Speaking Proficiency of Young Language Students: A Discourse-Analytic Study

By

Ching-Ni Hsieh
Educational Testing Service
Princeton, New Jersey,
United States

And

Yuan Wang
Department of Tourism,
East China Normal University,
Shanghai,
China

ABSTRACT

This study investigated a range of fluency, grammar, vocabulary, and content features of young language students' speaking performances, using a discourse-analytic approach. In total, 179 test takers' responses to the speaking section of the TOEFL Junior® Comprehensive test were selected for analysis. Mixed-design ANOVAs were used to compare 21 spoken discourse features across four proficiency levels and two task types (i.e., a picture narration task and an integrated listen/speak task). The discourse features largely differentiated test takers across proficiency levels. Task types showed some impact on measures of grammar, vocabulary, and content, but had no influence on features of fluency. Findings of the study have implications for the language development of young second and foreign language students and provide insights into language assessment task design for this population.

KEYWORDS: Speaking Proficiency, Language, Students, Discourse-Analytic Study

Introduction

With more and younger students learning English-as-a-foreign-language (EFL) worldwide, standardized language examinations designed for young language students (YLSs) have become increasingly popular (Nikolov, 2016). In light of this fast-growing trend, the need for a better understanding of YLSs' language developmental patterns and the linguistic profiles of their language performances has never been greater. Such information is critical to provide guidance and direction for the creation of assessment tasks for young learners (Bailey & Heritage, 2014). Empirical studies that systematically examine the progression of English language proficiency (ELP) among YLSs are similarly imperative in order to validate YLS assessment tasks and scoring rubrics and provide validity evidence for the claims that are based on test results (Kane, 2013).

Components of adult second language (L2) learners' speaking proficiency have been widely researched within the contexts of rating-scale development and test validation (Brown, Iwashita, & McNamara, 2005; Frost, Elder, & Wigglesworth, 2011). Many standardized YLS oral performance assessments have been developed that are based on our understanding of this body of adult L2 learner research, and construct definitions and task designs often draw upon aspects of speaking that are derived from research on which the

focus was adult learners (So, Wolf, Hauck, Mollaun, Rybinski, Tumposky, & Wang, 2015). It can be argued that the unique characteristics of young learners, with their developing cognition, varying degrees of socio-emotional maturation and world experience, may affect how they interpret and understand assessment tasks. Thus, the developmental patterns and linguistic profiles of YLSs may differ from those of adult L2 learners.

The purpose of the study was twofold. First, we examined components of speaking performances of YLSs at different levels of proficiency in order to inform construct definitions for YLS oral performance assessments. We focused on the domain of speaking because, as McKay (2006) noted, oral language is the essence of young learners' language learning and central to the language ability of young EFL learners. Cameron (2001) suggested that classroom activities for young learners should focus on fostering oral language skills. Given the importance of speaking skills, our investigation into YLSs' spoken discourse aims to provide insights into the developmental patterns of young learner speech and inform the design of assessment tasks for the age group. Second, we explored the effects of task type on ratings of YLSs' speaking proficiency with the aim of gaining an understanding of how task design might affect performances of YLSs on speaking examinations.

The study was conducted within the context of a larger research effort that aimed to provide validity evidence for a YLS assessment developed by Educational Testing Service (ETS), the *TOEFL Junior* Comprehensive test (for a detailed description of the test, see (So et al., 2015). We used the test as an instrument for examining the broader construct of speaking proficiency for YLSs, given the availability of data. We also aimed to provide empirical evidence to support score interpretations of the test.

Performance tasks that demand different language (e.g., listening and speaking) and cognitive (e.g., picture-based descriptions, information retelling based on input) skills may affect the performance of YLSs, whose cognition is still developing (Bailey & Heritage, 2014). Although the use of integrated tasks has garnered significant research interest in the field of language testing, to the best of our knowledge, no publicly available study has examined children's responses to integrated speaking tasks. Existing literature that explores the development of children's speaking proficiency predominantly uses picture narrations (e.g., Djigunović, 2016; Heilman, Miller, & Nockerts, 2010; Wolf, Lopez, Oh, & Tsutagawa, 2017) or oral interviews (Djigunović, 2016). We aim to fill this gap by including two task types, a picture narration and an integrated listen/speak task from the *TOEFL Junior* Comprehensive test, to shed light on the impact of task design on YLS oral performances. Detailed descriptions of the tasks are given in the Methodology section.

Literature Review

In this section we review literature that provides the theoretical foundation of the study. In particular, we focus on the performance features of *fluency*, *vocabulary*, *grammar*, and *content*, and the effects of *task type* on test takers' performances.

Fluency

In this study, the concept of fluency is defined according to the seminal work by Lennon (1990, 2000). Lennon (2000) stated that fluency "can be measured both impressionistically and instrumentally by speech rate, and by such dysfluency markers as filled and unfilled pauses, false starts, hesitations, lengthened syllables, retraces, and repetitions" (p. 25) and proposed that "a working definition of fluency might be the rapid, smooth, accurate, lucid,

and efficient translation of thought or communicative intention into language under the temporal constraints of on-line processing" (p. 26). This definition of fluency works well for our context because of its focus on the different measurable dimensions of fluency that we are interested in investigating and is thus adopted for the current study.

Fluency is researched primarily from three aspects: *breakdown fluency*, which concerns the pausing features of continuous speech (Ginther, Dimova, & Yang, 2010; Kormos & Dénes, 2004); *speed fluency*, which is characterized as the rate of speech delivery (Ginther et al., 2010); and *repair fluency*, which relates to the number of self-corrections and repetitions or reformulations present in speech (Iwashita, Brown, McNamara, & O'Hagan, 2008). Speech rate has been shown to be a consistently strong predictor of L2 fluency (Ginther et al., 2010; Kormos & Dénes, 2004). Mean length of run (Ginther et al., 2010; Towell, Hawkins, & Bazergui, 1996), word stress (Kormos & Dénes, 2004), filled or unfilled (silent) pauses, and repairs or repetitions (Kormos & Dénes, 2004) are also related to oral fluency to varying degrees.

