Measuring the Power of Learning.®

Research Memorandum
ETS RM–16-16

Effects of Printed Option Sets on Listening Item Performance Among Young English-as-a-Foreign-Language Learners

Edward Getman

Yeonsuk Cho

Christine Luce

December 2016

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Effects of Printed Option Sets on Listening Item Performance Among Young English-as-a-Foreign-Language Learners

Edward Getman, Yeonsuk Cho, and Christine Luce
Educational Testing Service, Princeton, New Jersey

December 2016

Corresponding author: Edward Getman, E-mail: egetman@ets.org

Suggested citation: Getman, E., Cho, Y., & Luce, C. (2016). *Effects of printed option sets on listening item performance among young English-as-a-foreign-language learners* (Research Memorandum No. RM-16-16). Princeton, NJ: Educational Testing Service.

Action Editor: James Carlson

Reviewers: Robert French and Guangming Ling

Abstract

The *TOEFL® Primary™* tests from Educational Testing Service (ETS) are intended to assess the English language proficiency of young students 8 years of age and older who are learning English as a foreign language (EFL). Separate tests assess listening, reading, and speaking skills. The TOEFL Primary Listening test includes selected-response items in which the option sets are both printed in a test book and read aloud. Concern was raised by test administrators that presenting the options in print might unduly impact young test takers, especially those whose reading ability in English is low, as it may require the ability to process both visual and auditory information concurrently. The current study investigated the effects of the presence of printed option sets on TOEFL Primary Listening test performance. In a sample of 747 test takers from Mongolia, Colombia, and Brazil, we examined whether any difference in performance on listening items existed when the option sets were presented both aurally and visually versus being presented aurally only. The participants took one of four experimental test forms consisting of (a) a reading section, (b) a listening section with options sets presented in both aural and visual modes, and (c) a second listening section with option sets presented in an aural mode only. The listening test items were counterbalanced across test forms to control for order effects. Participants also completed a brief survey on their perceptions about the relative difficulty of listening items under the two conditions. A repeated-measure general linear model (GLM) was employed to compare performance of listening items under both conditions. Results indicated that the presence of printed text options did not make items more difficult for test takers. In addition, test takers perceived that the presence of text options made the listening items easier, contrary to the concern that motivated the current study.

Key words: *TOEFL® Primary™*, young language learners, listening tests, EFL, assessment

Aimed at young learners 8 years of age and older, the *TOEFL® Primary™* tests are intended to be low-stakes tests that support English language learning. The TOEFL Primary Reading and Listening tests are paper-and-pencil tests composed of approximately 36 reading items and 36 listening items, respectively. Each form is available in two steps to accommodate the varying levels of beginning English learners (Step 1) and more advanced English learners (Step 2). The items are all selected-response (multiple-choice) format with a stem and three response options. A separate computer-based speaking test is also available. Since the launch of TOEFL Primary in 2013, some teachers and test administrators have reported that when young test takers move from items that are presented aurally only to ones in which the option sets are also printed in the test book, some test takers, especially those who have lower reading proficiency levels, appear distracted.

Currently, stems and option sets for a majority of the listening item types are printed in the test book as well as being read aloud (on a recording). The purpose of this study is to evaluate whether the presence of the printed option sets is a source of construct-irrelevant variance, to inform future test development activities (especially for the young English language learner population), and to potentially add more evidence to support the validity and usefulness of the TOEFL Primary tests.

Literature Review

The effects of different response formats (i.e., selected response versus constructed response) on language test task performance have been examined in prior studies (see In'nami & Koizumi, 2009), and researchers have also looked specifically at the influences of various presentations of selected-response items on second language listening task performance (e.g., Chang & Read, 2013; Hemmati & Ghaderi, 2014; Yanagawa & Green, 2008). Yanagawa and Green (2008) investigated the influence of different presentation formats of multiple-choice listening items on test performance. The three formats included in the study were full question preview (FQP), in which the stem and options are visible before the presentation of any stimulus material; answer option preview (AOP), in which only the options are visible; and question stem preview (QSP), in which the item stem only is presented visually. The full items—stem and options—were made available after the stimulus had concluded. The 279 adult Japanese participants were separated into three groups—one for each format condition. Thirty items from an old *TOEIC®* test were adapted for each condition. Using pretest scores as a covariate to

control for variations in listening ability, an analysis of covariance revealed significant differences between FQP and AOP and between AOP and QSP, but not between FQP and QSP. Yanagawa and Green concluded that previewing stems was beneficial to test takers but previewing options alone was not.

