

THE EFFECTS OF ITEM PREVIEW ON VIDEO-BASED MULTIPLE-CHOICE LISTENING ASSESSMENTS

Dennis Koyama, New York University

Angela Sun, Stanford University

Gary J. Ockey, Iowa State University

Multiple-choice formats remain a popular design for assessing listening comprehension, yet no consensus has been reached on how multiple-choice formats should be employed. Some researchers argue that test takers must be provided with a preview of the items prior to the input (Buck, 1995; Sherman, 1997); others argue that a preview may decrease the authenticity of the task by changing the way input is processed (Hughes, 2003).

Using stratified random sampling techniques, more and less proficient Japanese university English learners ($N = 206$) were assigned one of three test conditions: preview of question stem and answer options ($n = 67$), preview of question stem only ($n = 70$), and no preview ($n = 69$). A two-way ANOVA, with test condition and listening proficiency level as independent variables and score on the multiple-choice listening test as the dependent variable, indicated that the amount of item preview affected test scores but did not affect high and low proficiency students' scores differently. Item-level analysis identified items that were harder or easier than expected for one or more of the conditions, and the researchers posit three possible sources for these unexpected findings: 1) frequency of options in the input, 2) location of item focus, and 3) presence of organizational markers.

Keywords: Listening, Multimedia, Testing, Video

APA Citation: Koyama, D., Sun, A., & Ockey, G. J. (2016). The effects of item preview on video-based multiple-choice listening assessments. *Language Learning & Technology*, 20(1), 148–165. Retrieved from <http://llt.msu.edu/issues/february2016/koyamasunockey.pdf>

Received: July 21, 2013; **Accepted:** November 29, 2014; **Published:** February 1, 2016

Copyright: © Dennis Koyama, Angela Sun, & Gary J. Ockey

INTRODUCTION

Multiple-choice (MC) response formats have been a popular technique for assessing listening comprehension for decades. Although they have been criticized for not representing natural listening conditions, MC test formats are still the prevailing mode of measuring listening ability (Buck, 2001; Thompson, 1995; Weir, 1990). This is no doubt, in part, because MC tests are a familiar format of assessment for test takers and because the tests can easily be scored by machines or hand-scored by raters with little or no training. Despite the widespread use of MC listening tests, consensus has not been reached on how much information to provide test takers through item preview.

Research investigating the effect of information preview on the listening construct has altered the information preview of items by controlling the amount of information available to test takers (e.g., previews of answer options only or question stem only) and by controlling the number of questions to which test takers have access (e.g., presenting a subset of items or all items at once). Some researchers claim it is essential to have a purpose for listening, so test takers must be provided with a preview of the items prior to listening to the input (Buck, 1995; Sherman, 1997; Thompson, 1995; Vandergrift, 1999). Other researchers contend that allowing test takers an item preview may decrease the authenticity of the task by changing the way test takers process the input (Freedle & Fellbaum, 1987; Hughes, 2003; Wu,

1998). For instance, test takers may focus only on catching key words found in the answer options rather than trying to comprehend the information provided in the input.

Until recently, practical and technological issues have limited what test designers can present to test takers prior to the input. For example, with paper-and-pencil based tests, if a test designer wanted to postpone the presentation of items until after test takers have listened to the input, it was difficult to prevent test takers from turning ahead in the test booklet. It was also problematic to present the answer options one-by-one in audio format because test duration may become extended to the point of making the test administration impractical. Fortunately, technological advances have made it more feasible to choose whether or not to provide test takers with an item preview by controlling what test takers can see while taking a test. One way to do this is through video-based presentation of test items. This form of video-based assessment employs a video or computer screen as a canvas for the listening input and the question stem and answer options. The present study utilizes video-based test presentation to investigate the effects of the amount of item preview on MC listening test performance.

LITERATURE REVIEW

Multiple-choice Listening Assessments

When developing an assessment instrument, a test designer must consider several factors such as the language proficiency of test takers and the effects of using a particular method of assessment (e.g., MC, open-ended format) on the interpretations of test takers' scores (Brown, 2005; Carr, 2011; Fulcher, 2010). This means interpretations of listening test scores are a function of not only the test takers' ability but also the method of assessment and various contextual factors.

To help identify how test takers interact with listening tests, Wu (1998) identified the linguistic (e.g., vocabulary knowledge) and non-linguistic (e.g., world knowledge) test-taking strategies of 10 Chinese EFL learners with "intermediate-level proficiency or above in listening" through immediate retrospection to questions such as "What made you select that option?" (p. 28). Wu reported that some test takers relied on world knowledge and personal impressions to help them select the correct response, which means the key was selected for reasons other than comprehension of the information provided by the aural input. She concluded that the use of MC formats threatens the construct validity of listening assessments because test takers can be (mis)guided by their world knowledge, personal beliefs, and a preview of question stem and options. Hansen and Jensen (1994) also noted how having access to question stems and options can distract test takers from focusing on the input, and that long passages can tax their working memory. Other researchers have noted how test formats might cloud score interpretations, as performance on open-ended questions is poorer than performance on MC questions, suggesting that the recall of information from passages is easier than the generation of accurate information (Brindley, 1998).

Information Preview and Audio-only Listening Assessments

Research on the effects of information preview on L2 listening comprehension has largely been limited to audio-only assessments. This line of research has included investigation of the effects of the distribution of information (e.g., Sherman, 1997), the form of pre-listening support (e.g., Chang & Read, 2006), and the type of available information (e.g., Yanagawa & Green, 2008). The effects of information preview on audio-only forms of listening assessment remain unclear. Sherman (1997) used a Latin square design to investigate the performance of test takers ($N = 78$) on an audio-based listening exam. In her study, the test takers listened to an audio prompt twice and were given listening comprehension questions at three times: (a) before the audio sample played, (b) between the first and second playing of the sample, and (c) after the sample played twice. A fourth condition, the control, had test takers listen to an aural prompt twice without any preview of information, and respond to open-ended questions about the aural input. The Latin square design randomized conditions across test takers to account for exposure and practice effects, and the results showed that test takers scored highest under the "sandwich condition" (p.192) in which the

questions came after the first but before the second playing of the sample. The control condition, no information preview, yielded the lowest scores and was identified in follow-up interviews as the least preferred mode of input. Follow-up interviews and questionnaires also revealed that test takers preferred to preview the question before hearing the audio input because they believed it helped them focus their attention.