Research on L1 children's fluency development suggests that disfluencies are common and tend to reduce in frequency at the later period of language development (Ambrose & Yairi, 1999; Boscolo, Ratner, & Rescorla, 2002; Kowal, O'Connell, & Sabin, 1975). Thus, it is unsurprising to see frequent disfluencies, such as utterances of partial words or word-level repetitions in the speech of young ELF learners, and it seems reasonable to consider certain speech disfluencies developmentally normal for young EFL learners.

Relatively few empirical studies have attempted to measure the development of fluency among young EFL learners. In a recent longitudinal study, Djigunović (2016) examined the development of oral production among 24 young Croatian EFL learners and traced the participants' performances over four years, from the age of 11 to 14 years. Compared to their task achievement, vocabulary, and accuracy, the participants' global fluency showed the most consistent and steady developmental pattern over time. The children's fluency ratings were almost identical in the two speaking tasks used, a picture description and an oral interview, suggesting that oral fluency was a relatively stable feature of the participants' speech, regardless of task types.

Grammar

Grammar-related studies in language testing and second language acquisition (SLA) research broadly focus on two aspects: grammatical accuracy and grammatical complexity (Norris & Ortega, 2009; Skehan & Foster, 1999; Wigglesworth & Elder, 2010). Two levels of grammatical accuracy are examined: *global accuracy*, considering any and all types of grammatical errors in learner language (e.g., Djigunović, 2016; Foster & Skehan, 1996); and *specific types of error*, such as verb tense, subject–verb agreement, article use, and prepositions (e.g., Brown et al., 2005; Wolf et al., 2017). Grammatical complexity is conceptualized as the elaboration and variation of syntactic patterns that appear in learner language (e.g., Biber, Gray, & Staples, 2016; Iwashita, 2006; Iwashita, McNamara, & Elder, 2001).

Djigunović (2016) traced the development of global grammatical accuracy among a group of young EFL learners over a period of four years and found a non-linear pattern. The students' accuracy progressed as expected from grade 5 to grade 6, dipped in grade 7, and picked up again in grade 8 for the two speaking tasks that were examined. Specific types of grammatical errors were examined in Wolf et al. (2017). The researchers compared the oral performances of English language learners (ELLs) and their native English speaker

counterparts in grades K–2, using one picture-retelling and two picture description tasks. The most frequently observed grammatical error types were verb forms, tenses, subject–verb agreement, and the omissions of subjects, verbs, or objects. It is interesting, if not unsurprising, to note that these error types were present in the responses of both ELLs and non-ELLs, irrespective of grade level.

To the best of our knowledge, no studies have systematically examined the development of grammatical complexity using young EFL learners' speech data, possibly owing to the difficulty involved in segmenting children's often short, fragmented speech into meaningful units for analysis. Our study attempted to fill this gap by adopting a systematic speech segmentation method, the Analysis of Speech Unit (ASU) (Foster, Tonkyn, & Wigglesworth, 2000), to examine the grammatical complexity of YLS speech.

Vocabulary

Research has persistently found that (1) as language proficiency level increases, so does the general state of L2 learners' lexical knowledge, and (2) lexical competence is directly related to a language learner's ability to communicate effectively (Nation, 2001; Read, 2000). Aspects of lexical knowledge such as *lexical range*, that is, the range of a learner's vocabulary as displayed in his or her language use (deBoer, 2014), and *lexical sophistication*, that is, "the proportion of relatively unusual or advanced words in the learner's text" (Read, 2000, p. 203), are employed as indicators of lexical knowledge in discourse-based analysis of speaking proficiency. Word types, tokens, and type–token ratio are often used as indicators of lexical range, and word frequencies as indicators of lexical sophistication (Nation, 2001).

Similar to adult learners, YLSs develop their lexical knowledge along the dimensions of range and sophistication. On the one hand, they learn more and more new words in the language classroom, increasing the size of their vocabulary (*range*). At the same time, they are confronted with words of varying frequencies of use (*sophistication*). Young students know and acquire high-frequency words such as *bird* and *book* earlier than low-frequency words such as *falcon* and *fiction*. Vermeer (2000) examined lexical richness among young learners of Dutch and found that the number of words, types, type–token ratio, and word frequency were good measures of children's lexical knowledge. Roessingh and Elgie (2009), in their investigation of early language development among ELLs in grades K–2 in Canada, found that ELLs acquired a few hundred high-frequency English words and developed basic academic language very quickly; however, they lacked the low-frequency words that were observed in the speech of young native speakers. Their results revealed that the young ELLs in their study tended to depend heavily on the first 250 high-frequency words to convey meaning.

Recent developments in computational linguistics and automated text analysis have expanded the investigation of lexical proficiency to different dimensions of lexical knowledge; in particular, the psycholinguistic properties of words such as concreteness, familiarity, imageability, and age of acquisition (Crossley, Salsbury, & McNamara, 2010; Kyle & Crossley, 2015). Word concreteness is the perception of how abstract a word is, as judged by raters, based on how easy it is to describe the word's meaning (Brysbaert, Warriner, & Kuperman, 2014). Word familiarity is evaluated on the basis of judgments of how familiar words are to raters, and imageability scores are derived from judgments of how easy it is to create an image of a word (Kyle & Crossley, 2015). Age-of-acquisition indices are based on raters' estimates of the age at which a given word is learned (Kuperman, Stadthgagen-Gonzalez, & Brysbaert, 2012). These psycholinguistic properties of words can be evaluated

by using the online text analysis tool *Coh-Metrix* (Graesser, McNamara, Louwerse, &Cai, 2004), which allows researchers to explore the deeper cognitive functions of lexical acquisition. These lexical features have proven useful in predicting adult L2 learners' language proficiency and lexical knowledge and are targeted for analysis in the current study.