Hemmati and Ghaderi (2014) also investigated whether significant differences in performance on selected-response listening items could be found depending on the item format. In addition to FQP, AOP, and QSP, a fourth format—no preview (NP)—was included. A 20-item test was devised and administered under one of the conditions to each of 60 intermediate Iranian university students studying EFL. ANOVA analyses revealed a significant difference in performance between the NP condition and the other three conditions. However, in contrast with the findings of Yanagawa and Green (2008), no significant differences were found among the FQP, QSP, and AOP conditions. Thus, Hemmati and Ghaderi concluded that previewing any item content before listening to a stimulus can facilitate better performance on those items.

However, very few studies have examined whether there are any differences in test-taker performance on listening items when selected-response item stems and options are presented aurally, visually, or both. One study by Chang and Read (2013) investigated whether there were any differences in performance between selected-response listening items that were presented aurally (i.e., read aloud) and those that were presented visually (i.e., printed in a test book). A sample of 87 Taiwanese EFL university students answered 60 listening items, with half the items presented under each of the two conditions. No significant differences in performance were found between the two conditions. However, when examining potential interactions related to ability level, the benefit of printed selected-response items over aural ones among students with low ability became apparent. Interestingly, a majority of all students preferred the printed item format.

These findings, however, are not necessarily generalizable to young learners, such as those who constitute the target population for TOEFL Primary. The developing cognitive capacities of young learners must be considered in language assessments (Hasselgreen, 2005; McKay, 2006). According to the Cognitive Load Theory (Sweller, 1994), if the cognitive demands of a task exceed the processing capacities of the student, then performance may be compromised. It is not yet clear whether a heightened reading load, such as presenting printed options sets to selected-response listening items in addition to presenting them aurally, or a

heightened listening load, as occurs in the absence of such printed option sets, leads to cognitive bottlenecking for young language learners.

The theory of multimodal learning (Mayer & Moreno, 1998; Moreno & Mayer, 1999) suggests that presenting information both visually and aurally may support and enhance cognitive processing, as has been observed among adult language learners (e.g., Al-Seghayer, 2001; Chun & Plass, 1996; Jones & Plass, 2002). However, some evidence indicates that this observation may not hold for younger learners (e.g., Acha, 2009). Verbal information, if presented as text, may be carried by both the verbal and the visual channels, both of which have limited capacities (Mayer & Moreno, 2003). A redundancy effect occurs when redundant media input overloads such information-processing channels (Kalyuga, Chandler, & Sweller, 1998, as cited in Mayer, Heiser, & Lonn, 2001). Therefore, cognitive resources, which may already be limited due to age or lack of reading ability, may be diverted by reading the option sets and could affect subsequent performance on listening items. We are unaware of any published research that examines this possibility.

Research Questions

Previous literature has focused on the effects of various formats of selected-response listening items on the performance of adult learners. The current study is one of the first to examine the effects of various formats of selected-response listening items on the performance of young learners. The following research questions are addressed:

1. Does the way in which selected-response listening items are presented (i.e., with or without printed option text) affect the performance of young language learners?
2. What is the relationship between young language learners' age and performance difference between the two conditions?
3. What is the relationship between young language learners' reading ability and performance difference between the two conditions?
4. How do young language learners perceive the relative difficulty of listening items under the two conditions?

Method

The research questions were addressed through a quasi-experimental research design.

Participants

A total of 747 students from three countries—Brazil (252), Colombia (192), and Mongolia (303)—participated in the study. Participants ranged from 7 to 15 years of age (median = 10.75, *SD* = 1.34), and a vast majority (98%) were between 8 and 13 years of age. Both genders were well represented, with 51.8% of the participants identifying as male. The sample is believed to be reflective of the population of EFL learners who might take the TOEFL Primary test.

Materials

Four experimental test forms were created and included (a) a reading section, (b) a listening section with selected-response option sets presented both aurally and visually, and (c) a listening section with selected-response option sets presented aurally only. Each form had the same reading section with 19 items (in addition to three unscored practice items). The reading section was included to measure participants' reading ability to be used in the analysis. The four test forms also had the same sets of listening items (Listening A and Listening B), each consisting of 17 discrete selected-response items with three options (in addition to two sample items). These listening item sets contained the same item types with similar estimated difficulties. Item Sets A and B were further adjusted so that options were both written and spoken (WS) or spoken only (S). Although each form contained the same listening item sets, they varied in terms of option presentation and order. As shown in Table 1, the four experimental test forms counterbalanced the order of the option conditions (WS or S) and the order of the listening sets (Listening A first followed by Listening B or vice versa). Because the listening item sets were assumed to be parallel, the form differences were not considered as a factor in the study design. Recordings of the listening stimuli and options were made to standardize the administrations of those sections.

Table 1. Configuration of Experimental Test Forms

Form	Section 1 (19 items)	Section 2 (17 items)	Section 3 (17 items)
1	Reading	Listening A – WS options	Listening B – S options
2	Reading	Listening A – S options	Listening B – WS options
3	Reading	Listening B – WS options	Listening A – S options
4	Reading	Listening B – S options	Listening A – WS options

Note. Each section had two or three sample items that were not scored. They are not counted toward the numbers in the table. S = speaking only; WS = written and speaking.