Chang and Read (2006) investigated the performance of 160 EFL students on a MC listening comprehension test that utilized four types of listening support: (a) preview of questions, (b) repetition of input, (c) provision of background knowledge, and (d) vocabulary instruction. They found that a preview of the question stem and options before hearing the listening input increased scores of test takers with higher levels of proficiency but not the scores of test takers with lower proficiency. This finding contrasts that of Sherman (1997) who found that low-level learners benefited more from question and options previews compared to high-level learners. While the participants of both studies were EFL learners, there may have been design issues that led to this difference in results. In Chang and Read's (2006) study, all test takers took the same test with differing information preview conditions. In Sherman's (1997) study, each of the four participant groups experienced all conditions of information preview, but they took different versions of a set of tests. This means that while all participants in Sherman's study had the question and options preview, each of the four participant groups was administered a test with a different topic. Since there was no control for test topic and the test forms contained different information, conclusions about the effects of listening preview should be interpreted with caution.

One study that controlled for content and listening ability was conducted by Yanagawa and Green (2008). Yanagawa and Green investigated the effect of having a question stem and options preview, a question stem only preview (i.e., no options), and an options only preview (i.e., no question stem) for 279 TOEIC test takers. Test takers were randomly assigned to one test format, and an ANOVA of TOEIC listening scores verified the language skill equivalence of treatment groups. They found that test takers with a preview of the question and options and test takers with a preview of the question stem only scored significantly higher than test takers with a preview of the options only. No difference was found, however, when scores of test takers with the maximum information preview were compared to the group with the question only preview. Yanagawa and Green (2008) suggested that test takers may (mis)use lexical matching strategies when they become distracted by options that are recognizable as part of the listening input. Thus, depending on the test taker, using lexical matching strategies is not always beneficial to listening scores.

Information Preview with Video-based Listening Assessments

Only one study that investigated video-based listening assessment with information preview could be identified. Berne (1995) manipulated the type of information available to test takers before viewing the video input. Berne investigated the effect of multiple showings of a video prompt across 62 tertiary Spanish learners tested in one of three conditions: (a) preview of the question stem, (b) preview of vocabulary words from the video input, and (c) no listening support at all. The purpose of the presentation of the question was to give test takers some information to organize their thoughts and predict what they may hear in the input (i.e., a top-down approach), while the purpose of giving the vocabulary word list was to provide assistance with lexical familiarity issues that had been identified as a source of listening comprehension difficulties for test takers (i.e., a bottom-up approach). The results showed that the question preview group outperformed the no listening support group, but not the vocabulary preview group. While Berne's (1995) study suggests that access to some information related to the input can facilitate listening comprehension, other research has shown that access to questions while listening does not significantly affect test scores.

Though not directly pertaining to the effects of item preview, Wagner's (2013) study is also relevant to this review. Wagner investigated the effects of test takers' access to the test question while the video or

audio-text was presented. In his study, test takers ($N = 192$) were assigned one of four treatment conditions: (a) video-based input with access to test questions while the stimulus was presented, (b) video-based input without access to the test questions while the stimulus was presented, (c) audio-only input with access to the questions while the stimulus was presented, and (d) audio-only input without access to the test questions while the stimulus was presented. Wagner found that mode of input had a weak ($r^2 = .04$) effect on test taker scores, whereas access to test questions while listening did not significantly affect test scores.

Overall, the research suggests that item preview may increase test scores under certain conditions, but it is not clear which forms of item preview (e.g., question stem and options, question only, options only) benefit test takers most, whether high- and low-level learners are differentially affected by item preview, or (at the item level) which types of items might be affected differently by the amount of information preview. The purpose of this study was to shed some light on these unanswered questions. Our study addressed the following research questions.

RESEARCH QUESTIONS

1. What is the relationship among test condition (preview of question stem and options, preview of question stem only, and no preview), test taker listening ability, and test score?
2. Are any items harder or easier than expected when presented in a condition with a preview of the question and options as compared to a condition with a preview of the question only, and are any items harder or easier than expected when presented in a condition with a preview of the question and options as compared to a condition with no preview?
3. If items are found to be unexpectedly harder or easier depending on the information preview provided, what characteristics might explain these findings?

Because the most common practice has been to present both the question stem and options prior to listening, we compared this condition to the question only and no preview conditions (i.e., research question #2) for the item level analysis. Given that question only and no preview formats are less common, we did not make a comparison between these two formats.

METHODS

Participants

The participants in the study were 206 first-year Japanese university students in Japan who were majoring in English. They were studying in eight different classes at the time of the current study, had intermediate levels of listening ability as measured by scores on the listening section of the TOEFL ITP. Their scores ranged from 38 to 60, with a mean of 48 and a standard deviation of 3.81.

Instrument

Following the advice of Dunkel (1991), a video-based listening test was created to include a variety of listening input. It is important to note that to facilitate ease of reading and to avoid redundancy of terms, the remainder of the current paper will refer to the listening input that corresponds to a set of questions as a scene. When the whole listening test is addressed, the term input will be used to remain congruent with the literature review. With this in mind, the input employed in this study had six loosely related scenes with between six and nine questions per scene for a total of 44 MC items. Three items were not included in the final analysis (two were dropped due to unclear answer options and one was excluded because it was too easy). To avoid confusion, the test is referred to as a 41-item MC test throughout the paper. Two of the scenes were monologues in the form of presentations. The other four scenes were casual conversations discussing situations with which university students are likely to be familiar, such as asking

a classmate about a missed class session and planning a holiday tour. The scenes lasted between 1 min 32 s and 3 min 35 s. They featured actors and actresses who were Australian, British, and American English speakers as well as one highly proficient Japanese speaker of English. A description of the scenes can be found in [Table 1](#).