Content

In this study, speech content is defined as content elaboration, relevance, and accuracy, indicating the degree to which the content of a spoken performance is relevant and well-elaborated on given topics and is an accurate reflection of source materials in integrated tasks. This definition was developed on the basis of the content-related rating criteria of the speaking section of the *TOEFL Junior* Comprehensive test (see Appendices A and B) and informed by previous research in discourse-based analysis of speech content (Brown et al., 2005; Frost et al., 2011; Sato, 2011).

Few empirical studies have investigated content aspects of speech, despite the fact that research has repeatedly shown that raters pay considerable attention to the content of examinee responses (Brown et al., 2005; Sato, 2011) and that rating scales invariably include content as one of the major rating criteria (Inoue, 2009). In a comprehensive analysis of 200 responses of adult L2 learners on the *TOEFL iBT* speaking test, Brown et al. (2005) examined the content quantity (i.e., the number of T-units and clauses in the speech sample) and quality (i.e., the schematic structure of the spoken discourse) of responses at different score levels. The researchers found mixed results for the quantity measures, as the number of T-units was not consistently different across the five proficiency levels that were investigated, but the number of clauses did increase with proficiency level as expected. A similar finding was reported by Inoue (2009), in which the quantity and quality of speech content of performances on the Health Sciences Communication Skills Test were found to improve along with the level of proficiency.

The speech content of YLSs is rarely empirically examined in language testing literature. Wolf et al. (2017) explored content-related errors made by ELLs, by examining the coverage of major events and logical development in picture-based retelling and description tasks. The researchers found that children's speech content was often fragmented, with limited elaboration and accuracy.

Evaluating content aspect of performances in integrated speaking tasks that call for the use of multiple language modalities (e.g., listening and speaking) has triggered some research interest recently, owing to the challenges involved in rating content appropriately and accurately. Researchers have approached content-related aspects of speech through the dimensions of schematic structure (Brown et al., 2005; Frost et al., 2011; Inoue, 2009), content elaboration and development (Sato, 2011), key points coverage (Frost et al., 2011), and reproduction of source materials (Frost et al., 2011). Frost et al. (2011) evaluated aspects of spoken data elicited from a listen/speak task in the Oxford English language test and found that the number of key points covered and the schematic structure of the test takers' performances were related to the test takers' proficiency levels, and that the number of ideas accurately reproduced from the source text increased with proficiency.

Although the use of integrated speaking tasks in large-scale language assessments has become increasingly popular, to date no publicly available empirical study has taken a discourse-analytical approach to the investigation of the content features of YLS spoken responses to integrated tasks. This area warrants further research in order to describe more

fully the developmental progressions of YLSs' speech content elicited from tasks that call upon different language skills.

Task types

Different task types are characterized by their design features (e.g., a storytelling task using pictures versus an oral interview), performance conditions (e.g., tasks with or without planning time), and the demands they place on test takers. In this study, we were interested in the effects of task types on test takers' performances and included two types of tasks: a picture narration task and an integrated listen/speak task. We hypothesized that the differences in the design characteristics of the two tasks would place different levels of cognitive demands on young students and affect their performances.

Tavakoli (2009) investigated the effects of task structure and storyline complexity of oral narrative tasks on task performances and found that test takers produced more accurate and fluent responses when responding to the more structured tasks, which involved a clear sequence of events. Elder, Iwashita, and McNamara (2002) examined performance differences on eight different narrative tasks with varying degrees of task demands. The tasks were manipulated by the researchers to make them either easier (i.e., less cognitively demanding) or more difficult (i.e., more cognitively demanding). The researchers found no marked differences in the quality of test takers' performances across task conditions. It should be noted that these two studies did not compare tasks of different types, but instead compared different task conditions for only one task type, that is, narrative tasks. In contrast, Brown et al. (2005) found little difference in test takers' performances across two independent and three integrated *TOEFL iBT* speaking tasks.

Little empirical research in language assessment is available to describe the effects of different task types on oral performances of YLSs (e.g., Djigunović, 2016; Wolf et al., 2017). Djigunović (2016) compared test takers' performances on a picture narration and an oral interview task using a five-point, holistic scale to assess task achievement, vocabulary, accuracy, and fluency. The researcher found some effects of task type on vocabulary and grammar with better performances observed for the interview task. Wolf et al. (2017) examined test takers' performances on one picture retelling and two picture description tasks. The picture retelling task required students to listen to a story (196 words long) while looking at four pictures depicting the story and then retell the story. This task appeared to be more challenging than the two picture description tasks, which asked students to describe a sequence of events presented in an animation with a shorter narration (56 and 89 words, respectively). The researchers speculated that the animated input could have reduced the level of processing load or increased the level of task engagement and thus led to better task performances. This study provides an insight into how tasks with varying degrees of cognitive load could affect the performances of young students and highlights the importance of taking into consideration age-appropriateness in task design.

The goal of this study was to examine components of speaking proficiency among YLSs and the effects of task type on test takers' performances. We focused our research inquiry on features of fluency, grammar, vocabulary, and content, and we used responses elicited from two speaking tasks on the *TOEFL Junior* Comprehensive to address the following research questions:

1. How do performance features of fluency, grammar, vocabulary, and content distinguish YLSs at different levels of speaking proficiency?

2. Does the elicitation of YLS's spoken features differ across a picture narration and an integrated listen/speak task? If so, how?

Methodology

We adopted a discourse-analytic approach for analyzing actual test performance data in aspects of fluency, grammar, vocabulary, and content. This research methodology has been employed in several studies that examined rating-scale development and test validation (e.g., Brown et al., 2005; Frost et al., 2011; Iwashita et al., 2008). Results of previous empirical studies have collectively shown that spoken discourse features could differentiate test takers between score bands and illuminate differences in aspects of oral performances across task types. This line of research highlights the applicability of discourse-based analysis as a means of gaining insights into the linguistic profiles of young students and the demands that assessment tasks place on this population. Given the scarcity of empirical research on components of YLSs oral proficiency, our study makes an important contribution by analyzing YLSs' performance discourse based on a large-scale, standardized YLS assessment, that is, the speaking section of *TOEFL Junior* Comprehensive.