Procedure

Students took one of four experimental test forms randomly assigned across the three countries with the goal to fairly represent participants across the test forms (see Table 2). Participants completed the test in approximately 1 hour under standard testing conditions. The reading section was administered first, and the two listening sections were administered along with the accompanying recordings. Because the test was a paper-and-pencil test, any printed option text was available before, during, and after the audio was presented. A brief postadministration survey asked participants which of the two listening sections (the first or the second) was felt to be easier.

Table 2. Percentage of Participant Distribution by Country

Form	*N*	Brazil	Colombia	Mongolia
1	174	30.5	26.4	43.1
2	187	33.2	26.2	40.6
3	183	38.3	20.2	41.5
4	203	33.0	29.6	37.4
Total	747	33.7	25.7	40.6

For performance data, three test scores including one reading score and two listening scores were recorded for each participant. The two listening scores reflected the option conditions (WS and S). In addition, dummy variables were created to reflect the parameters of each of the four experimental test forms administered to students. Although the order and item sets were counterbalanced by the study design, we considered them as between-subjects factors in our analysis to examine the potential effect on the results. Thus, a dummy variable of *order*

was created to indicate whether a participant's listening score was obtained from the first or second listening section—in other words, to identify which item set was presented first. The other dummy variable of *item set* was used to indicate whether a listening score represented performance on Listening A or Listening B. It should be noted that the current study design was not intended to examine the effect of listening item sets (A and B) because the two item sets were considered parallel. To investigate the effect of listening item sets (A and B) in the current study, it would have been ideal to have groups of students taking both item sets in one of the two conditions. Such design required four additional experimental forms each in the same option conditions (that is, one group taking both Item Sets A and B with the S option condition, another group taking both Item Sets A and B with the WS and S option condition, and two additional groups with Item Sets A and B reversed in order but with the same option conditions). Nevertheless, having the *item set* variable allowed us to compare the mean score of the two item sets within the same option condition, providing the overall difference in the difficulty of the two item sets (University of North Carolina at Chapel Hill, 2015).

Prior to the main analysis, two participants who showed extreme score profiles (i.e., 0/15/9 and 16/11/1 for the reading and two listening scores) were detected. The two participants were also identified as outliers according to the Mahalanobis distance, an index used to detect multivariate outliers (Tabachnick & Fidell, 2013) and were therefore excluded. Analyses were performed on the remaining 745 participants. In addition to listening and reading scores, each participant's age, country, and survey response were available for the analyses.

The general linear model (GLM) was employed to answer the first three research questions. For the first research question, a repeated-measures GLM was performed with the two listening scores as the study design produced two listening scores for each participant. The two listening scores (i.e., WS and S) were specified as dependent variables for a within-subjects factor called *option.* The two dummy variables, called *order* and *item set*, were included as between-subjects factors to examine their potential effects. For the second and third research questions, participants' ages and reading scores were included as covariates to examine how those factors interacted with listening performance given the two option conditions. Lastly, a crosstabulation was used to address the final research question.

Results

The overall descriptive statistics and correlations suggest that the participants' listening performances were comparable between the two option conditions. On average, the listening scores between the two option conditions were not considerably different (Table 3). To understand the effect of the option conditions at the individual level, we computed a score difference between the two listening sections for each participant. The score difference ranged between −8 and 9. A positive difference indicates that participants performed better on the listening section when the option sets were presented both aurally and in writing. A negative difference indicates that the participants did better when the options were read aloud but not printed. The average score difference indicates very little difference in listening performance.

Table 3. Descriptive Statistics of Section Scores and Score Differences

Statistic	WS	S	Reading	Listening score difference (WS – S)
Mean	9.45	9.01	13.66	.44
Median	8.00	8.00	14.00	.00
SD	4.239	4.214	3.860	2.57
Minimum	2	1	3	−8.00
Maximum	17	17	19	9.00

Note. WS = written and spoken; S = spoken.

This result was also supported by a strong correlation between the two listening scores (Table 4). Reading scores also showed a strong correlation with the listening scores, confirming that the reading and listening skills are related. However, the correlation between reading and score differences (WS – S) was very weak (.021), suggesting that the participants' reading ability may not explain the score difference between the two option conditions. Furthermore, the correlation between age and score difference (.008) suggests that performance differences between the two option conditions do not relate to the participants' age.

Table 4. Correlations Among Score and Age (*N* = 745)

Option	WS	S	WS – S	Age	Reading
WS		.815*	.313*	.038	.697*
S			−.296*	.033	.688*
WS – S				.008	.021
Age					.154*

Note. Age was missing for two participants, so correlations are based on 743. WS = written and spoken; S = spoken.
*Correlation is significant at the 0.01 level (2-tailed).

