Language Assessment Quarterly

ISSN: 1543-4303 (Print) 1543-4311 (Online) Journal homepage: https://www.tandfonline.com/loi/hlaq20

The Relative Difficulty of Dialogic and Monologic Input in a Second-Language Listening Comprehension Test

Spiros Papageorgiou , Robin Stevens & Sarah Goodwin

To cite this article: Spiros Papageorgiou , Robin Stevens & Sarah Goodwin (2012) The Relative Difficulty of Dialogic and Monologic Input in a Second-Language Listening Comprehension Test, Language Assessment Quarterly, 9:4, 375-397, DOI: 10.1080/15434303.2012.721425

To link to this article: https://doi.org/10.1080/15434303.2012.721425

Language Assessment Quarterly, 9: 375–397, 2012

ISSN: 1543-4303 print / 1543-4311 online
DOI: 10.1080/15434303.2012.721425

Routledge
Taylor & Francis Group

The Relative Difficulty of Dialogic and Monologic Input in a Second-Language Listening Comprehension Test

Spiros Papageorgiou
Educational Testing Service

Robin Stevens
Lidget Green Inc.

Sarah Goodwin
Georgia State University

Listening comprehension tests typically include both monologic and dialogic input to measure listening ability. However, research as to which type of input is more challenging for examinees remains limited and has provided inconclusive results (Brindley & Slatyer, 2002; Read, 2002; Shohamy & Inbar, 1991). A better understanding of the comparative difficulty of items associated with both input types is important, as it has implications for developing test content at the desired levels of difficulty. This study explores this issue by analyzing examinee performance on test items developed to accompany three pairs of stimuli on the same topic. Each pair of stimuli consists of a monologue and a dialogue with identical content and vocabulary. The test items associated with these stimuli were embedded in 3 test forms taken by 494 examinees as a part of a routine administration of the Michigan English Test. Test results were analyzed with the Rasch computer program WINSTEPS (Linacre, 2009) to investigate the relative difficulty of the items associated with the two versions of the input and the measurement characteristics of the item options. To interpret statistical findings, a content analysis of the stimuli and items was also performed. Findings provide partial support to the hypothesis that items associated with dialogic input may be easier for examinees than the same items associated with identical monologic input. The implications of these findings for developers and users of listening comprehension tests are discussed.

INTRODUCTION

To support valid inferences about examinee ability in relation to a specific target language use (TLU) domain (Bachman & Palmer, 1996, pp. 44–45), developers of second-language listening comprehension tests must select audio input representative of typical target domain tasks. For example, listening tests for college admission typically include input aligned to the types of tasks

Correspondence should be sent to Spiros Papageorgiou, Educational Testing Service, 660 Rosedale Road, MS 04 R, Princeton, NJ 08541. E-mail: spapageorgiou@ets.org

that learners will perform in the higher education domain, such as listening to lectures or participating in a conversation with a professor on an academic topic. Listening comprehension test input generally comprises two types of stimuli. The first type is monologic discourse, which consists of a single speaker giving a talk or presentation. The second type is dialogic, with two or more speakers participating in a conversation.

Research by language testers as to whether monologic or dialogic input is more difficult for examinees to comprehend is limited and has provided inconclusive results (Brindley & Slatyer, 2002; Read, 2002; Shohamy & Inbar, 1991). Therefore, a better understanding of the comparative difficulty of monologic versus dialogic input is important, as it has implications for developers and users of listening comprehension tests, in particular with regard to test content and intended test difficulty level. The study reported in this article attempts to shed light upon this issue by analyzing examinee performance on three pairs of stimuli that were adapted to be parallel and comparable. Each pair of stimuli consists of a monologue and a dialogue containing the same content and vocabulary. Three identical sets of test items were developed and administered to examinees with the stimuli. A review of the literature related to monologic and dialogic discourse follows.