Instrument

TOEFL Junior Comprehensive was launched in July 2012 and is a computer-delivered test consisting of four sections: reading comprehension, listening comprehension, speaking, and writing. The test is designed for students age 11 years and older. It assesses the academic and social English-language communication skills that are representative of English-medium instructional environments. The test measures language proficiency in situations and tasks that are representative of English-medium school contexts (So et al., 2015). The main uses of the test are as follows: to determine the ELP levels of students on the basis of their test performances; to support decisions regarding placement of students into English-language programs (Papageorgiou & Cho, 2014); and to provide information about student progress in developing ELP over time (Gu, Lockwood, & Powers, 2015).

The speaking section of the *TOEFL Junior* Comprehensive test consists of four tasks: read aloud; picture narration (PN); non-academic integrated listen/speak (L/S); and academic integrated Listen/Speak. The PN and the non-academic L/S tasks were chosen for analyses because they represent two task types and cover two target language use (TLU) domains. The PN task measures a test taker's ability to communicate in the social, interpersonal TLU domain. In this task, test takers are presented with a six-picture sequence, have 60 seconds to prepare their answers, and are asked to narrate a story based on the pictures in a 60-second recording. The L/S task measures the ability to communicate in the navigational TLU domain within a typical classroom setting. An example of the navigational domain would be students communicating with peers about homework assignments to obtain some details or to get key information from school-related announcements. Students are expected to listen to a short lecture, prepare a response within 45 seconds, and recount the key information conveyed in the lecture in a 60-second recording.

Responses to the speaking test are scored by two human raters on two task-specific scoring rubrics that assess three construct areas: delivery, language use, and content. Each task is evaluated on a four-point scale, from 1 to 4; responses that are off topic are scored as 0. The speaking section score of the *TOEFL Junior* Comprehensive test is the sum of the four item scores and ranges from 0 to 16. The scores are mapped to four levels of proficiency on the Common European Framework of Reference (CEFR) (Council of Europe, 2001): below A2 (1–7 points), A2 (8–10 points), B1 (11–13 points), and B2 (14–16 points) (Papageorgiou, Xi,

Morgan, & So, 2015). Sample responses were selected and grouped on the basis of these four levels of proficiency and were labeled as Level 1, Level 2, Level 3, and Level 4 in this study.

Spoken responses

The spoken responses were chosen from the *TOEFL Junior* Comprehensive test database. Initially, a sample of 180 test takers (45 at each of the four proficiency levels) was selected from a large pool of test data. To the greatest extent possible, these test takers were chosen to represent the global *TOEFL Junior* Comprehensive population in terms of age, gender, L1, and native country. After data cleaning, the responses of one Level 3 test taker were discarded owing to poor audio quality (Level 3 $N = 44$). The final data set contained 358 responses produced by 179 YLSs of different ages ($M = 13.6$, $SD = 2.87$), and included 85 males and 94 females. The test takers came from a variety of first language backgrounds, including Arabic, Chinese, German, Indonesian, Japanese, Korean, Portuguese, Spanish, and Vietnamese.

The audio files of the selected spoken samples were transcribed verbatim by professional transcribers. Features of breaths or grunts were not retained in the transcripts. If a word was unintelligible, that word was annotated with a <%> symbol. The annotation symbols were removed from the transcripts prior to data analysis.

Data analysis

The transcripts and audio files were analyzed by using human coders and three types of software: (1) *SpeechRater*^{SM}, an automated scoring engine developed by ETS; (2) the public-domain software *VocabProfile* (Heatley, Nation, & Coxhead, 2002); and (3) *Coh-Metrix* (Graesser et al., 2004). These tools have been used in many previous studies as measures of spoken proficiency and lexical knowledge (e.g., Crossley et al., 2010), and have proved useful for our analysis based on a small-scale pilot study (Hsieh &Gu, 2015). In the pilot study, we explored a wide range of spoken features, using discourse-based analysis to determine the extent to which differences in YLS spoken features were observable or measurable across levels and task types. We used automated tools and human coding to analyze a separate, smaller pool of *TOEFL Junior* Comprehensive test takers' responses. The results helped us to identify a list of measures of fluency, grammar, vocabulary, and content that were sensitive to children's developmental levels. The most salient features were included in the current study and were selected largely on the basis of construct relevance and their correlational strength and predictive power with respect to the test takers' *TOEFL Junior* Comprehensive speaking section scores.

Certain aspects of grammar and content were coded by the two authors using a coding scheme that was developed and refined in light of the pilot study results and our review of the literature. The spoken responses were first segmented into Analysis of Speech Units (ASU) (Foster et al., 2000). In this study, we define an ASU as a single speaker's utterance that consists of either an independent clause, or subclausal unit (e.g., "Oh poor woman," "Thank you very much," "Yes") with any subordinate clause. We used // to mark the end of each ASU, and: to mark the clause boundaries within an ASU. An example of a segmented response is shown below:

First, a lot of people they were buying tickets to the game, to the game, the soccer game.//

Then um, well, people started buying um, //

then uh, two boys is they doesn't buy um, tickets:: because there //

and then the um, the boys left //

the boy sat //

and then start to rain //

and the people start getting cold //

and then a lot of people start um, //

they were watching TV//

To code the feature of content quality, we adapted the "key points" approach utilized by Frost et al. (2011), based on the results of our pilot study that examined the utility of this measure as an indication of content quality. To derive a list of key points for coding, we first examined the six pictures in the PN task, the listening stimulus material, and the prompt in the L/S task. We jointly derived six key points for the PN task, each associated with one picture. We also derived six key points for the L/S task, three associated with the three keywords provided in the prompt and three related to the detailed content information provided in the audio input. The numbers of key points accurately mentioned by the test takers in their responses were tallied and used as an indicator of the content quality of the responses. It is important to note that these key points were not specified in the official scoring guides. There was a detail of all the spoken features analyzed in the study.