Table 5 presents the mean and standard deviation of section scores by test form. First, similar reading scores across forms indicate that the study samples were comparable with respect to the reading ability. The mean reading score for Form 3 was the lowest, but it was not significantly different from the others. However, the listening scores present a complex picture. To represent the study design shown in Table 1, bold text and shading are shown in Table 5. The shaded cells indicate the listening section in which options were presented both visually and aurally (WS). The listening section that had Item Set A is marked with the bold text. In addition, the last three columns were created to make it easy to examine score differences associated with different conditions of the listening sections. The signs in each of the last three columns denote a score contrast between the conditions. For example, a positive mean difference (>) between the item sets (A – B) indicates that participants, on average, performed better with Item Set A. As shown in Table 5, except for Form 3, there was some variation in the average listening scores between the two listening sections. Mixed results were observed in the direction of the score differences associated with the option conditions (WS – S) and the order (L1 – L2). The signs in the WS – S column indicate a positive effect of having both written and spoken options in Forms 1 and 4 but not in Form 2. Similarly, the average score of the first listening section is higher than that of the second listening section in Forms 1 and 2, but the same pattern is not found in Form 4. Nevertheless, the effect of the item set (A – B) showed more consistent results: Students generally performed better on Item Set A regardless of the option conditions or order of the listening section.

Table 5. Mean and *SD* of the Three Section Scores by Experimental Test Form

Form	*n*	Reading Section	Listening Section 1 (L1)	Listening Section 2 (L2)	WS – S	L1 – L2	A – B
Form 1	172	14.01 (3.78)	**10.34 (4.14)**	8.98 (4.17)	>	>	>
Form 2	187	14.12 (3.46)	**9.56 (4.35)**	8.96 (4.34)	<	>	>
Form 3	183	13.10 (3.93)	8.93 (4.05)	**8.92 (4.09)**	≈	≈	≈
Form 4	203	13.46 (4.15)	8.64(4.21)	**9.62 (4.28)**	>	<	>

Note. WS – S = written and spoken - spoken; A – B = Item Set A - Item Set B. Bolded cell information indicates item set A; shaded cell information indicates WS option sets.

Participants' test scores indicate wide performance differences between the samples from Colombia and from the other two countries (Figure 1). The performances of the student sample from Colombia on the three test sections were considerably higher than those from the other two countries. For example, the median reading score for the students from Columbia was 18 whereas the median reading scores for the students from Brazil and Mongolia were 12 and 13, respectively. The same pattern was observed in the listening sections. However, the performance differences across the countries were not considered an issue in the main analyses because a representation of the three countries was similar across test forms as shown in Table 2.

