The Validity of Speaking Measures:

Relative Impact of Assessment Methods on Learners' Production of Grammatical Complexity

Alyssa A. D. Kermad

Northern Arizona University

Abstract

Major standardized testing companies and Intensive English Programs consistently rate learners' spoken output on grammatical aspects, including grammatical complexity. However, based on the communicative purpose of a task, differences can occur with what learners actually produce and what raters pay attention to. In the present study, previous research on grammatical complexity is referenced in order to explain register differences related to spoken task-type variation. Specifically, two sets of linguistic analyses were carried out to analyze the differences in grammatical complexity across four tasks: spontaneous speech tasks, picture-description tasks, elicitation tasks, and read-aloud tasks. The quantitative analysis investigated the use of sixteen grammatical features associated with grammatical complexity, and the descriptive analysis placed texts into the hypothesized developmental stages of grammatical complexity. Although there were no statistically significant differences in task-type variation from the quantitative analyses, there were large-sized effects of mean differences between the spontaneous speech and picture-description tasks. Descriptive analyses further confirmed these differences as all four texts corresponded with different developmental stages of grammatical complexity. The implications of the results are discussed in light of assessment purposes and how they suggest a need for more clearly defined rubrics measuring grammatical complexity.

Keywords: task-variation, task-design, grammatical complexity, speaking assessment

The Validity of Speaking Measures:

Relative Impact of Assessment Methods on Learners' Production of Grammatical Complexity

Background

Across sub fields of applied linguistics, the consideration of tasks for language teaching, learning/acquisition, assessment, and research has been considered important and necessary; however, with such a variety of tasks available to measure diverse constructs related to language learning, it cannot be said that all tasks are equal in their cognitive complexity (Biber, Gray, and Staples, 2014). Specifically in the field of second language (L2) speaking and pronunciation, a multiplicity of speaking tasks have been used in research and/or in the classroom based on popularity or convenience of the task.

Major standardized testing companies consistently rate speakers' proficiency level on grammatical aspects: on the IELTS test, the highest band level for "Grammatical range and accuracy" includes descriptors of the speaker's use of "a full range of structures." Similarly, the highest band of the TOEFL iBT test under "Language Use" includes a description of the speaker's "good control of basic and complex grammatical structures that allow for coherent, efficient (automatic) expression of relative ideas." Even in Intensive English Programs across the nation, speaking rubrics often include descriptors of the speaker's grammatical production (complexity and/or accuracy).

Despite the fact that many speaking rubrics include ratings of L2 speakers' production of grammatical complexity, previous studies in the field of L2 speaking and pronunciation have inadequately established how different speaking tasks require different degrees of linguistic variation in terms of grammatical complexity. Based on the goal of the task, speakers are required to frame their spoken production in a way which satisfies the demands of the task.

However, speaking rubrics often include vague descriptors of grammatical complexity which are oftentimes not applicable to the language being generated, or which are otherwise too vague to be pertinent. To this end, the present study investigates the production of grammatical complexity in four of the most commonly used speaking and pronunciation tasks (Thomson & Derwing, 2014): read-aloud tasks, spontaneous speech tasks, elicitation tasks, and picture-description tasks.

Research Questions

The present study seeks to respond to the following research questions: (1) In what ways does the nature of a speaking task affect spoken features of grammatical complexity? And (2) To what extent does the nature of a speaking task reflect the developmental stages of grammatical complexity?

Methods

Speakers

Fifteen participants were recruited from the Program in Intensive English (PIE). Five participants were enrolled in Level 3, four participants were enrolled in Level 4, and six participants were enrolled in Level 5.

Speaking Tasks

The study incorporated the four most commonly used assessment methods in the field of L2 pronunciation (see Thomson & Derwing, 2014): read-aloud tasks (Celce-Murcia, Goodwin, & Brinton, 2010), spontaneous speech tasks, elicitation tasks (Trofimovich, Lightbown, Halter, & Song, 2009), and picture-description tasks (Derwing, Munro, Thomson, & Rossiter, 2009).