View larger version

To establish inter-coder reliability for the coding of the numbers of clauses, ASUs, and key points covered, we first carried out a trial coding of 12 test takers' speech samples. After all problematic cases were resolved by consensus, the authors independently coded another six test takers' responses to establish inter-rater reliability. Spearman's *rho* correlations were calculated to compare the numbers of clauses, ASUs, and key points coded by each coder for the two tasks. The reliability indices ranged between .83 and .91, showing that the speech samples were coded reliably.

Statistical analyses

We performed a series of mean-comparison, inferential statistical analyses to address the research questions. For all analyses, we set the significance level at .05. We report the results with exact *p*-values and partial eta squared (η^2_p) effect sizes to indicate the percent of the variance accounted for by the main effect. The criteria for small, medium, and large effect sizes are .01, .06, and .14, respectively, following Cohen (1988). We also conducted post-hoc pairwise comparisons to examine the differences between the four levels of proficiency. The Bonferroni adjustment procedure ($p < .083$) was applied to control for family-wise Type I error (.083 = .05/6). All statistical analyses were performed using PASW Statistics 18.

A two-by-four (two task types by four proficiency levels) multivariate analysis of variance (MANOVA), including the 21 spoken features as the dependent variables, was first performed to detect whether there were group differences along the combined set of variables. Multivariate outliers were first checked by assessing Mahalanobis distances among all the cases; 15 outliers were identified and removed for further analyses. The 21 dependent variables were checked for linearity using scatterplot matrices and for multicollinearity using Pearson correlations; these assumptions were met ($r < .90$, Tabachnick & Fidell, 2001). The homogeneity of variance-covariance matrices was examined by using Box's M test ($p <$

.001). Due to the violation of this assumption, we used Pillai's Trace to report the main effects given that this test statistic is robust to the violation when group sizes are equal, which was the case (Field, 2005). Results of the MANOVA analysis with Pillai's Trace showed a significant main effect for proficiency, $F(63, 951) = 5.797$, $p < .001$, $\eta^2_p = .277$; a main effect for task type, $F(21, 315) = 28.986$, $p < .001$, $\eta^2_p = .659$; and a significant interaction effect, $F(63, 951) = 1.832$, $p < .001$, $\eta^2_p = .108$.

Based on the significant MANOVA results, we conducted a series of mixed–design ANOVA tests for each of the 21 variables to determine where the group differences were located. Task type was used as the within-subjects variable and proficiency level served as the between-subjects variable. Prior to running the ANOVAs, we checked for univariate outliers and those identified were removed. Outliers were defined as values more than three standard deviations away from the mean. There was a provision of the means and standard deviations for all measures at each proficiency level and for each task after outliers were removed

Six fluency measures were included in the ANOVA analyses. All fluency measures were significantly influenced by proficiency with medium to large effect sizes. The descriptive statistics showed that higher-level test takers tended to speak faster with shorter pause duration, relatively fewer filled and unfilled pauses, and produced longer stretches of chunks. No significant effect of task type was found for any of the fluency measures, suggesting that test takers' fluency features were relatively consistent across task types. Three fluency measures (MSD, NSS, and NRR) showed a significant interaction effect between proficiency and task type. Test takers at Levels 2, 3, and 4 were found to pause more frequently in the PN task, whereas Level 1 test takers paused more frequently when responding to the L/S task. Test takers at Levels 3 and 4 produced a larger number of repetitions and repairs when responding to the L/S task, whereas Levels 1 and 2 test takers had more prominent problems with breakdown fluency when responding to the P/N task.

Grammar

Grammatical accuracy as measured by EFA was significantly influenced by proficiency level with a large effect size. Higher-proficiency test takers tended to produce fewer grammatical errors compared to lower-proficiency test takers. No effect for task type was found, suggesting that the test takers did not produce more or fewer grammatical errors in a particular task. An interaction effect between proficiency and task type with a medium-effect size was observed. The descriptive statistics showed that the difference was marked at Level 4, where test takers produced a significantly larger mean number of EFAs in the PN task ($M = 4.90$, $SD = 2.99$) than in the L/S task ($M = 2.62$, $SD = 3.37$).

Grammatical complexity was assessed by CPA and WPC. A significant main effect for proficiency was observed for CPA with a large effect size, revealing that higher-level speakers produced more clauses per ASU, thus more complex grammatical structures. No significant difference, however, was observed for WPC, suggesting that this granular level of analysis may not be a sensitive measure to assess grammatical complexity across proficiency levels for the test takers' spoken responses. Both CPA and WPC were significantly influenced by task type with medium effect sizes. Regardless of proficiency, test takers produced more clauses per ASU and more words per clause in the L/S task, indicating that the young speakers' language was more complex in the L/S task than in the PN task.

Vocabulary

Nine vocabulary variables were analyzed, covering measures of lexical range, sophistication, and psycholinguistic properties of words. All lexical range measures showed a significant proficiency effect, indicating that more proficient speakers produced longer words, more words, and a wider range of word types in their responses. However, no proficiency effect was found for 1K and 2K word frequency measures.

The four psycholinguistic properties of words, AFR, ACR, AIR, and AAA, albeit not used as rating criteria in the speaking section of *TOEFL Junior* Comprehensive, all showed significant proficiency effects. With increasing proficiency, test takers produced words that were less familiar, more abstract, harder to associate with an image, and having a later age of acquisition, revealing that these word properties are relevant to YLS speaking proficiency. Significant task effects were found for seven vocabulary features: NLW, NTK, 2K, AFR, ACR, AIR, and AAA. Test takers produced words that were longer, larger in total number, more unfamiliar, more abstract, less imageable, and acquired at later ages when responding to the L/S task. It is interesting to note that test takers used a larger percentage of 2K words when responding to the P/N task. This result could be attributable to the fact that some more sophisticated words were required to describe the various events depicted in the six pictures, whereas the L/S task focused on a typical middle-school classroom setting and the vocabulary produced by test takers may have been constrained by the listening input. Significant interactions were observed for word length, types, tokens, 1K, 2K, ACR, and AIR, suggesting that the effects of task type on these lexical measures varied for test takers of different proficiency levels.