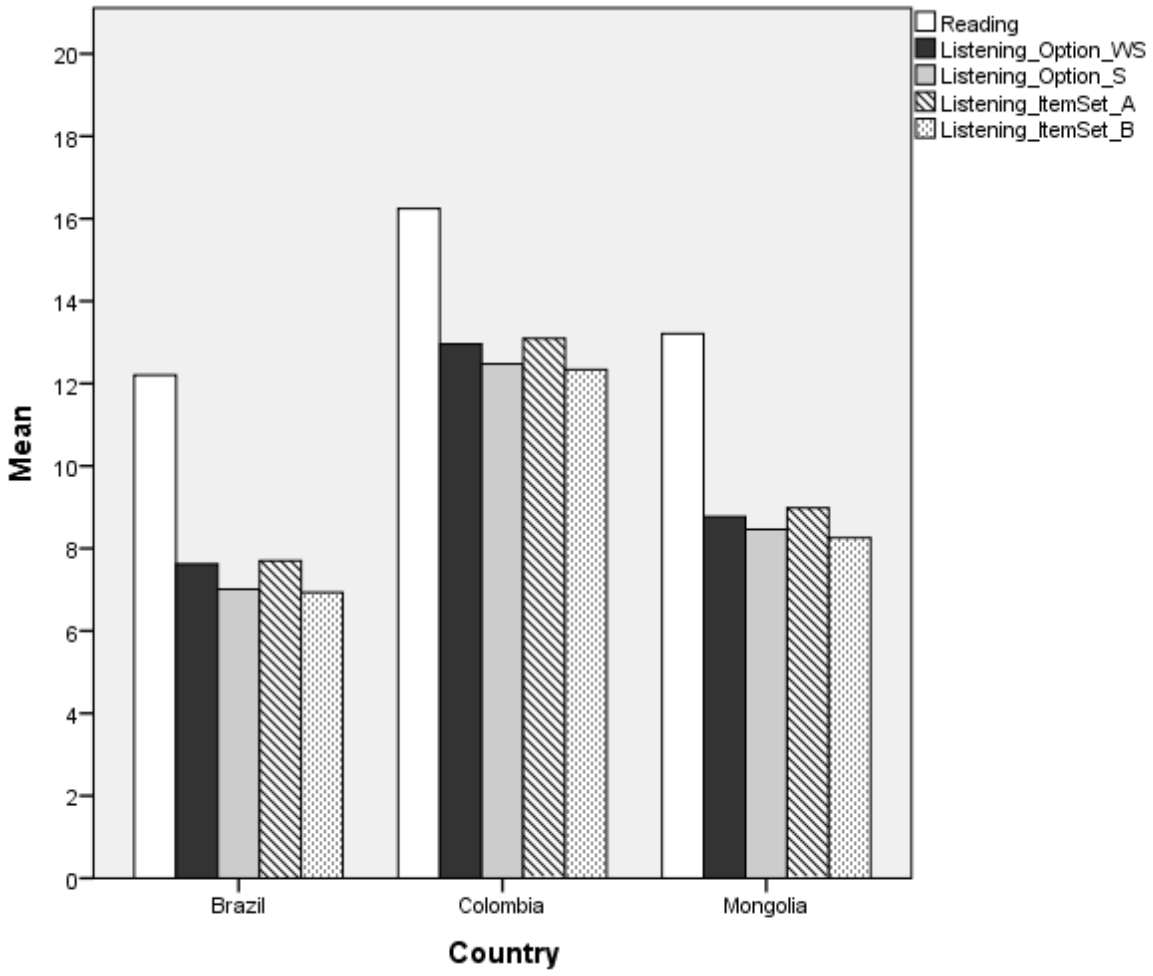

Figure 1. Comparisons of average scores among the countries by test sections.

Reliability of the three test sections was calculated for internal consistency using Cronbach's alpha, resulting in greater than α = .80 for each section, and no items, if excluded, would have raised these values (see Table 6). This information suggests that the items are functioning as expected. Furthermore, the reliability coefficients of subsets according to the option conditions are very similar.

Table 6. Reliabilities of the Different Test Sections

Cronbach's alpha	Reading	Listening A (S and WS)	Listening A (S only)	Listening A (WS only)	Listening B (S and WS)	Listening B (S only)	Listening B (WS only)
α	0.817	0.821	0.816	0.825	0.816	0.815	0.819

Note. S = spoken; WS = written and spoken.

Research Question 1: Does the way in which selected-response listening items are presented (i.e., with or without printed option text) affect the performance of young language learners?

To examine whether the same young EFL learners' listening performance differs between the two option conditions, a repeated-measure GLM was applied to the two listening scores, WS and S. In the analysis, the two listening scores were the within-subjects variables as they were from the same participant. The order and item set were specified as between-subjects factors to allow us to examine the possible interactions of the option conditions with the order (i.e., Option × Order) and item set (i.e., Option × Item Set) in which they appeared in the test booklet. Results of the GLM analysis indicate a significant within-subject main effect for Option F(1, 741) = 23.91, $p < .001$, $\eta_p^2 = .03$ and interaction effects of Option × Order F(1, 741) = 7.70, $p = .006$, $\eta_p^2 = .01$ and Option × Item Set F(1, 741) = 65.91, $p < .001$, $\eta_p^2 = .08$, but all with small effect sizes except for Option × Item Set. The effect of Option × Order × Item Set was nonsignificant [$F(1, 741) = .40$, $p = .525$]. We plotted the two interactions to clarify the results (see Figures 2 and 3). Figure 2 reveals that the young EFL learners' listening scores tended to be slightly higher in the first listening section regardless of the option condition, but these differences are not meaningful as is evident from a visual inspection. The interaction effect of Option × Item Set was moderate ($\eta_p^2 = .08$), indicating the largest amount of systematic variability in the listening scores. The significant effect of item set suggests that the interpretation of the score differences between the two listening sections should be made in light of the differences between the item sets. Figure 3 illustrates that the main effect for option

conditions was observed with Item Set A but not with Item Set B and that Item Set A was generally easier than Item Set B. The average listening scores varied more between the item sets than between the option conditions themselves. Thus, the presence of printed text options did not make items more difficult for test takers.