Table 1. Description of Video Input by Scene

Scene	Length	No. of actors	Type of scene	No. of items
1	1'32"	2F 2M	Self-introductions	9
2	3'00"	1M	Monologue	7
3	2'35"	2M	Conversation	6
4	2'19"	1F 2M	Conversation	6
5	3'35"	1F	Monologue	8
6	3'00"	2F	Conversation	8

Notes. ' = minutes, " = seconds, F = female, M = male

Each of the 41 MC items had four options and included one of two types of listening focus, explicit or implicit (Buck, 2001). Explicit listening foci required test takers to answer questions about specific and local information in the scene (e.g., time, location, or activity of a character). An example of an explicit focus detail question is:

1. What does Amy plan to do on Saturday?

The object of some listening foci in the input was not stated in exact lexical terms but as a synonym. An example of a synonym-based item, noted in italics for the following sentences, is: Amy says, "I'm going to the gym after lunch" and the item reads,

2. What will Amy do *this afternoon*?

The portion of the dialogue "...after lunch" and the portion of the question stem 2 "...this afternoon" are synonymous.

The other type of listening focus was implicit, which required test takers to synthesize information from an entire scene and make generalizations or conclusions about the main idea or the purpose of the scene, or about the attitude or personality of a character. An example of an implicit question is,

3. The most suitable title for this presentation is _____.

Another implicit item type required test takers to make inferences from parts of the scene. For example, in a scene in which the character, Amy, uses phrases like "I got it, Mom" and "Just leave me alone." The item reads,

4. Amy's attitude was _____.

As sample question 4 shows, test takers must infer that Jane was frustrated with her mother from the phrases used throughout the scene. To balance the format of the question types, approximately half of the items were written in the interrogative form, as seen in example questions 1 and 2, while the other half were in the sentence completion, as seen in example question 3 and 4.

Procedure

Students from each of the eight classes were assigned to one of three testing conditions (descriptions are provided in the next section). Stratified random sampling was used to ensure that the three conditions had

test takers with similar listening abilities. Each class had between 24 and 28 students, and these classes were split into thirds with approximately eight students from a given class randomly assigned to all three conditions. To help determine whether this stratified random sampling procedure selected students of similar listening proficiency for each of the conditions, the listening abilities of the three conditions were compared based on their TOEFL ITP listening scores. A one-way ANOVA was conducted using grouping as the independent variable and TOEFL ITP score as the dependent variable. The average listening abilities of the test takers in the three conditions were not found to be significantly different: $F(2, 203) = 1.34, ns$. Although this procedure provided further evidence the three conditions had test takers with similar listening abilities, TOEFL ITP listening scores were not used as a covariate in the study because these scores were obtained approximately six months prior to the other measures used in the study. Although all test takers received the same amount of instruction during the six months, their listening proficiency could not be assumed to have developed at exactly the same rate.

The test was presented on large screen TVs to test takers in each testing room for the three separate conditions. A teacher was present in every testing room to administer the test and ensure that the test followed consistent procedures. The tests were administered over a two week period to the different classes.

Description of Conditions

The test takers in each condition were given a different test booklet according to their input condition. A summary of the test booklet contents of and procedure for each test condition can be seen in [Table 2](#).

Table 2. Description of Formats and Test Booklets for each Condition

Relevant information	Preview condition		
	Q-option	Q-only	No-prev
Items given to student before test	1) Test booklet 2) Picture and name guide for actors	1) Test booklet 2) Picture and name guide for actors	1) Test booklet
Test booklet contents	Question stem, options, and scratch paper	Question stem and scratch paper	Scratch paper only
Information provided in preview	All question stems, options for the given scene, and directions for turning the page for each scene	All question stems and directions for turning the page for each scene, but no options	Directions for turning the page for each scene, but no other information
Booklet sample	1) What time will the train come? a. 11:14 b. 11:40 c. 12:30 d. 12:13	1) What time will the train come?	"This page is for notes"

Description of video shown before input	A timer with 24 s per question was displayed on the screen to indicate time left for question stem and option preview before the scene played.	An adjusted timer with 8 s per question was displayed on the screen to indicate time left for question stem preview before the next scene played. For example, a scene with 5 questions would provide a preview timer of 40 s.	None
Description of video shown after input	A timer was displayed on the screen to indicate time left for answering questions prior the next scene playing. The timer was set to 17 s before a bell sounded twice to indicate that a new item was going to play.	The question stem and options appeared on the screen along with a 24 s timer to indicate time left for a given question. A bell sounded once to alert students to the next question, and another bell sounded twice to indicate a new item was going to play.	The question stem and options appeared on the screen along with a 24 s timer to indicate time left for a given question. A bell sounded once to alert students to the next question, and another bell sounded twice to indicate a new item was going to play.
Summary of sequence for participants	Preview question and options in booklet (8 s per question), view scene and take notes in booklet, choose option from booklet (calculated at 17 s per question)	Preview question in booklet (8 s per question), view scene and take notes in booklet, read question and options on video screen, choose option from video screen (24 s limit per question)	View scene and take notes in booklet, read question and options on video screen, choose option from video screen (24 s limit per question)

Notes. *Q-option* = preview of questions and options, *Q-only* = preview of questions only, *No-prev* = no preview condition.

Test takers in the first condition, referred to as question and options (Q-option), were given a test booklet that included the question stem and options printed on one page for each scene and some scratch paper for taking notes. The Q-option booklet was designed to show all information for only one scene at a time; additionally, a sheet of paper with the names and photographs of the characters in the video was inserted in the test booklet. Test takers in the Q-option condition were given a set amount of time to preview their test booklets before viewing each scene. It is important to note that the amount of time to preview information varied according to the number of questions for a given scene, but the time was determined systematically by allotting 8 s per question (e.g., for a six-question scene, 48 s of preview time was provided before the video played). This time allotment was determined based on time given and shown to be appropriate on previous versions of the video-based test. After viewing each scene, test takers in the Q-option condition had 17 s to select a response to each item, but they were not permitted to turn ahead in the test booklet (e.g., for a six-question scene, 1 min 42 s were provided to answer all six questions). During this time, the video screen showed the phrase “Please answer questions ___ to ___” and a timer counting down the total number of minutes and seconds for this set of questions. Thus, a scene with six questions would have a timer that counted down from “1:42”. After the timer expired, a bell sounded twice to indicate a new passage was going to play and directions to turn the page were displayed on the screen. The question and options were not displayed on the video screen at any time for this condition. Test takers in this condition experienced a traditional listening MC test with maximal amounts of information preview (Buck, 2001; Freedle & Fellbaum, 1987; Rost, 2002), and close invigilation was

required to prevent test takers from looking ahead in the test booklet. With the exception of a countdown timer indicating the amount of time left for an item on the screen and the use of video for the input, there was no technological control of information preview for test takers in this condition.