Content

Significant proficiency effects for measures of content quantity and content quality were observed. In terms of quantity, more proficient test takers produced larger numbers of ASUs and clauses (NAS and NCZ) compared to the less-proficient ones. Only NAS was significantly influenced by task type with a small effect size.

numbers of accurately covered key points (NKP) with a very large effect size (η^2_p = .548). Post-hoc Bonferroni pairwise comparisons showed that all proficiency levels differed from each other on the NKP measure, except between Levels 3 and 4. The NKP measure also yielded a significant difference across task types with a large effect size (η^2_p = .218). Test takers, regardless of proficiency levels, produced better content quality in the PN task than in the L/S task. Nevertheless, as the proficiency went up, the difference became smaller as indicated in the descriptive statistics.

Significant interaction effects between proficiency and task type were detected for all content features with large effect sizes. Levels 1 and 2 test takers produced larger mean numbers of ASUs and clauses in the PN task than in the L/S task. Contrastively, Levels 3 and 4 speakers produced larger NAS and NCZ in the L/S task than in the PN task. The interaction effect was also detected for content quality measure. Test takers in general produced a larger NKP in the PN task than in the L/S task. The difference decreased with increasing proficiency with the smallest difference found at Level 4, suggesting that the more proficient test takers were better able to tackle the complex integrated L/S task and produce a similar degree of content quality across task types.

Discussion

Results of the study show that the great majority of the features examined significantly differentiated test takers across proficiency levels with moderate to strong effect sizes. With

increasing proficiency, test takers' performances displayed a higher degree of speech fluency, grammatical accuracy and complexity, a wider range of vocabulary, and improved content quantity and quality. These results corroborate findings of previous studies, suggesting that aspects of fluency, grammar, vocabulary, and content are important components of speaking proficiency among YLSs.

Contrary to research findings on adult L2 fluency that show no relationship between holistic scores and the length and frequency of pauses and disfluencies (e.g., Ginther et al., 2010; Kormors & Dénes, 2004), our findings indicate that the frequency and duration of pauses, speaking rate, and features of disfluencies are all associated with children's speaking proficiency and should be considered in scale descriptors when evaluating children's speaking proficiency.

In terms of grammatical complexity, we observed that higher-proficiency test takers uniformly produced more clauses per ASU than less proficient ones. However, the measure of number of words per clause did not follow the same pattern. Level 3 test takers produced fewer words per clause than Levels 1 and 2 test takers on both tasks, a finding comparable to Iwashita (2006), who found non-significant difference in words per clause in the speech of learners of Japanese. As Foster et al. (2000) suggest, keeping track of complex micro-units such as clauses requires attention to the syntactic requirements and the constraints of syntactic constructions when the speech planning unfolds simultaneously. The clauses that we segmented were relatively short, sometimes containing only a few words, which is a characteristic of YLS speech (e.g., Wolf et al., 2017). The results imply that the use of clauses as a unit of measurement may not be very robust especially when analyzing the syntactic complexity of YLS speech. It also suggests that productive use of clauses in spontaneous speech may take a longer time to develop.

Results of the lexical measures collectively showed that more proficient YLSs used significantly more words, longer words, and a wider range of words, and had better access to core lexical items as indicated by the psycholinguistic properties of words (Crossley et al., 2010). The two word frequency measures, 1K and 2K, however, did not differ across proficiency levels. This result is somewhat unexpected, given that lexical sophistication as measured by word frequency has consistently proven to be a strong indicator of adult L2 learners' language proficiency (Nation, 2001; Read, 2000). Our results indicate that children's lexical development does not always follow the same pattern as adults and that young EFL learners may largely rely on high-frequency words as the building blocks for their growing communicative abilities. We also speculate that the results could have been different if narrower vocabulary bands (e.g., first 250, first 500, etc.) had been used because, as Roessingh and Elgie (2009) found, young ELLs rely heavily on the first 250 commonly used English word family in their daily communication. It is also possible that there was a *qualitative* shift in how the test takers used their vocabulary that was not captured in our analysis. We reason that while high-proficiency students may not be using new, low-frequency words in their responses, they may have acquired secondary meanings of high-frequency words and are using these words in deliberate ways to communicate meanings. This depth dimension of lexical knowledge was not examined in the current study. In future work, we believe that the inclusion of measures of depth of lexical knowledge can provide interesting insights into the developmental patterns of YLSs' lexical acquisition (Cremer, Dingshoff, de Beer, & Schoonen, 2010; Schoonen & Verhallen, 2008).

The four psycholinguistic properties of words resulted in significant differentiation of proficiency levels, suggesting that these are valuable measures of speaking proficiency for

YLSs. Recent research in L2 vocabulary has seen an increasing interest in exploring how these lexical properties are related to language proficiency and lexical development in adult L2 speech (Crossley et al., 2010; Kyle & Crossley, 2015). To the best of our knowledge, our study is the first empirical investigation that systematically uses these lexical features to examine young EFL learners' oral proficiency. Our data have yielded promising results that enable us to examine the relationship between these lexical properties and the speaking proficiency of YLSs and pointed to an interesting area for future research. The development of these dimensions of lexical knowledge needs to be examined in other more naturalistic settings and tasks to allow researchers to gather empirical data on how words are actually being acquired and used by this population.