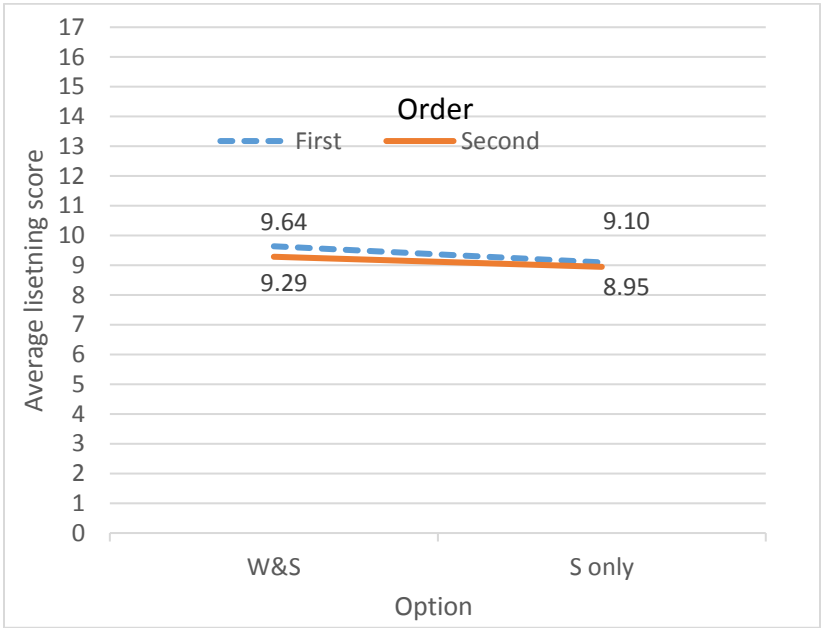

Figure 2. Option Condition × Order Interaction. W&S = written and spoken; S = spoken.

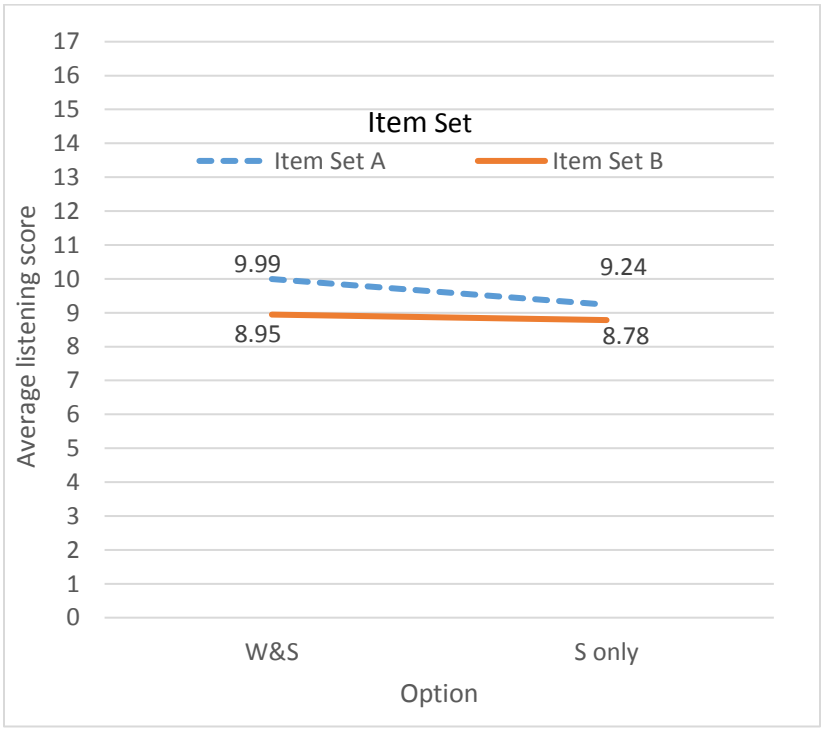

Figure 3. Option Condition × Item Set interaction. W&S = written and spoken; S = spoken.

Research Question 2: What is the relationship between young language learners' age and performance difference between the two conditions?

We were interested in examining if younger participants had greater observed score differences between the two option conditions—that is, do older students have an advantage over younger students with a particular option condition or were younger students more adversely affected by the difference in the option presentation? Thus, the *age* variable was added as a covariate in the analysis. Given the result of the analysis to the first research question, the *item set* factor remained as a between-subjects factor, but the *order* factor was no longer considered in the analysis. Because the relationship between age and score difference is the focus of the second research question, the Option × Age interaction is of primary interest. Results indicate no significant interaction for Option × Age $F(1, 740) = .24$, $p = .621$, $\eta_p^2 = .00$, suggesting that no systematic relationship exists between age and differences in listening scores between the two conditions. As the partial eta squared (η_p^2) refers to the variance explained by a factor, the effect size of 0 for the Option × Age interaction corresponds to the correlation of .01 between age and the score difference presented in Table 4. Therefore, there is little or no relationship between the young learners' ages and the variations in performance between the two option conditions.

Research Question 3: What is the relationship between young language learners' reading ability and performance difference between the two conditions?

The analysis for this research question was analogous to the previous one except that the reading score (*reading*) was used as a covariate instead of age. Results indicate no significant interaction for Option × Reading $F(1, 742) = .26$, $p = .611$, $\eta_p^2 = .00$. Therefore, there is little or no relationship between the participants' reading ability and any difference in performance between the two option conditions.

Research Question 4: How do young language learners perceive the relative difficulty of listening items under the two conditions?