LITERATURE REVIEW

In a discussion of the practical issues related to the creation of listening test recordings, Buck (2001, pp. 165–166) noted that one drawback of monologic speech is that it contains few features that are characteristic of interactive discourse; back-channels and listener feedback are absent. However, discourse researchers point out that monologues do not completely lack interactivity. Similar to participants in a conversation, most lecture presenters do not simply tell their thoughts to an audience without consideration of the listener. Rather, they monitor their speech and observe whether the audience comprehends what they are saying[1] (Fox Tree, 1999). Even when solo speakers are unable to adjust their language based on audience feedback (e.g., a radio broadcast), they realize, as they speak, that there is another party listening and modify their speech accordingly. Because of this auto-monitoring feature, monologues normally contain two types of speech: main discourse, which involves the actual topic being discussed, and subsidiary discourse, talk that is concerned with the reception of subject matter and reflects on and monitors the discussion (Coulthard & Montgomery, 1981). In a similar vein, Enyedy and Hoadley (2006) pointed out that the critical factor that distinguishes a monologue from a dialogue is not the number of participants but the degree of participation of different parties to produce a text.

Diverse hypotheses have been expressed with regard to whether listeners can understand more when listening to monologues or dialogues (Fox Tree, 1999, pp. 39–41). On one hand, monologues are not usually tailored to a specific addressee. Therefore, they might contain extraneous information and lack overlaps in speech or interruptions, both of which may facilitate comprehension. On the other hand, comprehension of dialogues might be facilitated by the additional information conveyed by the speakers, including nonverbal cues such as gestures and

[1] However, it is difficult to replicate this feature with scripted texts used in a listening comprehension test even if they are added in the course of development, as the stimuli are typically written and then read aloud.

facial expressions (Wagner, 2010, p. 495). Two discourse studies support the second hypothesis. Listeners performed nonlinguistic tasks (the ordering and selection of various abstract shapes) more successfully with dialogic input, which was attributed to the larger number of discourse markers and the additional perspectives that multiple speakers introduced into the dialogue (Fox Tree, 1999; Fox Tree & Mayer, 2008).

Unlike the two studies just mentioned, research in the field of language testing has provided inconclusive results as to whether examinees find monologic or dialogic input more difficult. Shohamy and Inbar (1991) described a "continuum of orality" (Tannen, 1982), which proposes that an intimate conversation would fall at the oral end of the continuum, whereas a formal lecture would be found closer to the literate end, despite both forms of language being delivered orally. They found that listening comprehension input more closely associated with the oral end of the continuum yielded higher test scores than input associated with the literate end of the continuum. A news report was the most difficult for participants in their study, followed by a lecture, and finally by a consultative dialogue, which was the least difficult type. The authors attributed this finding to the larger density of propositions and complex grammatical and syntactic structures in the news broadcast and to the interaction with the audience in the lecture and the dialogue types. Questions posed by the audience may have facilitated learners' comprehension of the listening input. Shohamy and Inbar also hypothesized that the learners might have been more familiar with lectures and dialogues than with news broadcasts.

Contrary to Shohamy and Inbar, Read (2002) found that a monologue was easier to understand than a nonscripted dialogic version of the same content discussed by three speakers. Read attributed these results to the fact that the items in his study were originally intended to accompany the monologic stimulus and also that the dialogue was unscripted. The unscripted dialogue produced, according to Read (2002, p. 116), a more genuine sample of spontaneous and colloquial speech than Shohamy and Inbar's scripted consultative dialogue, but was also very demanding for nonnative listeners (half of the study participants reported that the three speakers in the unscripted dialogue spoke too fast). Moreover, he suggested that there may have been a practice effect, as learners who were administered the monologue task were given a similar task the previous week. Finally, the complexity of variables involved when comparing the two types of input might have made the results more difficult to interpret.

In a third study, Brindley and Slatyer (2002) were unable to draw any significant conclusions regarding the effect of dialogic versus monologic input on second-language listening comprehension; thus, they provided no support either to Shohamy and Inbar's (1991) or Read's (2002) findings. The researchers investigated five variables that, based on the literature, they hypothesized would affect difficulty: speech rate, text type (monologic or dialogic), number of hearings, input source, and item format. Brindley and Slatyer expected the dialogue version of one of their stimuli to be easier than the monologic version because of recycling of information and greater redundancy. The absence of any difference between the two text types was attributed to the interaction of the text type variable and the speech rate variable (Brindley & Slatyer, 2002, p. 388). One of the participants in the dialogue used a fast speech rate, and the majority of the responses were also spoken by this participant. As a result, any characteristics of the dialogue that could have made the input easier to comprehend were negated by the fast speech rate.