Content quality as measured by NKP appears to be a crucial component of test takers' performances given its very large effect size. The quality of accurate descriptions of the pictures in the PN task and the recount of key information from the source material in the L/S task both increased consistently and significantly according to proficiency levels. Congruent with findings from Frost et al. (2011), we believe that, for human raters, speech content quality is critical for not only (1) integrated listen/speak tasks that involve appropriate reproductions or synthesis of source materials, but also for (2) picture narration tasks that require relevant and adequate descriptions of the events or elements illustrated in the pictures. The results also provide empirical support for the *TOEFL Junior* Comprehensive scoring rubrics that embody this content-related aspect of the speaking construct by appropriately accounting for content accuracy and relevancy in the evaluation of both the picture narration and listen/speak task performances. More importantly, the study provides a much needed insight into the use of integrated tasks in YLS assessments, which is a relatively uncharted territory. Integrating listening and speaking skills to respond to assessment tasks is complex and requires huge efforts on the part of young learners who have limited memory capacity and short attention spans. Examining the demands of integrated tasks on young students using retrospective cognitive interviews can be a fruitful focus in future research.

Regarding the second research question that addresses the impact of task type, unexpectedly, no difference was found for any of the fluency measures. We had assumed that the greater cognitive load of the more complex L/S task would lead to worse performances in aspects of fluency given that previous research has found that the more cognitively demanding integrated tasks would leave less attention resources available to aspects of fluency such as speaking rate and mean length of run (e.g., Brown et al., 2005). As found in Hsieh and Gu (2016), which explored YLSs' strategy use when responding to the same *TOEFL Junior* Comprehensive speaking tasks, we observed that YLSs, regardless of proficiency levels, relied heavily on a note-taking strategy during the pre-task planning to jot down words and sentences in order to help organize their thoughts and assist in production. Within the context of the current study, we speculate that the cognitive load that the L/S task imposed on the test takers could have been reduced with the assistance of pre-task planning time that induced a facilitative and beneficial effect on fluency. On the other hand, the absence of performance differences in fluency measures between task types corroborates findings from some empirical studies on the impact of task demands on both adult L2 test takers' performances (Elder et al., 2002; Wigglesworth & Elder, 2010) and on YLSs' global fluency (Djigunović, 2016). Similar to Djigunović (2016), results of the current study reveal an overall trend showing that children's fluency is consistent when responding to different types of speaking tasks, demonstrating that fluency is a robust component of YLS speaking proficiency.

With regard to grammatical complexity, performances were consistently more complex in the L/S task than in the PN task. We observed that our test takers tended to mimic the sentences they heard in the input material of the L/S task, which consisted of some long stretches of chunks and compound sentences; language of this type is generally more complex than the simple narration young learners often produce. This observation partially explains the differences we found and suggests that the more demanding L/S task could potentially push the young speakers to produce linguistically more complex sentences than they would normally do.

The effect of task type on performance is most pronounced on lexical measures. The prevalence of longer words seen in the L/S task can be attributed to a high proportion of words reproduced from the source material (e.g., the repeated use of the key words *homework* and *assignment* that appeared in the lecture and the written prompt). Interestingly, the PN task elicited more 2K words than the L/S task, albeit with a very small percentage difference. This could be due to the fact that test takers needed to employ some advanced words in order to describe the various events taking place in the pictures. Finally, significant performance differences were found for the four psycholinguistic properties of words across tasks, further demonstrating that task design has an impact on the kind of words being elicited.

Several significant interaction effects between proficiency and task type detected require further discussion. Lower-proficiency test takers appeared to have more sophisticated use of vocabulary and better content when responding to the PN task, while higher-proficiency test takers produced more advanced vocabulary and content when responding to the L/S task. The result indicates that task types affected test takers' vocabulary and content measures differently depending on their level of proficiency. It could be the case that the more cognitively demanding L/S task pushed the more proficient test takers to produce more advanced language, leading to a wider range of lexical use and finer content. On the other hand, some features of the performances of lower-proficiency test takers were raised in the PN task, perhaps because these young students could direct most of their attention to studying the pictures and producing a narrative without having to commit attentional resources to memory when responding to the task. We should note, however, that explicating the nature of the phenomenon is beyond the scope of the study. Future research should continue to investigate the inherent features of task design as a source of variance in YLS speaking proficiency.

Findings of this study have implications for language development and task design for YLSs. The wide range of spoken features that were examined provides a credible developmental pattern of YLS speech at various stages of language acquisition. We believe that this pattern can inform rating-scale development for tests that are intended for YLSs in other testing contexts. Our study also extended the scope of research in speaking proficiency by exploring an integrated task and the psycholinguistic properties of words, which are areas rarely examined in YLS assessment contexts. More research is needed to examine these dimensions of lexical knowledge and integrated speaking tasks in future discourse-based studies of YLS speech. Finally, our study delineates the importance of gaining further insights into the effects of task type on oral performances of YLSs whose speech production is susceptible to influence by the cognitive demands of assessment tasks.

Conclusion

This study explored the speaking proficiency of YLSs by analyzing response characteristics of *TOEFL Junior* Comprehensive test takers. Results of the study show that the spoken features examined largely differentiated test takers across proficiency levels and provide important empirical evidence to support score interpretations for the test.

A few limitations need to be pointed out. First of all, the study investigated only two task types and findings may not be generalizable to other task types not included. Second, we only investigated aspects of fluency, grammar, vocabulary, and content. Future research should look into other features such as pronunciation and intonation. Finally, it should be noted that some of the ANOVA assumptions were violated and the results should be interpreted with caution. With these points in mind, we conclude the study by recommending that future research should continue to investigate the different dimensions of spoken features for young learners in different language learning and testing contexts, and that further studies should seek to determine how these features differ across proficiency levels and task types. A better understanding of the interactions between task types and language proficiency will help to extend our knowledge about the speaking abilities of YLSs and inform the development of language assessments designed for young language learners

Recommendations

1. Learners should be remarkably exposed to training that can enhance them well established inklings that can promote their performance features of fluency, vocabulary, grammar, and content, and the effects of task type on test takers' performances.
2. School management should endeavor to engage learners in debates, speak-out and quiz in order to help develop speaking proficiency among them.

REFERENCES

Ambrose, N. G., Yairi, E. (1999). Normative disfluency data for early childhood stuttering. *Journal of Speech and Hearing Research*, 42(1), 895–909.