Test takers in the second condition, the question stem only (Q-only), were given a test booklet that included only the question stems for the scenes and some scratch paper for taking notes. As in the Q-option condition, test takers in Q-only had access to information for only one scene at a time; however, the Q-only condition presented test takers with the question stems and not the options. Test takers had 8 s to preview each question stem before viewing the scene (e.g., for a six-question scene, 48 s of preview time was provided before the video played). After the scene played, the video screen showed one item with the question and its four options as well as the names and the photographs of the relevant characters, and a countdown timer in the bottom right corner of the screen helped the test takers manage their time (see [Appendix](#) for a screen shot of an item). The test takers had 24 s to respond to each item, after which a bell sounded indicating that a new item was to be displayed. After all the questions were sequentially shown, another bell sounded twice to indicate a new scene was going to play and directions to turn the page were displayed on the screen. The design for Q-only demonstrates how technology can control the amount of information available to test takers by limiting them to a question stem only preview. Technology controls testing conditions by preventing test takers from seeing the options, which may otherwise have helped them predict what they would need to listen for (Wu, 1998; Yanagawa & Green, 2008).

Test takers in the third condition, no preview (No-prev), were given a test booklet that included only blank pages for taking notes. These test takers watched each scene having previewed neither the question stems nor the options. After each scene, the video screen displayed each question and its options as well as the names and the photographs of the relevant characters. This screen was identical to the Q-only condition ([Appendix](#)). Also similar to the Q-only condition, there was a 24 s countdown timer and a bell prompting the test takers to watch the screen and move on to the next item. The design for the No-prev condition utilized technology to prevent any information preview, regardless if a test taker were to look ahead in the booklet. On one hand, by removing the preview of the questions prior to listening to the input, listeners would be deprived of a specific purpose for listening (Buck, 1995; Sherman, 1997; Thompson, 1995; Vandergrift, 1999). On the other hand, not providing an item preview might increase the authenticity of the listening task as question and options are not available to interact with test takers' processing of the input (Freedle & Fellbaum, 1987; Hughes, 2003; Wu, 1998; Yanagawa & Green, 2008).

ANALYSIS

To investigate the relationship among test formats (the three conditions), test taker listening ability, and test score, a two-way analysis of variance (ANOVA) was conducted. To prepare the data for the analysis, scores were separated into three listening proficiency groups (high, middle, and low) based on corresponding scores on the TOEFL ITP listening assessment. The middle third was excluded from this part of the analysis to make an extreme groups design. The purpose of using this design was to increase the power in the study for uncovering a possible relationship between test condition and test taker listening ability, since previous research discussed in the literature review had provided mixed findings on this issue. In the two-way ANOVA, test condition (Q-option, $n = 47$; Q-only, $n = 48$; and No-prev, $n = 43$) and test taker listening ability (based on upper one-third and lower one-third of TOEFL ITP listening scores) were used as independent variables, and score on the MC listening test was used as the dependent variable.

While a test-level analysis is important for understanding general effects of a set of items on scores, it cannot provide a clear indication of the effects of certain item types on scores. Thus, an item level analysis was also conducted. For this analysis, an item response theory (IRT) approach was used as implemented in Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 2003). The procedure was used to

identify items that were unexpectedly harder or easier when comparing the commonly encountered Q-option test condition ($n = 67$) and the Q-only condition ($n = 70$), and when comparing the Q-option condition and the No-prev condition ($n = 69$). Difficulty of items was not compared between the less commonly encountered Q-only and No-prev conditions. All test takers were included in this analysis ($N = 206$).

Item response theory (IRT) is an approach that links observed performance, in this case, scores on the MC listening test, to a location on an underlying latent continuum of ability, in this case, listening ability. IRT is generally believed to provide more accurate estimates of listening ability than raw test scores (e.g., Embretson & Reise, 2000). IRT uses a logit measurement scale that is set to have a center point of zero. Items with positive logits are more difficult than average while items with negative logits are easier than average (Ockey, 2012). A 1-parameter logistic model (de Ayala, 2009) was selected, and only item difficulty was modeled in the analysis. We used this simplest IRT model because the number of participants in the study was too small for a more complex IRT model. Simpler models require fewer participants because there are fewer parameters to estimate (Ockey, 2012). A principal components analysis to assess the IRT assumption of unidimensionality suggested that the data set was appropriate for the analysis. However, it should be noted that while IRT is commonly used with passage-based inputs which use more than one item, such data may not completely satisfy the IRT assumption of local independence (So, 2010).

To determine if certain item types were more difficult than expected for one or more of the formats, we used differential item functioning, a procedure for identifying items that are harder or easier for particular test taker conditions, as implemented in BILOG MG (see Geranpayeh & Kunnan, 2007; Ockey, 2007 for examples in applied linguistics). This procedure can also be used to compare item difficulty across test formats (Embretson & Reise, 2000). In the first step, the overall effects of condition are estimated. In the second step, the items in each condition are adjusted for this overall effect. In step three, the difficulty of these items is compared to see if any is easier or harder than expected after accounting for a condition effect at the test level¹ (du Toit, 2003; Embretson & Reise, 2000). Step 1 indicated that at the test level, that is, when averaged across items, the Q-only condition was 0.19 logits more difficult than Q-option condition, and the No-prev condition was 0.81 logits more difficult than Q-option condition. In step 2, BILOG MG adjusted the difficulty of each of the items to make them comparable across formats. Thus, 0.19 logits was subtracted from each of the 41-item scores of the test takers in the Q-only condition, and 0.81 logits was subtracted from each of the 41-item scores for those in the No-prev condition. In step 3 of the process, these adjusted difficulties were compared for each item between the scores of the test takers who took the Q-option and Q-only condition and the scores test takers who took the Q-option and No-prev condition. After these adjustments, items found to be significantly more or less difficult than expected were identified.