A number of studies (Freedle & Kostin, 1999; Kostin, 2004; Nissan, DeVincenzi, & Tang, 1996) have examined listening item difficulty in the Test of English as a Foreign Language (TOEFL). Although these studies did not address the issue of the relative difficulty of monologic

and dialogic input, they are relevant to the study presented in this article because of the similar item types (see the discussion in the upcoming Instruments section and see the appendices) and the insights offered into the effect of the audio input on item difficulty. For example, Nissan et al. (1996) found that five variables had a significant effect on the difficulty of dialogic items: word frequency in the stimulus, role of speaker, utterance pattern (i.e., combinations of statements and questions), use of negatives in the stimulus, and explicit or implicit information in the stimulus. The last three variables were also found to be significant in a second study investigating the difficulty of items associated with dialogic input (Kostin, 2004). Given that the utterance pattern variable (statements and questions) is typically associated with the dialogic format and that monologic input variables were accurate predictors of item difficulty in Freedle and Kostin (1999), the comparative difficulty of dialogic and monologic input merits further investigation.

It is widely accepted in the literature that listening is a highly complex, individual, and interactive process, during which listeners use a variety of skills and strategies (Brindley, 1998, p. 181) as well as background knowledge, past experience, feelings, and intentions to create an interpretation of the input (Buck, 2001, p. 29). The plethora of such variables might explain why results differed in the studies examining the relative difficulty of monologic and dialogic input and has implications for investigating the difficulty of monologic and dialogic input in language tests, as Brindley and Slatyer (2002, p. 390) concluded. Such an investigation is further complicated by the relationship between the stimulus and the items (Freedle & Kostin, 1999). There are variables that affect understanding of the stimulus, such as speed, pausing, pronunciation, explicitness of ideas, familiarity of topic, and frequency of vocabulary, that are very difficult to control and reproduce on a test (Buck, 2001, pp. 149–151). At the same time, the difficulty of items might depend on the listening subskills they test, their question type (e.g., selected response, open-ended questions), the amount of lexical overlap between the correct answer and the stimulus, and the syntactic complexity and frequency of the vocabulary used in the correct answer and distracters (Buck & Tatsuoka, 1998, pp. 120–126). It should also be recognized that although listeners perform a wide range of tasks in real life, examinees in most cases listen to a recorded stimulus and respond to test items without the opportunity to participate orally, thus functioning as overhearers of monologues and dialogues (Flowerdew & Miller, 2005, p. 89). Overhearers, according to Schober and Clark (1989), have a disadvantage over participants in a conversation, as they cannot collaborate with other participants to reach understanding. This lack of participation has important implications for the inferences that can be made about examinee ability in relation to the TLU domain, especially when it comes to dialogic input, and could question the construct validity of such listening tasks. In conclusion, due to the numerous variables that can affect examinee performance, comparing the relative difficulty of monologic versus dialogic input in the language testing context becomes a complex endeavor. We turn now to the research questions for this study.

RESEARCH QUESTIONS

As discussed earlier, research in the fields of language testing and discourse studies have not provided conclusive evidence as to the relative difficulty of the items associated with monologic and dialogic input. This issue is investigated in this study to help support the theoretical rationale for including more than one type of spoken text on tests of listening comprehension, thus assuring fuller coverage of the TLU domain. It is also a critical issue for examination providers who need to estimate test difficulty levels and make decisions related to test content and for users

whose test scores are affected by these decisions. Therefore, to investigate the relative difficulty of items associated with identical monologic and dialogic stimuli, the study addressed the following research questions:

RQ1: What is the relative difficulty of the same listening comprehension items when discourse in the stimulus is monologic versus dialogic?
RQ2: Do the identical options of these items demonstrate satisfactory measurement characteristics when discourse in the stimulus is monologic versus dialogic?

A description of the methodology used for the study follows.

