical Analysis of Distinguishing Features of Young English Language Learners' Oral Performance

Benjamin Sanchez Murillo

Temple University

Japan Campus

Tokyo Center

EPSY 8625 Introduction to Research Methodology

Dr. James Sick

March 11, 2021

A Critical Analysis of Distinguishing Features of Young English Language Learners' Oral Performance

In the article "Distinguishing Features of Young English Language Learners' Oral performance," Lin Gu and Ching-Ni Hsieh (2019) examined the distinguishing features of spoken proficiency of young English learners as it develops. Gu and Hsieh (2019) were interested in learning if the speaking proficiency features found to differentiate adult learners were also present in young learners.

Gu and Hsieh (2019) used 57 TOEFL (Test of English as a Foreign Language) Junior Speaking test takers' spoken samples from young children between 9 and 12 years old. The TOEFL Junior Speaking test is a test developed by Educational Testing Service (ETS) designed for children aged 11 and above. The samples the researchers used were between July 2012 and December 2013 and consisted of speech elicited from three different tasks, a picture narration task, a non-academic listen-speak task, and an academic listen-speak task.

Gu and Hsieh (2019) scored the speaking tests, organized the results according to score and age, and assigned them into score bands. The researchers examined four speaking proficiency categories: fluency, vocabulary, grammar, content, and their 19 corresponding speaking proficiency features in adult speech and calculated the mean and standard deviation of each feature by score band level, resulting in four bands. Furthermore, Gu and Hsieh (2019) obtained a separate one-way analysis of variance (ANOVA) analysis for each feature and calculated the bands' effect size using Cohen's *d*. Gu and Hsieh (2019) found that two features, speech rate (fluency) and proportion of critical points covered (content), were "sensitive to the growth of speaking proficiency" (p. 190).

This critique aims to point to some improvements that the researchers could have implemented in their literature review and their methodological approach and identify some of the study's strengths. In other words, the literature review could have used more recent studies, and the methodology approach could have used better sampling techniques and a better statistical method to measure inter-rater agreement. The study could have also more accurately reported data. Nevertheless, the study does expand the current knowledge of speaking proficiency in children.

Literature Review

The literature review in Gu and Hsieh (2019) shed light on the different aspects of English language learners' spoken proficiency by providing examples of these features within six empirical studies ranging from 2004 to 2011. However, the researchers' use of old empirical studies to classify spoken proficiency features could have used more recent articles to represent current findings appropriately. While proficiency changes would probably not change significantly in 15 years, it would still be essential to learn newer changes, if any, as it might better inform readers who would like to expand on current knowledge. It would have also been beneficial for the reader to know the justification for using these older studies as I was able to find more recent studies between 2011 and 2018 regarding adult learners and spoken proficiency features.

Nevertheless, besides the questionable date range in their choice of supporting empirical studies, the literature review was generally well written. To help the reader appreciate a holistic view of the studies, Gu and Hsieh (2019) summarized the categories and features with the corresponding studies in a table. In addition to a visual summary, Gu and Hsieh (2019) also provided a textual overview of the reviewed articles. In their summary, the researchers stated that in addition to identifying the spoken performance categories and features, the review articles also showed that some of the features are better used to describe speaking proficiency at all levels. In contrast, other features are unique to an adjacent level of speaking proficiency. Gu and Hsieh (2019) further explained that the reason for this phenomenon is that raters use different criteria when making scoring decisions, and those decisions depend on how raters associate the performance level with a set of features. The combination of a visual and textual overview in their literature review offers readers a brief synopsis of prior work that makes the remaining work easy to follow and seamlessly transitions into the research gaps and their research questions.

The empirical studies revealed two research gaps. First, Gu and Hsieh (2019) believed that it was necessary to differentiate which features are more prevalent in a particular proficiency level to understand better how speaking proficiency develops in young children. Second, the previous studies focused on adult learners. These research gaps reveal the researchers' observations of the differences between adult and children's speaking proficiency. The researchers state that adult learners' cognitive maturity, learning contexts, and communicative needs do not necessarily generalize to young children.