To determine possible explanations for the unexpected level of difficulty of items, a micro-level item content analysis was conducted. Two researchers carefully analyzed the video input by cross-referencing the scripts used by the actors to the dialogue in the video. The researchers worked together and examined the eight items for (a) explicit or implicit focus (inclusively called the "item focus"; Buck, 2001); (b) exact match, synonymous match, or no match between words in the item and words spoken in the scene; (c) frequency of words in the question stem or options spoken by characters in the scene; (d) location of item focus in the scene measured by minutes and seconds elapsed; and (e) use of organizational markers by characters in the scene.

RESULTS AND DISCUSSION

The descriptive statistics and estimates of reliability (based on Cronbach's Alpha) for the three test conditions are presented in Table 3. As can be seen, reliability estimates were all near .70, suggesting that all conditions similarly distinguished test takers on their listening abilities. One item which had a

difficulty of .98 and a point-biserial of .05 was excluded from the analysis based on the reliability analysis (leaving 41 items in the final data set). After excluding this item, difficulty ranged from .26 to .95. Items were spread throughout the difficulty range, although the average of .75 suggested the test was easy for the students and likely resulted in the somewhat marginal reliabilities of the test conditions. Classical Test Theory point-biserials were between .08 and .45, with the large majority near or above .30. Analysis indicated that excluding one or more of the items with low point-biserials had minimal impact on the reliability, so no other items were excluded from the analysis.

Table 3 Descriptive Statistics and Estimates of Reliability

	TOEFL listening	All conditions	Q-option	Q-only	No-prev
Number of students	206	206	67	70	69
Reliability	-	0.69	0.68	0.71	0.69
Mean	47.54	29.93	31.27	30.43	28.13
Standard deviation	3.81	4.58	4.29	4.63	4.27
Skewness	0.30	-0.15	-0.44	0.39	-0.01
Kurtosis	0.75	-0.57	-0.42	0.71	-0.07

Note. Reliability could not be calculated for the TOEFL listening test because only total scores were available.

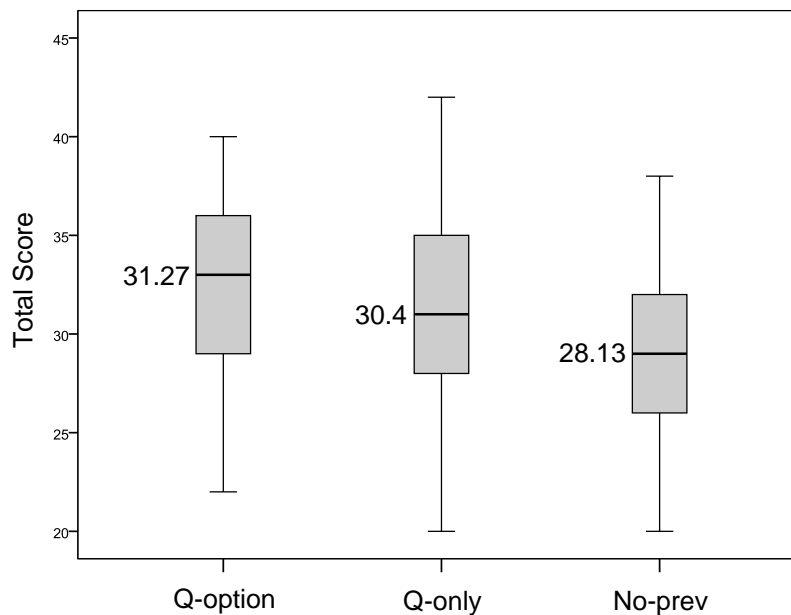


Figure 1. Boxplots of scores by amount of information preview with median indicated by bold line and the mean displayed next to each boxplot.

As can be seen in [Table 3](#), the overall mean on the MC listening test was 29.93 out of 41 with a standard deviation of 4.58. The means and standard deviations of the Q-option, Q-only, and No-prev were 31.27, 30.43, and 28.13, respectively, indicating that as more preview was available to the test takers, the easier the test became. This result is presented in boxplots (see [Figure 1](#)); for each condition, the median is indicated by a bold line and the mean has been provided in text format. Skewness and kurtosis values were all indicative of a normal distribution. Skewness values were less than the absolute value of 1, and

kurtosis values were all less than the absolute value of 3, suggesting that the assumption of normality was tenable (Kline, 2005), and therefore the data were appropriate for ANOVA procedures. Levine's test indicated that the assumption of homogeneity of variance was also met for each ANOVA procedure.

Test Level Analyses

To determine the extent to which condition, test taker listening ability, and test score were related, a two-way ANOVA, as implemented in PASW 18, was conducted with test condition (Q-option, Q-only, and No-prev) and test taker listening ability (as measured by TOEFL ITP) as independent variables, and the score on the MC listening assessment as the dependent variable. Not surprisingly, there was a statistically significant main effect for test taker listening ability, $F(1, 132) = 71.8, p < .05, \eta^2 = 0.35$, indicating that test takers with higher listening ability performed significantly better than those with lower listening ability. This finding provides some indication that the test, independent of condition, was a valid indicator of listening ability.

An effect was also found for test condition, $F(2, 132) = 71.8, p < .05, \eta^2 = 0.10$, indicating that 10% of the variance in test scores could be attributed to the condition the test takers were assigned. We did not, however, find a significant interaction between test condition and test taker listening ability, $F(2, 132) = 1.3, ns$, suggesting that test condition did not affect high- and low-level test takers differently. That is, test takers at different ability levels were equally impacted by their test condition. This result does not imply, however, that no test taker was more or less affected by one of the conditions; it only indicates that on average more and less proficient test takers were not impacted differently by test conditions. These findings do not support Chang and Read's (2006) conclusion that test takers with high levels of listening ability benefited more from Q-option than those with low levels of listening ability. Neither do the findings support those of Sherman (1997), who noted that more information benefited lower ability test takers.

A possible explanation for these contradictory findings is that some item types affect test takers with certain ability levels differently while other item types do not. Such an effect would only manifest itself at the test level when a sufficient number of items that have such an impact are present. This suggests the importance of investigating specific item types to determine their effects on high- and low-level test takers. Given that our findings contradict those of Chang and Read (2006) and Sherman (1997), further research is recommended before such procedures for adapting an assessment are used.