METHOD

Participants

Data were collected in 2010 from 494 examinees during a routine administration of the Michigan English Test (MET), which is described next. The test was administered at nine exam centers in a Latin American country. All examinees indicated on their scannable answer sheets that Spanish was their first language. Information about the examinees' gender and age is presented in Table 1.

The examinees' English proficiency levels ranged from upper beginner (A2) to lower advanced (C1), based on the proficiency descriptions described in the Common European Framework of Reference (CEFR; Council of Europe, 2001). The examinees' listening proficiency scores on this test were distributed as follows: A2 level, 107 (21.66%); B1 level, 189 (38.26%); B2 level, 127 (25.71%); and C1 level, 71 (14.37%). More information on the relationship between test scores and the CEFR levels is provided in a relevant standard-setting study (Papageorgiou, 2010).

Instruments

The examinees were administered the MET, a paper-and-pencil examination of general English language proficiency provided by Cambridge Michigan Language Assessments (http://www.cambridgemichigan.org/met). It consists of two sections: Section I—Listening and Section II—Reading and Grammar. Only the Listening section is described here for the purpose of the study. The Listening section contains 60 multiple-choice items, divided into three parts: short dialogues between two interlocutors followed by one question, longer dialogues with two interlocutors preceding three to four questions, and monologues followed by four to five questions. Items have

TABLE 1
Examinee Gender and Age

Gender			*Age*			
Male	*Female*	*Missing*	*M*	*SD*	*Min*	*Max*
289 (58.5%)	195 (39.5%)	10 (2%)	22.5	8.33	14	63

TABLE 2
Content of the Three MET Listening Test Forms

Listening Stimuli	*No. Items per Stimulus*	*Items Used Across Forms*	*Form A*[a]	*Form B*[b]	*Form C*[c]
22 short dialogues	1	Common items	17	17	17
		Unique items	5	5	5
6 long dialogues	3–4	Common items	17	17	17
		Unique items	4	4	4
4 monologues	4–5	Common items	12	12	12
		Unique items	5	5	5
Total			60	60	60

[a]$N = 257$. [b]$N = 138$. [c]$N = 99$.

TABLE 3
Descriptive Statistics and Listening Test Form Reliability

Form	*k*	*N*	*α*	*M*	*SD*	*Min*	*Max*	*Skewness*	*Kurtosis*
A	60	257	.94	33.70	13.14	7	60	.15	−1.07
B	60	138	.95	35.34	14.56	9	58	−.10	−1.38
C	60	99	.94	30.63	13.90	7	58	.28	−1.07

four answer options, with one key and three distracters. All stimuli are played once, and questions and options are printed. Examinees are given the option to take notes when listening to the stimuli.

Data from the administration of three test forms were used in this study; the forms were administered to 257, 138, and 99 examinees, respectively, for a total of 494 examinees. As shown in Table 2, all forms contained the same number of items (60) and stimuli (32). Most items were common across forms (46) with a small number of unique items (14).

Table 3 presents a summary of the reliability and descriptive statistics for this administration of each listening form. Reliability indicated by Cronbach's alpha ranged from .94 to .95, suggesting high internal consistency of all three test forms. Performance varied from very low to very high scores as indicated by the mean, standard deviation, and minimum and maximum scores. The skewness and kurtosis values are within the –2 to +2 range; thus, a normal distribution is assumed for all forms.

For the study, analyses were performed only on examinee responses to items associated with the unique long dialogues and monologues on each form. The short dialogues were not included in the study because the stimuli were shorter than the stimuli of the other two types and were followed by one item, as opposed to three or more for the longer stimuli. Due to the brevity of these dialogic stimuli, developing comparable monologic stimuli would have been difficult.

Procedures

Three pairs of long dialogue and monologue stimuli were created for the purpose of the study, following the same item writing guidelines used to craft other MET items. To ensure that a range

of domains and topics were included in the study, stimuli pairs were developed in three of the four domains described in the CEFR (Council of Europe, 2001, p. 14): educational, occupational, and public. Three different writers crafted the stimuli and item sets, and the testlets underwent the same rigorous review as all other MET items.