At the end of each test, participants were asked to indicate which of the two listening sections was easier. Because the question instructed the participants to mark the listening sections as they appeared in the test booklet—that is, if the first section was easy, they marked a bubble corresponding to the first listening section on the answer sheet—the participants were not explicitly instructed to compare the difficulty of the listening sections between the option

conditions (WS versus S). Thus, prior to the analysis, the responses to the questions were transformed so that their response indicated a preference for WS or S. Large proportions of the participants (66% to 89%) indicated that they preferred having the WS option sets (see Figure 4). As shown in the pattern in Figure 4, regardless of the difficulty of item sets, many participants preferred the WS option; in Forms 2 and 3, the WS options were used with Item Set B, which was shown to be more difficult than Item Set A, but 66% and 89% still responded that they found the listening section where the options were presented both visually and aurally easier.

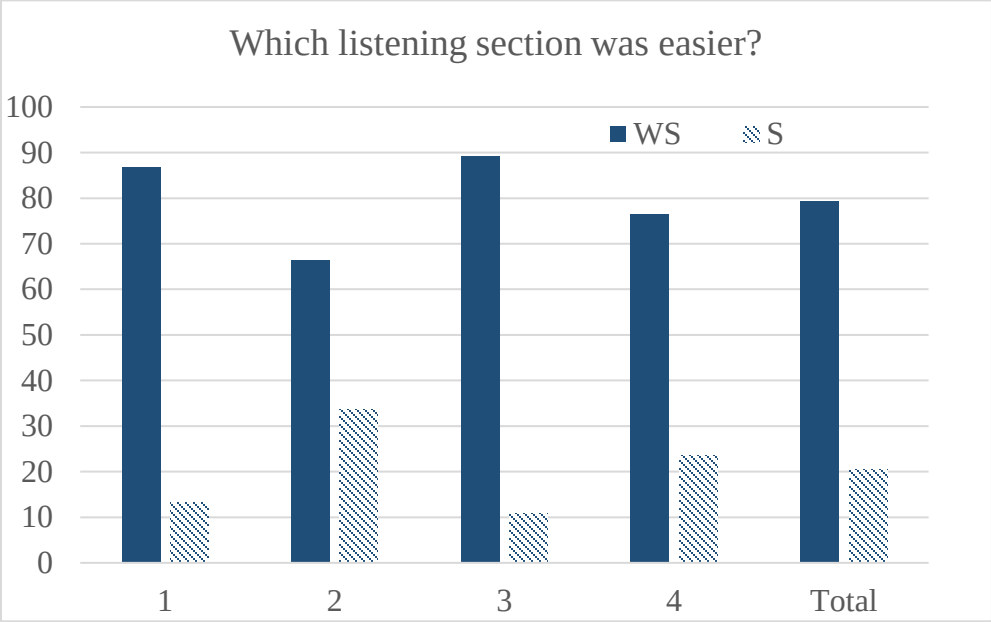

Figure 4. Participant preference for the written and spoken (WS) option versus the spoken (S) option on Test Forms 1–4.

Discussion

The current investigation was motivated in part by our effort to address the anecdotal observations that some young test takers, due perhaps to their age or weak reading ability, may struggle with some of the TOEFL Primary Listening test items in which the options are presented in writing and read aloud at the same time. In other words, it is likely that test takers who can read without difficulty while listening to a question have an advantage over those who have difficulty in reading or are less cognitively developed because the test task is less demanding for the former group. As suggested by some previous evidence (Cho & So, 2014), young students' test performance may be more susceptible to the effects of construct-irrelevant factors. Thus, to evaluate the impact of this potential bias on young test takers' EFL listening

performance, we conducted the quasi-experimental study to compare whether young EFL test takers' performance on the listening test is affected by two different option presentation conditions and whether young EFL learners' age and reading ability can explain performance difference if it exists. We analyzed a sample of 745 EFL students' listening and reading scores, ages, and perceptions from a sample collected from Brazil, Colombia, and Mongolia. We did not observe clear evidence that suggested young test-takers' performance on an EFL test is influenced by the different format of options. Results indicate that there is little difference between the two listening scores and that the variability between the two listening scores for the same participants cannot be explained by the difference in the presentation of options on the test in the current study. Furthermore, the participants' reading ability and age does not explain any score difference in the study sample. Nevertheless, we observed that the majority of young EFL participants expressed a preference for having option sets presented in both written and spoken formats. Although participants' preference did not result in better performance on the listening tests, it confirms the current design of the listening test and provides supporting evidence for the validity of test scores as the format of an item appears to have little influence on performance.

However, the results of the current study should not be taken to preclude the potential impact of EFL test-takers' reading ability and their age on test performance in listening or other language skills without more empirical evidence. Because the TOEFL Primary test is for young EFL learners, the length of options is relatively short in terms of the number of options and the amount of language used. Unlike large standardized assessments, the TOEFL Primary test uses three options instead of four to minimize the effect of memory. An investigation with a different test instrument with longer options may produce different results.