Because scores were not found to be significantly associated with test condition, the complete data set was analyzed with test condition (Q-option, Q-only, and No-prev) as an independent variable and test score as the dependent variable in a one-way ANOVA. The results further confirmed that test condition led to significantly different scores on the test, $F(2, 203) = 9.30, p < .05, \eta^2 = 0.09$. Tukey's HSD post hoc analysis indicated that the Q-option condition was significantly easier than the No-prev condition and that the Q-only condition was significantly easier than the No-prev condition. Scores on the Q-option and Q-only conditions were not found to be significantly different.

Item Level Analyses

Table 4 presents the results for percent correct on each of the conditions. Based on the test level analyses, the percent correct for the Q-option condition would be expected to be somewhat, although not significantly, higher than the Q-only condition, which would be somewhat higher than the No-prev condition. Items that did not follow this pattern were identified in the IRT analysis as unexpectedly harder or easier across the test conditions.

Table 4 Percent Correct on the Test for each Condition

Item #	Q-option	Q-only	No-prev	Item #	Q-option	Q-only	No-prev
1	66%	57%	64%	21	91%	94%	86%
2	93%	93%	86%	22	76%	64%	57%
3	70%	71%	45%	23	27%	30%	20%
4	87%	81%	80%	24	27%	26%	26%
5	87%	86%	77%	25	81%	81%	78%
6	84%	79%	74%	26	93%	99%	94%
7	81%	79%	64%	27	91%	83%	58%
8	75%	73%	62%	28	96%	97%	90%
9	72%	84%	94%	29	82%	69%	51%
10	85%	84%	93%	30	94%	94%	97%
11	84%	91%	86%	31	70%	71%	60%
12	93%	89%	86%	32	57%	50%	48%
13	91%	91%	88%	33	73%	61%	49%
14	61%	69%	74%	34	70%	61%	65%
15	52%	59%	51%	35	73%	74%	54%
16	88%	77%	78%	36	87%	76%	81%
17	73%	83%	88%	37	96%	84%	78%
18	93%	86%	90%	38	96%	77%	64%
19	82%	83%	75%	39	61%	50%	39%
20	94%	97%	87%	40	31%	34%	30%
				41	49%	54%	48%

IRT estimates for Q-option and Q-only as well as between Q-option and No-prev are presented in Table 5. The IRT adjusted item difficulty differences, the standard errors (SE) for these difference estimates, and the difficulty differences divided by their SEs are provided. Asterisks are used to indicate the items that are unexpectedly harder or easier for one of the conditions, after adjusting for expected differences due to condition ($\alpha < .05$, indicated here by a z-score with an absolute value of 1.96 or greater).

Table 5 Comparison of Scores Between Q-option v/s Q-only and Q-option v/s No-prev

Item	Difference between Q-option and Q-only			Difference between Q-option and No-prev		
	Adjusted difficulty difference	SE of difference	Difficulty difference/SE	Adjusted difficulty difference	SE of difference	Difficulty difference/SE
1	0.47	0.60	0.78	-0.64	0.60	-1.07
2	-0.26	1.09	-0.24	0.53	0.95	0.56
3	-0.29	0.66	-0.44	1.08	0.64	1.69
4	0.49	0.80	0.61	0.11	0.80	0.14
5	-0.06	0.84	-0.07	0.40	0.78	0.51

6	0.39	0.72	0.54	0.26	0.71	0.37
7	0.04	0.74	0.05	0.74	0.70	1.06
8	-0.02	0.69	-0.03	0.28	0.66	0.42
9	-1.50	0.72	*-2.08	-3.92	0.98	*-4.00
10	-0.08	0.80	-0.10	-2.10	0.96	*-2.19
11	-1.45	0.93	-1.56	-0.99	0.80	-1.24
12	0.62	1.03	0.60	0.53	0.99	0.54
13	-0.26	0.99	-0.26	-0.24	0.94	-0.26
14	-0.76	0.65	-1.17	-1.81	0.65	*-2.78
15	-0.64	0.6	-1.07	-0.70	0.60	-1.17
16	1.17	0.82	1.43	0.49	0.82	0.60
17	-1.19	0.70	-1.70	-2.52	0.78	*-3.23
18	1.06	1.01	1.05	-0.15	1.05	-0.14
19	-0.27	0.78	-0.35	-0.05	0.73	-0.07
20	-1.47	1.47	-1.00	0.73	1.06	0.69
21	1.00	1.14	0.88	0.20	0.93	0.22
22	0.83	0.65	1.28	0.80	0.63	1.27
23	-0.42	0.66	-0.64	-0.21	0.69	-0.30
24	-0.04	0.65	-0.06	-0.78	0.65	-1.20
25	-0.28	0.77	-0.36	-0.5	0.74	-0.68
26	-3.04	1.84	-1.65	-1.17	1.17	-1.00
27	1.10	0.92	1.20	2.70	0.85	*3.18
28	-0.96	1.56	-0.62	0.76	1.20	0.63
29	1.12	0.68	1.65	1.80	0.68	*2.65
30	-0.26	1.23	-0.21	-1.97	1.47	-1.34
31	-0.29	0.65	-0.45	0.05	0.61	0.08
32	0.31	0.61	0.51	-0.17	0.61	-0.28
33	0.77	0.67	1.15	1.03	0.64	1.61
34	0.51	0.62	0.82	-0.38	0.63	-0.60
35	-0.29	0.68	-0.43	0.73	0.65	1.12
36	1.08	0.77	1.40	-0.05	0.80	-0.06
37	2.17	1.11	*1.96	2.29	1.08	*2.12
38	2.98	1.07	*2.79	3.53	1.05	*3.36
39	0.64	0.63	1.02	0.78	0.62	1.26
40	-0.39	0.63	-0.62	-0.77	0.65	-1.18
41	-0.54	0.58	-0.93	-0.71	0.58	-1.22

Note: * indicates significant at .05 level.