These gaps in the research introduced a couple of straightforward research questions. The two research questions were the following: First, "What are the linguistic features that differentiate young English learners across proficiency levels?" (p. 183). Second, "To what extent is the spoken performance of young English language learners characterized by different features at different proficiency levels?" (p. 183). The research questions on linguistic features and spoken proficiency did reflect the literature presented, so it was easy to follow how the researchers' literature review related to their research questions.

Methodological Approach

Gu and Hsieh (2019) conducted basic research with an exploratory study design using a quantitative approach. The aim of exploratory research is not to find conclusive answers to research questions but rather to better understand the problem. Using an exploratory approach to understanding young English learners' linguistic features is a good approach in this study as it gives researchers opportunities to generate discoveries for further learning.

In their aim to measure the speaking proficiency features, the researchers relied on percent agreement to rate performance samples, one-way ANOVA analysis to compare the means of the samples, and Cohen's *d* to determine the effect size between two means, in this case, the means between adjacent score band levels. Gu and Hsieh's (2019) used SpeechRater, a tool developed by ETS, to compute different conceptual category features such as productivity, pause phenomena, and automaticity in fluency. They used VocabProfile to analyze lexical features for vocabulary breadth, vocabulary sophistication, and percent of AWL (Academic Word List) in vocabulary. Finally, Gu and Hsieh's (2019) used human coding to analyze grammatical complexity, grammatical accuracy, and content coverage for content.

Non-probability Sampling

Gu and Hsieh (2019) selected the test samples from July 2012 to December 2013 from an earlier, more extensive study. The researchers mention that a high percentage of these test-takers during that period were Korean L1 speakers, and the researchers opted to base their current research on Korean L1 speakers. The study, however, did not elaborate on the selection method of the test samples. Therefore, the researchers' sampling choice was most likely nonprobability sampling. If so, the type of nonprobability sampling that Gu and Hsieh (2019) used can be classified as convenience sampling because the samples were already available from an earlier study. However, the use of convenience sampling can limit the generalizability of results.

In studies using convenience sampling, researchers use readily available data or individuals to explore a topic further but with a disadvantage. According to Paltridge and Phakiti (2015), convenience sampling can provide informative results; however, one of its drawbacks is that the results cannot be generalized to a larger population. Gu and Hsieh (2019) concede that including only one native language limited their study results' generalizability; however, they do not mention that nonprobability sampling was another factor limiting their study. The researchers' focus on native language as a limiting factor for generalizability gives the impression that L1 was the main factor when sampling issues could have also played a role.

Inter-rater Reliability

The researchers used different measurement methods to analyze the features of fluency, vocabulary and grammar used automatization tools (i.e., SpeechRater and VocabProfile) and human coding. Gu and Hsieh (2019) used these tools to analyze fluency and vocabulary and used human coding to analyze grammar and content features. While I did not see any issues using SpeechRater and VocabProfile to analyze fluency and vocabulary in this study, the use of human coding to rate grammar and content coverage brings inter-rater reliability concerns for researchers wishing to replicate similar results. For example, Gu and Hsieh (2019) relied on percent agreement to analyze their convenience samples' grammar and content features. The researchers used two raters to code these features, and the raters agreed that the grammar features ranged from 83% to 91% and 91% for content.

However, one problem with percent agreement is raters' potential to arrive at similar answers by chance. Arriving at similar answers by chance can affect the reliability of the measures as it brings into question whether raters provided reasonably consistent ratings. Gu and Hsieh's (2019) research used ratings to explore the extent of young learners' spoken performance; however, it would have been helpful if researchers had included a statistical analysis to account for chance. To account for this possibility, the researchers could have included Cohen's *kappa coefficient (k)* to support their findings. According to Brown (2016), Cohen's kappa coefficient (k) accounts for chance agreement by "providing an estimate of the proportion of agreement in classifications beyond what would be expected to occur by chance alone" (p. 142). Gu and Hsieh possibly did not intend to use a more accurate statistical analysis in an exploratory study. However, the fact that two of the four categories used percent agreement can weaken the study's reliability.