Furthermore, although we did not observe strong evidence to suggest that the presence of written options while the options are read aloud adversely impacts performance on the listening tests, several limitations should be acknowledged. First, the study sample might have been inadequate for detecting the impact of the item format difference. The participants in this sample appeared to be more advanced in their reading abilities than in their listening abilities. They had a higher average score in reading than listening in an absolute sense. Thus, reading and listening to the same information might not have been as much of a problem to many of the participants in the current sample as it could have been to young EFL test takers who have much less developed or no reading ability.

A similar issue can be raised regarding the age represented in the sample. The TOEFL Primary test is designed for EFL learners 8 years of age and older. Research in child development indicates a substantial difference in the cognitive development between an 8-year-old and a 12-year-old (e.g., Anderson & Lajoie, 1996; Fry & Hale, 1996; Martins et al., 2005). Our study sample had a small number of young participants; more than 60% of students were 11 years of age or older. A series of well-controlled studies should be undertaken to more fully understand any relationships between the factors investigated in the current study.

In addition, the findings of the study may not hold with young EFL students in different countries. In this study, because the effect of *country* was not the focus of our interest, we did not investigate whether the results varied across the three countries. However, considering that countries differ in terms of the focus of their EFL instruction and the age of introducing EFL education, different results may emerge across countries.

Finally, the results of the current study should not be interpreted to dismiss the potential effect of these factors on an individual test taker because the study was about examining a general pattern—that is, comparing group-level performances. However, the results are still encouraging in that they suggest that the effect of item format variation seems insignificant for the majority of participants in the study sample. Yet, at the same time, large performance differences were observed for some participants in the data. Our data and analysis cannot tell us why some students showed quite different performances between the two listening sections. As the test was not an official test, some students might have lacked motivation to do their best on the testing instrument used for the study. However, the presence of large score difference for some students, no matter what the reason, suggests that explicit metacognitive instruction and activities (e.g., test-taking strategy, test familiarization) may help mitigate the impact that construct-irrelevant factors may have on young test takers.

Overall, the study did not yield any evidence that presenting option sets to selected-response listening items both aurally and in print made them more or less difficult for young learners than if they were presented aurally only. Also, at least in our sample, age and reading ability did not account for any significant variations in performance between the two option conditions. It is interesting, however, that we found a widespread preference on the part of young test takers to have the option sets available in print in addition to being read aloud. Although more research on assessing young language learners in general and on the effects of different

item presentations in particular is needed, the results of the current study support the practice of presenting option sets to selected-response listening items both aurally and visually as occurs on some of the TOEFL Primary Listening items instead of aurally only.

References

Acha, J. (2009). The effectiveness of multimedia programmes in children's vocabulary learning. *British Journal of Educational Technology*, *40,* 23–31.

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning & Technology*, *5*, 202–232.

Anderson, V. A., & Lajoie, G. (1996). Development of memory and learning skills in school-aged children: A neuropsychological perspective. *Applied Neuropsychology*, *3*, 128–39.

Chang, A. C.-S., & Read, J. (2013). Investigating the effects of multiple-choice listening test items in the oral versus written mode on L2 listeners' performance and perceptions. *System*, *41*, 575–586.

Cho, Y., & So, Y. (2014). *Construct-irrelevant factors influencing young English as a foreign language (EFL) learners' perceptions of test task difficulty* (Research Memorandum No. RM-14-04). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/s/research/pdf/RM-14-04.pdf

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal*, *80*, 183–198.

Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science*, *7*(4), 237–241.

Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing*, *22*, 337–354.

Hemmati, F., & Ghaderi, E. (2014). The effect of four formats of multiple-choice questions on the listening comprehension of EFL learners. *Procedia - Social and Behavioral Sciences*, *98*, 637–644.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*, 219–244.

Jones, L. C., & Plass, J. L. (2002). Supporting listening comprehension and vocabulary acquisition in French with multimedia annotations. *The Modern Language Journal*, *86*, 546–561.

Martins, I. P., Castro-Caldas, A., Townes, B. D., Ferreira, G., Rodrigues, P., Marques, S., . . . Derouen, T. (2005). Age and sex differences in neurobehavioral performance: A study of

Portuguese elementary school children. *International Journal of Neuroscience*, *115*(12), 1687–1709.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology*, *93*, 187–198.

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology*, *90*, 312.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*, 43–52.

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

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology*, *91*, 358.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*, 295–312.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

University of North Carolina at Chapel Hill. (2015, October 27). *Counterbalancing in the design of experiments.* Retrieved from http://www.unc.edu/courses/2008spring/psyc/270/001/counterbalancing.html

Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System*, *36*, 107–122.