Items with negative values indicate that the test condition that was expected to be easier was not. For

instance, for item 9, it can be seen in Table 4 that a larger percentage of test takers in the No-prev condition answered correctly than those in the Q-only condition and a larger percentage of test takers in the Q-only condition answered correctly than those in the Q-option condition. The asterisk by item 9 in Table 5 indicates that this effect was significant. An asterisk by a positive number indicates an opposite effect, in which the item was significantly easier than expected for the Q-option condition as compared to the Q-only condition or as compared to the No-prev condition. An example is item 38 for which the differences in the percentage of test takers that answered the item correctly on the conditions were larger than expected.

Table 6. Breakdown of Items that were Harder or Easier than Expected for a Condition by Scene, Distractor, and Key with Rationale for Behavior

Scene	Item	Key type	Multiple mention of stem	Distractors mentioned (3 maximum)	Multiple mention of key	Possible reason for behavior
2	9	Time	Yes	3	Yes	Frequency of focus
2	10	Location	Yes	3	No	Key after focus
2	14	Object	No	2	No	Key after focus
3	17	Inference (feeling)	No	1	No	Not watching screen
5	27	Time	No	0	No	Timing in script
5	29	Activity	No	2	Yes	Organizational marker
6	37	Location	No	1	No	Temporal marker
6	38	Activity	No	2	No	Temporal marker

As can be seen in Table 6, eight items were found to be unexpectedly harder or easier, after adjusting for test conditions, for at least one of the test conditions. An item-level analysis was performed and three possible reasons were found which may explain why the test takers performed differently under the three conditions. The three factors were (a) the frequency of mention of the options, (b) the placement of the item focus, and (c) the presence of organizational markers in the listening scene. The frequency of mention refers to the number of times words in the option were spoken by the characters in the scene. If a character mentioned an option more than twice, it was marked as frequently mentioned. For instance, if the option is "backpack" and a character says the word "backpack" three times in the scene, the option was coded as "frequently mentioned". The placement of the item focus indicates the position of the necessary information to answer an item in the script. In other words, it refers to whether the character spoke the line containing the necessary information to answer a given question at the beginning, in the middle, or at the end of the scene. Using video editing software, this was marked by the minutes and seconds elapsed from the beginning of the passage to the first time words pertaining to the item were spoken by a character. Organizational markers refer to phrases that signal the organizational pattern of the passage, such as "first," "second," and "to conclude." Four of the items and possible reasons for the differences in test takers' performance will be discussed. The other four items share the same three factors and possible reasons for differences in test taker performance; therefore, these items will not be further explained.

The first factor that may have affected test takers' performance across the three test conditions is the frequency of option mentions in the listening scene. The effect of this factor can be seen in item 9, which corresponded to a monologue in which a school official presented information to new students during an

orientation session. Item 9 is an example of explicit focus items for which options were frequently mentioned in the scene in the form of exact lexical overlapping or of synonyms (e.g., example question 2 from above). On item 9, the Q-only and the No-prev test takers performed better than expected when compared to the Q-option test takers. This may have occurred because incorrect options, which were only continuously visible in the Q-option condition, were mentioned multiple times throughout the scene. Thus, the test takers who had previewed the question stem and options may have relied on a lexical matching strategy of listening for key words or phrases, thereby choosing a frequently mentioned option. This finding is analogous to the findings of Wu (1998) and Yanagawa and Green (2008).

The second factor that could have caused items to behave differently is the placement of the item focus in the input. Item 27 is the first item corresponding to a scene that was an academic presentation on an eco-village. This item required test takers to select the season during which the character visited an eco-village. The Q-option condition scored higher than expected when compared to the No-prev condition on this item. This may have occurred because the focus of the item appeared relatively early in the listening scene (approximately 6 s into the monologue) and was not frequently mentioned. The No-prev test takers might have been negatively affected because they were not prepared to answer this item, whereas the Q-option test takers may have been prepared to listen for language related to seasons of the year. It is possible that the No-prev test takers did not pay much attention or did not take detailed notes while listening to this early part of the presentation.

Items 29 and 37 illustrate how the presence of organizational markers in the listening scene may have helped the Q-option group score higher than expected as compared to the No-preview group. Item 29, which corresponded to the eco-village presentation described above, had a focus related to two main ideas in the presentation. In the listening scene, the character stated, "There are two things that are very important to people living in an eco-village: one is making their own food, and the other is not wasting resources like water." Because the character used organizational markers such as "there are two," "one is," and "the other is" to introduce these main ideas, there was strong lexical overlap between the scene and the question stem. This may have caused test takers in the Q-options condition to score higher than expected when compared to test takers in the No-prev condition. These findings are in line with those of Chaudron (1983) and other researchers (e.g., Freedle & Fellbaum, 1987; Vandergrift, 1999; Wu, 1998).

Similar to item 29, item 37 corresponded to a scene in which a character used an organizational marker. The scene is a dialogue between two friends discussing several activities to do during a holiday tour, and one of the characters stated, "On day one, we can...". The temporal marker "day one" may have aided the Q-option test takers, who performed better than expected when compared to both the Q-only and the No-prev test takers. The Q-only condition had a preview of the question stems before listening to the scene, but they were not able to answer as accurately as the Q-option group. This was most likely because the characters discussed several activities to take place on "day one" and the Q-only group could only rely on their notes and memory, which may not have included the detail that was the key.

CONCLUSIONS

We found that presenting a preview of the question stems and options or question stem only made the test less difficult than presenting no preview at all. Our research did not support the contention that test takers with different English proficiencies were differentially impacted by information preview. We did find that some items were harder or easier than expected dependent upon the condition and the interaction between the content of the input and the item type. Based on an item level content analysis, we hypothesized that frequency of mention of the options, the placement of the item focus, and the presence of organizational markers in the listening input may be factors that affect the differential difficulty of the items across conditions.