Misreported Data

One of the reasons researchers use statistical analysis in their work is to help researchers and readers interpret data results. One could argue that misreported data results run the risk of affecting the validity of a study. Brown (2016) defined the validity of a study as "the degree to which the results of a study represent what the researcher thinks they represent" (p. 163). One such example of misreported data that can affect the validity of Gu and Hsieh's study is the inconsistent eta squared (η^2) values that the researchers reported in the vocabulary result section and the summary of ANOVA and pairwise comparisons table.

Brown (2016) defines eta squared (η^2) as "the proportion of variance associated with or accounted for by each of the mean effects, interactions, and error in an ANOVA study" (p. 205). The reported values were $\eta^2 = .22$ and $\eta^2 = .003$. Gu and Hsieh (2019) denoted the η^2 value in the table with three asterisks to show the significance of the percentage of words from the most frequent thousand-word families of English (Percent 1K) for the vocabulary sophistication feature. However, this reporting inconsistency leaves the reader unable to assess the vocabulary category results' true variance significance. One can either interpret the Percent 1K results as accounting for either 0.3% or 22% of the speaking proficiency variance.

Strengths of the Study

Out of the 19 adult features found in adult speech, Gu and Hsieh (2019) found 17 to be associated with young learners. Each of the features differentiated when compared against the different score bands indicating "how the performance on individual features changed as the speaking ability of the young learners progressed" (p. 190).

It was interesting to find that speech rate and proportion of key points covered features had the most significant effect size. These features also differentiated speakers across all the score band levels. This finding suggests that these features are a "strong predictor of speaking proficiency for young learners" (p. 190). Other features did not differentiate across all levels but instead exhibited different effect size power at certain levels. For example, a couple of pause features in the fluency category (i.e., silence per word and mean silence duration) were significant for the learners with the lowest scores. Gu and Hsieh (2019) also stated that other features (i.e., clauses per AS unit (a single speaker's utterance), silence per word, phonation-time ratio, mean length of chunk, number of error-free AS unit, proportion of error-free AS units, word types, number of words) also differed depending on their English proficiency level.

In sum, Gu and Hsieh's (2019) findings open the door for those who wish to explore English proficiency further in children as they develop. Not only can their results guide future researchers who want to build rating criteria to assess young learners' English proficiency, but also English teachers can use this study as a teaching guide. For example, knowing that young learners exhibit longer pauses and silence durations during speaking performances than older children will help English teachers adjust their lesson plans accordingly.

Conclusion

Gu and Hsieh's (2019) exploratory study offers the reader a detailed perspective as to the distinguishing features of spoken performances among adults and uses these adult features to identify those that are specific to children. They found 17 features associated with young English learners showing high levels of strength of association. They found some features that differentiated across band scores, indicating that some features are more associated with a particular band score level. Overall, Gu and Hsieh could have improved the reliability, validity, and generalizability of their research by providing probability sampling, using Cohen's kappa coefficient for accounting for chance when rating features, and reporting consistent data. However, the study adds to the current lack of research on young learners, which future researchers can use as a stepping-stone for further discoveries.

References

Brown, J.D. (2016). *Statistics Corner: Questions and Answers about Language Testing Statistics*. Tokyo: JALT Testing and Evaluation Special Interest Group.

Gu, L., & Hsieh, C. (2019). Distinguishing features of young English language learner's oral performance. *Language Assessment Quarterly*, 16(2), 180–195. https://doi.org/10.1080/15434303.2019.1605518

Paltridge, B. & Phakiti, A. (2015). *Research Methods in Applied Linguistics: A Practical Resource (Research Methods in Linguistics)*, 2nd edition. New York, New York: Bloomsbury USA Academic.