Bailey, A. L., Heritage, M. (2014). The role of language learning progressions in improved instruction and assessment of English language learners. *TESOL Quarterly*, 48(3), 480–506.

Biber, D., Gray, B., Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, 37(5), 639–668.

Boscolo, B., Ratner, N. B., Rescorla, L. (2002). Fluency of school-aged children with a history of specific expressive language impairment: An exploratory story. *American Journal of Speech-Language Pathology*, 11(1), 41–49.

Brown, A., Iwashita, N., McNamara, T. F. (2005). *An examination of rater orientations and test taker performance on English for Academic Purposes speaking tasks*. TOEFL Monograph No. MS-29. Princeton, NJ: Educational Testing Service.

Brysbaert, M., Warriner, A. B., Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(1), 904–911.

Cameron, L. (2001). *Teaching languages to young learners*. Cambridge: Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Council of Europe (2001). *Common European Framework of Reference for Languages*: Learning, teaching, assessment. Cambridge: Cambridge University Press.

Cremer, M., Dingshoff, D., de Beer, M., Schoonen, R. (2010). Do word associations assess word knowledge? A comparison of L1 and L2, child and adult word associations. *International Journal of Bilingualism*, 15(2), 187–204.

Crossley, S., Salsbury, T., McNamara, D. S. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, 60(3), 573–605.

deBoer, F. (2014). Evaluating the comparability of two measures of lexical diversity. *System*, 47(1), 139–145.

Djigunović, J. M. (2016). *Individual learner differences and young learners' performance on L2 speaking tests*. In Nikolov, M. (Ed.), Assessing young learners of English: Global and local perspectives (pp. 243–261). New York: Springer.

Elder, C., Iwashita, N., McNamara, T. F. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer. *Language Testing*, 19(4), 343–368.

Field, A. (2005). *Discovering statistics using SPSS*. Sage: London.

Foster, P., Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18(3), 299–323.

Foster, P., Tonkyn, A., Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3), 354–375.

Frost, K., Elder, C., Wigglesworth, G. (2011). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing*, 29(3), 345–369.

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.

Graesser, A., McNamara, D. S., Louwerse, M. M., Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193–202.

Gu, L., Lockwood, J., Powers, D. E. (2015). Evaluating the *TOEFL Junior® Standard* test as a measure of progress for young English language learners. ETS Research Report RR-15–22. Princeton, NJ: Educational Testing Service.

Heatley, A., Nation, P., Coxhead, A. (2002). *Range and frequency programs*. Retrieved from www.victoria.ac.nz/lals/staff/paul-nation.aspx

Heilmann, J., Miller, J. F., Nockerts, A. (2010). Sensitivity of narrative organization measures using narrative retells produced by young school-age children. *Language Testing*, 27(4), 603–626.

Hsieh, C.-N., Gu, L. (2015). *Comparing the discourse features of young language learners' responses to different oral language tasks*. Paper presented at the annual conference of the American Association for Applied Linguistics Conference (AAAL), Toronto, Canada.

Hsieh, C.-N., Gu, L. (2016). *Young EFL learners' strategy use during L2 speaking performance*. Paper presented at the 38th Language Testing Research Colloquium (LTRC), Palermo, Italy.

Inoue, M. (2009). Health sciences communication skills test: The development of a rating scale. *Melbourne Papers in Language Testing*, 14(1), 55–91.

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly*, 3(2), 151–169.

Iwashita, N., Brown, A., McNamara, T. F., O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24–49.

Iwashita, N., McNamara, T. F., Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning*, 21(3), 401–436.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.

Kormos, J., Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145–164.

Kowal, S., O'Connell, D. C., Sabin, E. F. (1975). Development of temporal planning and vocal hesitations in spontaneous narrative. *Journal of Psycholinguistic Research*, 4(1), 195–207.

Kuperman, V., Stadthagen-Gonzalez, H., Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978–990.

Kyle, K., Crossley, S. (2015). Automatically assessing lexical sophistication: Indices, tools, findings and application. *TESOL Quarterly*, 49(4), 757–786.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3), 387–417.

Lennon, P. (2000). *The lexical element in spoken second language fluency*. In Riggenbach, H. (Ed.), Perspectives on fluency (pp. 25–42). An Arbor, MI: The University of Michigan Press.

McKay, P. (2006). *Assessing young language learners*. Cambridge: Cambridge University Press.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nikolov, M. (2016). *Trends, issues, and challenges in assessing young language learners*. In Nikolov, M. (Ed.), Assessing young learners of English: Global and local perspectives (pp. 1–17). New York: Springer.

Norris, J., Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578.

Papageorgiou, S., Xi, X., Morgan, R., So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly*, 12(2), 153–177.

Papageorgious, S., Cho, Y. (2014). An investigation of the use of *TOEFL® Junior™* Standard scores for ESL placement decisions in secondary education. *Language Testing*, 31(2), 223–239.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Roessingh, H., Elgie, S. (2009). Early language and literacy development among young English language learners: Preliminary insights from a longitudinal study. *TESL Canada Journal*, 26(2), 24–45.

Sato, T. (2011). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, 29(2), 223–241.

Schoonen, R., Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, 25(2), 211–236.

Skehan, P., Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49(1), 93–129.

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., Wang, J. (2015). *TOEFL Junior® Design Framework*. TOEFL Junior Research Report No. 02. ETS Research Report, No. RR-15–13. Princeton, NJ: Educational Testing Service.

Tabachnick, B. G., Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn & Bacon.

Tavakoli, P. (2009). Assessing L2 task performance: Understanding effects of task design. *System*, 37, 482–495.

Towell, R., Hawkins, R., Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84–119.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17(1), 65–83.

Wigglesworth, G., Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, 7(1), 1–24.

Wolf, M. K., Lopez, A., Oh, S., Tsutagawa, F. S. (2017). *Comparing the performance of young English language learners and native English speakers on speaking assessment tasks*. In Wolf, M. K., Butler, Y. G. (Eds.), English language proficiency assessments for young learners. New York: Routledge.