As discussed earlier, variables such as topic familiarity, speech rate, and accent, as well as the interaction between the stimulus and the item options, may affect examinee performance. To the extent possible, the potential effects of such factors were minimized across the two input versions of each stimuli pair and in the creation of items. Each long dialogue stimulus had the same topic as one of the monologic stimuli. The educational domain topic was related to a lesson on how music is selected by band leaders, presented as a lecture by a professor in the monologic version and as a dialogue between the same professor and a student in the dialogic version. In this pair of stimuli, the professor includes the retelling of an experience he had as a director. Both stimuli include features of informal and spontaneous speech (e.g., pauses, connected speech, fillers, etc.) and are oriented toward the oral end of the continuum described by Tannen (1982).

The occupational topic was related to a sale in a retail store, presented as a comment by an employee during a store staff meeting in the monologic version and as a public dialogue between a customer and the same employee in the store. In this pair of stimuli, the dialogue is a spontaneous conversation between the employee and customer, who do not know each other; in the monologue, the employee retells the specifics of this encounter at the store during a staff meeting. Both include features of informal and spontaneous speech and are oriented toward the oral end of the continuum as well.

Finally, the public topic was related to a book about runners, presented as a radio book review in the monologic version and a conversation between two friends in the other. In this pair, the dialogue consists of a friend telling the other about the book. Although the dialogue is a spontaneous conversation, the monologue is a structured exposition of the content of the book with anecdotal comments by the speaker and falls onto the "literate" end of the oral continuum. The second and third topics contain some overlap because they both mention shoes. However, the focus of the second topic is on the confusion caused by the placement of a sale sign and the third topic is about the idea of running without running shoes.

Each of the three dialogic–monologic stimuli pairs shared identical items, with the same stem and four options (see the appendices). The stimuli contained identical key vocabulary and idiomatic expressions, and identical phrasing of details and main ideas, allowing the development of identical test items and answers. In addition, the solo speaker of the monologue was one of the two interlocutors in the dialogic versions of the stimuli. All speakers were professional, trained voice actors with consistent rates of speech. When possible, the answers for the four common items were located in the turns of the speaker who was heard in both the dialogic and monologic stimuli and were phrased in a similar or identical way. Although it was important for the dialogic and monologic versions to contain the same content, naturalness of each stimulus was also critical, so in some instances, phrasing of the tested content was not identical.

Items and options were also ordered similarly in the test forms. The only difference was that the monologic version contained five items as opposed to the dialogic version, which contained four items. This additional item (which appeared as the fourth item in the first and second monologues and as the third item in the third monologue) could not be avoided as it was dictated by the design of the test forms (see Table 2), and permission to use monologic stimuli in the test forms

TABLE 4
Distribution of the Unique Stimuli

Form A	*Form B*	*Form C*
DIA1	DIA2	DIA3
MON3	MON1	MON2

Note. DIA = dialogic; MON = monologic.

was allowed only for monologues with five items. Therefore, the additional item is not included in the presentation of the results of the study.

To avoid a potential memory effect, the design shown in Table 4 was adopted. The first dialogic stimulus (DIA1) was administered to examinees who took Form A, but the monologic version (MON1) was administered to examinees who took Form B. The two versions of the second stimuli pair, DIA2 and MON2, were administered to examinees who took Forms B and C, and the two versions of the third pair, DIA3 and MON3, were administered to examinees who took Form C and A, respectively.

Although memory effect was controlled using the aforementioned design, the estimation of the relative difficulty of the monologic and dialogic versions of the items was potentially problematic, given that not all of the 494 examinees would respond to every item. For this reason, analyses were performed using the Rasch model (Rasch, 1980) with the computer program WINSTEPS (Linacre, 2009), which is robust to missing data (Bond & Fox, 2007, p. 312) and more appropriate than other item response theory models due to the relatively small N size of Form C (McNamara, 1996, p. 295). Even though items were administered to three different populations, a comparison of the difficulty of all items is possible using WINSTEPS because the data are linked through the 46 common listening items on the MET. The Rasch model produces linear measures of item difficulty and person ability on a common interval scale of "log odds" units (McNamara, 1996, p. 165) centered on 0, the "logit" scale. Positive values indicate more proficient examinees or more difficult items, whereas negative values indicate less proficient examinees or easier items. Through analysis of the differences between observed responses and responses expected by the model, fit statistics were calculated, indicating the degree to which items fit the underlying construct. Estimates of item difficulty and fit are further discussed in the following section.