After taking into account the effect of test conditions, the result that some types of items are harder or

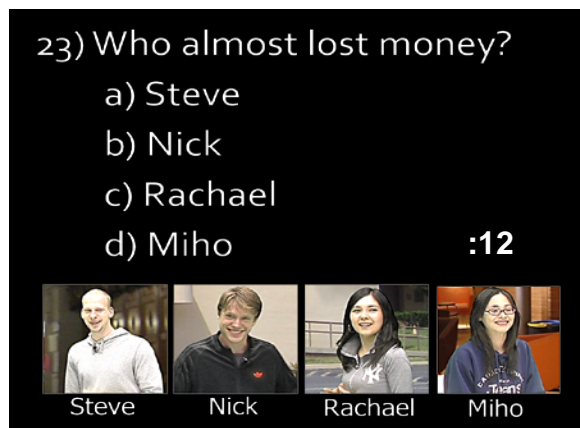
easier than expected across the three test conditions provides useful information for test designers and test users as they consider the test format to be used and as they carefully define the construct they wish to measure. For instance, if item writers aim to assess general listening comprehension, that is, to avoid word-level comprehension or lexical matching, it may be useful to construct a test using the No-prev format. On the other hand, if the aim is to assess specific word-level comprehension, it may be more appropriate to provide an item preview. Assessing a broader construct that includes both of these abilities may require the use of more than one of the conditions described in this study. Thus, the decision about whether or not to provide an item preview should be made based on the construct that the test developer aims to assess.

Limitations of the study should be noted. The test taker population was a fairly homogeneous group of Japanese EFL learners, and consequently the results may not generalize well to other populations. In addition, the sample size was rather small (N=206) for an IRT analysis, and because there were multiple items used to assess comprehension of one scene, the IRT assumption of item independence may not have been completely satisfied in the study (So, 2010).

The findings of our study suggest that test developers should carefully consider how much information about items they want to provide to test takers prior to listening because the amount and form of information preview might affect which items students get right and which items they get wrong. These differences in difficulty can be attributed to various content aspects of the items, indicating the importance of considering these factors when constructing items for listening tests. Building on the findings of this study, it might be informative to construct an empirical study that specifically investigates the frequency of mention—and perhaps type of mention (e.g., exact mention or synonym)—and test form (e.g., amount of information preview available to test takers). A carefully designed study would also consider the effects of the placement of the item focus (i.e., its location in the listening input) in relationship to the frequency of mention. For example, a study could compare the assessment of content mentioned frequently in the first half of the listening input to content mentioned in the last half of the listening input. It might be informative to understand the influence of grammatical structures on the assessment of the listening construct. To that end, another follow-up study might consider the effects of manipulating syntactic structures to include specific grammatical forms such as passives, anaphors, and relative clauses in the input versus in the item options.

APPENDIX

Sample screen shot of listening item #23 with question stem, options, graphic support, and onscreen timer.



23) Who almost lost money?

- a) Steve
- b) Nick
- c) Rachael
- d) Miho

:12

Steve Nick Rachael Miho

NOTE

1. This procedure places the three groups on a common logit scale by estimating the item parameters for the three groups simultaneously, making it possible to compare scores across formats (Embretson & Reise, 2000, p. 261.)
-

ACKNOWLEDGEMENTS

This research was graciously supported by the Sano Foundation and the English Language Institute at Kanda University of International Studies. We would like to acknowledge John M. Norris and Larry Davis for their comments and input on the earliest version of this paper, which was presented at the 2011 Language Testing Research Colloquium held in Ann Arbor, Michigan. Thanks are also given to the students of the seminar course on Publication in Second Language Studies at Purdue University for their feedback on an earlier draft of this paper, especially to Ghada Gherwash for her useful suggestions. We thank the anonymous reviewers for their feedback and suggestions. Any errors that remain are entirely ours.

REFERENCES

- Berne, J. E. (1995). How does varying pre-listening activities affect second language listening comprehension?. *Hispania*, 78(2), 316–329.
- Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, 18, 171–191.
- Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York: McGraw-Hill.
- Buck, G. (1995). How to become a good listening teacher. In D. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 113–131). San Diego: Dominie.
- Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.
- Carr, N. T. (2011). *Designing and analyzing language tests*. New York: Oxford University.
- Chang, A. C. S., & Read, J. (2006). The effects of listening support on the listening performance of ELF learners. *TESOL Quarterly*, 40(2), 375–397.
- Chaudron, C. (1983). Simplification of input: Topic reinstatements and their effects on L2 learners' recognition and recall. *TESOL Quarterly*, 17(3), 437–458.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford.
- du Toit, M. (2003). *IRT from SSI: Bilog-MG, multilog, parscale, testfact*. Skokie, IL: Scientific Software International.
- Dunkel, P. (1991). Computerized testing of nonparticipatory L2 listening comprehension proficiency: An ESL prototype development effort. *Modern Language Journal*, 75(1), 64–73.
- Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. London: Lawrence Erlbaum.
- Freedle, R., & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In R. Freedle & R. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 162–192). Norwood, NJ: Ablex.
- Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

- Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the Certificate in Advanced English Examination. *Language Assessment Quarterly*, 4(2), 190–222.
- Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 241–268). Cambridge, MA: Cambridge University.
- Hughes, A. (2003). *Testing for language teachers*. Cambridge, MA: Cambridge University.
- Kline, R. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.
- Ockey, G. J. (2007). Investigating the validity of math word problems for English language learners with DIF. *Language Assessment Quarterly*, 4(2), 149–164.
- Ockey, G. J. (2012). Item response theory. In G. Fulcher & F. Davidson (Eds.), *Routledge handbook of language testing* (pp. 316–328). New York: Routledge.
- Rost, M. (2002). *Teaching and researching listening*. London, UK: Pearson Education.
- Sherman, J. (1997). The effect of question preview in listening comprehension tests. *Language Testing*, 14(2), 185–213.
- So, Y. (2010). *Dimensionality of responses to a reading comprehension assessment and its implications to scoring test takers on their reading proficiency* (Unpublished doctoral dissertation). University of California, Los Angeles. Los Angeles, CA.
- Thompson, I. (1995). Assessment of second/foreign language listening comprehension. In: D. Mendelsohn & J. Rubin (Eds.), *A Guide for the teaching of second language listening* (pp. 31–58). San Diego: Dominie.
- Vandergrift, L. (1999). Facilitating second language listening comprehension: Acquiring successful strategies. *ELT Journal*, 53(3), 168–176.
- Wagner, E. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Language Assessment Quarterly*, 10(2), 178–195.
- Weir, C. (1990). *Communicative language testing*. New York: Prentice Hall.
- Wu, Y. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, 15(1), 21–44.
- 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(1), 107–122.
- Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). Bilog-MG (Version 3.0.2327.2) [Computer software]. Mooresville, IN: Scientific Software.