Each participant produced a total of 4 spoken texts each, however, only 2 out of the 4 texts were original production, being that all participants recorded the same read-aloud and

elicitation text. There was a total number of 30 novel spoken texts acquired from the participants: 15 from the spontaneous speech task and 15 from the picture-description task.

Linguistic Analyses

Once transcribed, all of the spontaneous and picture-description texts were tagged using the Biber tagger which counted and normalized linguistic features to a rate per 1,000 words of text. This ensured that quantitative measures could be comparable across texts, despite the actual length of the text.

The present study focused on the 23 linguistic features motivated by prior empirical research that have been associated with grammatical complexity (see Biber *et al.*, 2014). These features included grammatical classes, some dependent clauses, and phrasal structures of modification and elaboration.

Quantitative analyses. In order to compare mean differences of features of grammatical complexity for the spontaneous and picture description tasks, five independent *t*-tests were run for 5 of the linguistic features which met the assumptions of normality and homogeneity of variance. For the remaining 11 features, independent tests for non-parametric data were conducted through the Mann-Whitney *U* test. It should be noted that 7/23 features were not included in the quantitative analysis, as there were no occurrences of these features in either of the texts.

Descriptive analyses. Analyses of the elicitation and read-aloud texts were conducted descriptively. These tasks were analyzed in terms of how well they reflected the developmental stages of grammatical complexity (see Biber, Gray, & Poonpon, 2011). They were not included in the main statistical analysis being that there was no inter-speaker variation for these tasks.

All results of the quantitative linguistic analysis were also described descriptively following the statistical analyses. This was to investigate how the different tasks corresponded with the developmental stages described in Biber *et al*. (2011).

Results

Quantitative Features of Grammatical Complexity

Of the sixteen features of grammatical complexity analyzed in this study, five independent samples *t*-tests were conducted for normally distributed data and eleven Mann-Whitney tests were conducted for non-normally distributed data.

Five independent samples *t*-tests were conducted for normally distributed data of the two independent text groups: the spontaneous speech texts and the picture-description texts. The five dependent variables were (1) word length, (2) adverbs, (3) nouns, (4) prepositional phrases, and (5) attributive adjectives.

Table 1

Independent Samples t-test of Grammatical Complexity between Tasks

	Picture Description		Spontaneous Speech					
Grammatical Feature	*M*	*SD*	*M*	*SD*	*t*(28)	*p*	95% CI	η^2
Word Length	3.61	0.25	3.90	0.28	-2.97	.006	[-00.48, -00.09]	.24
Adverbs	42.67	14.18	51.08	21.34	-1.27	.214	[-21.97, 05.14]	.16
Nouns	191.53	41.80	236.29	43.82	-2.86	.008	[-76.79, -12.73]	.23
Prepositional Phrases	61.87	24.51	85.31	23.95	-2.65	.013	[-41.57, -05.33]	.20
Attributive Adjectives	17.69	13.05	32.35	18.16	-2.54	.017	[-26.49, -02.83]	.19

*$p < .003$

The results shown in Table 1 indicate that although not significant at the Bonferroni adjusted alpha level, $\alpha = .003$, on average, the spontaneous speech task generated longer words

(*M* = 3.90) than the picture description task (*M* =3.61), more nouns (*M* = 236.29) than the picture description task (*M* =191.53), more prepositional phrases (*M* = 85.31) than the picture description task (*M* =61.87), and more attributive adjectives (*M* = 32.35) than the picture description task (*M* =17.69). All differences had large-sized effects.

Eleven Mann-Whitney tests were conducted for non-normally distributed data of the two independent text groups: the spontaneous speech texts and the picture-description texts. The eleven dependent variables were (1) passives, (2) clausal connectors, (3) *of* genitives, (4) pre-modifying nouns, (5) *wh*- complement clauses, (6) finite adverbial clauses, (7) verb + *that* clauses, (8) verb + *to* clauses, (9) desire verb + *to* clauses, (10) verb + -*ing* clauses, and (11) finite relative clauses.

Table 2

Mann Whitney U tests of Grammatical Complexity between Tasks

	Picture Description	Spontaneous Speech			
Grammatical Feature	*M Rank*	*M Rank*	*U*(28)	*p*	η^2
Passives	17.60	13.40	81.00	.202	.11
Clausal connector	18.50	12.50	67.50	.061	.20
Of genitive	14.73	16.27	101.00	.653	.01
Pre-modifying Nouns	14.13	16.87	92.00	.412	.03
Wh- complement clause	16.00	15.00	105.00	.775	.03
Finite adverbial clauses	14.80	16.20	102.00	.683	.01
Verb + *that*- clause	17.70	13.30	79.50	.174	.08
Verb + *to* clause	12.47	18.53	67.00	.061	.17
Desire verb + *to*-clause	13.00	18.00	75.00	.126	.17
Verb + *ing*-clause	15.87	15.13	107.00	.838	.00
Finite relative clause	17.20	13.80	87.00	.305	.07

*$p < .003$

The results displayed in Table 2 indicated no significant effect of the task at the Bonferroni adjusted alpha level, $\alpha = .003$.

Descriptive Analyses of Grammatical Complexity

The spontaneous speech task generated longer word lengths and more adverbs, nouns, prepositional phrases, verb + *to* clause, and desire verb + *to* clauses, all having large sized effects. The picture-description task, on the other hand, generated more passives, clausal connectors, verb + *that* clauses, and more finite relative clauses, all having a medium to large sized-effect. The finite *that* and *wh* complement clauses controlled by verbs were both more frequent in the picture description task (although the *wh* complement clauses had a small effect size). These are characteristic of stage 1 of the developmental sequence. The picture-description task also made use of more finite relative clauses which are more characteristic of stage 3, and passives which are more characteristic of stage 4.

The elicitation task seemed to be at a pre-stage level. All six sentences were mostly mono-clausal, with no phrasal or clausal embedding, and almost all in the present tense. They were simplistic, and most likely representative of some sort of "teacher-generated" language that would occur in a pre-stage of the developmental sequences.

The read-aloud task had features which placed it early on in the developmental stages such as with finite complement clauses controlled by verbs (stage 1), finite adverbial clauses (stage 2), prepositional phrases as post-modifiers (stage 3), or *that* relative clauses (stage 3). However, it also made use of phrasal embedding in the noun phrase through the use of attributive adjectives or noun modifiers, which was more characteristic of stage 4. With its high density of compound nouns and informational topic, the focus of the paragraph more closely approximated informational writing than conversation.

Relevance to PIE and Second Language Learning

The present study has offered suggestions eluding to the fact that grammatical complexity is associated with different linguistic structures across tasks. This comes as no surprise since the communicative purposes of the given tasks are different. The spontaneous speech task approximates conversation and is therefore more representative of clausal complexity features, whereas the read-aloud task, more associated with informational writing, produced more phrasal noun modifiers and phrasal embedding. The picture-description task, instilling greater cognitive pressure, generated more passives and clausal connectors, which can be linked to the unknown characters in the task and the needed cohesive devises. Finally, the elicitation task seemed to generate extremely simplistic structures that did not fit well into the developmental sequence of complexity. For their simplicity and decontextualized subject matter, they might not be representative of naturally occurring language.

The implications gleaned from these task-type differences can be relative to PIE in how they can inform speaking assessment rubrics. Descriptors of "grammatical complexity" can more appropriately be defined on speaking rubrics, depending on the task at hand, especially for the spontaneous and picture description task. Although further research is needed to make generalizations about this point, preliminary findings suggest that commonly used speaking tasks have different communicative purposes, and thus a different output of grammatical complexity. Further defining of these terms includes a consideration of the grammatical complexity associated with the task and its communicative purpose. By doing so, there will be a closer agreement between the spoken production of language learners on a given task and the focus of listener ratings.

References

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. *TESOL Quarterly, 45*, 5-35.

Biber, D., Gray, B., & Staples, S. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, 1-31.

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation: A course book and reference guide*. Cambridge: Cambridge University Press.

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*, 533-557.

Thomson, R. I., & Derwing, T. M. (2014). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*, 326-344.

Trofimovich, P., Lightbown, P. M., Halter, R. H., & Song, H. (2009). Comprehension-based practice. *Studies in Second Language Acquisition*, *31*, 609-639.