The responses to all 88 items—46 common items and 42 unique items (14 for each form)—were imported from the scannable answer sheets to a Microsoft Excel spreadsheet. They were then exported to WINSTEPS along with the correct response, or key, and analyzed using the dichotomous Rasch model, where 1 indicates a correct answer and 0 a wrong answer. As is explained in the next section, when the Rasch analyses indicated differences in item difficulty and the performance of the item options, qualitative content analyses were conducted to explore the possible causes of these differences.

RESULTS

To answer the two research questions, the presentation of the results has been divided into three sections, each first providing an analysis of the relative difficulty of the listening items for each

TABLE 5
Item Statistics for Pair 1 DIA/MON

	N		*F.V.*		*Logit*		*SE*		*Infit MNSQ*		*Infit ZSTD*	
Item	*DIA*	*MON*	*DIA*	*MON*	*DIA*	*MON*	*DIA*	*MON*	*DIA*	*MON*	*DIA*	*MON*
01	257	138	0.72	0.72	−0.98	−0.98	0.15	0.21	1.01	1.02	0.1	0.3
02	256	137	0.49	0.42	0.30	0.75	0.14	0.21	0.86	0.92	−2.4	−0.8
03	254	138	0.64	0.67	−0.48	−0.66	0.15	0.21	1.01	1.17	0.1	1.8
04	256	138	0.50	0.49	0.26	0.37	0.14	0.20	1.27	1.29	4.2	2.8

Note. DIA = dialogic; MON = monologic; F.V. = facility value; MNSQ = mean square statistic; ZSTD = transformed, standardized fit statistic.

pair of stimuli and then an analysis of the measurement characteristics for each set of items when discourse in the stimulus is dialogic versus monologic.

Comparison of Pair 1 Items

Table 5 lists the items that were identical in each version (first column) and then presents the item statistics twice, once for the dialogic version and once for the monologic version. The number of examinees who responded to each item is shown first, followed by its facility value, that is, the percentage of examinees who responded correctly (Alderson, Clapham, & Wall, 1995, p. 80). The difficulty estimate is presented in logit values as is its associated standard error. Of the four fit statistics calculated by WINSTEPS, two are reported here because they rely more on responses of examinees whose ability is well matched with item difficulty on the logit scale (Bond & Fox, 2007, p. 57). The infit mean square statistic (infit MNSQ) has an expected value of 1. Although interpretation of acceptable values depends on the data (Linacre & Wright, 1994), typically values above 1.3 show significant underfit, which indicates a lack of predictability, signaling either that the items are problematic or that they do not measure the same trait (Bond & Fox, 2007, p. 240). Values below .75 show significant overfit, indicating a lack of variation, which suggests that the overall response pattern is too predictable and there might be content overlap with other items (McNamara, 1996, p. 175). A transformed, standardized fit statistic (infit ZSTD) is also reported. It tests the null hypothesis that the data fit the Rasch model after allowing for randomness predicted by the model (Linacre, 2009, p. 161). The standardized fit statistic ensures that Type I errors, occurring when the null hypothesis is rejected when it is actually true (Cohen, 2008, p. 131), are not affected by varying sample sizes (Smith, 2004, p. 83). Values between –2 and +2 are considered acceptable.

For the purposes of this study we consider a difference of half a logit between two item versions (dialogic and monologic) to be indicative of a substantive difference in difficulty. This is because when person ability and item difficulty are well matched, a decrease in difficulty by half a logit will increase the probability of a correct response by about 12.5% (McNamara, 1996, pp. 165–166). The mean logit value of the two versions only differs by one tenth of a logit, –0.23 for the dialogic items and –0.13 for the monologic items, suggesting that the two versions of Pair 1 are of similar difficulty. As shown in Table 5, none of the four items demonstrates a difference of more than 0.50 logits between the two versions. The two versions of Items 01, 02,

03, and 04 differ by 0.00, 0.45, 0.22, and 0.11 logits, respectively. Although all items demonstrated acceptable infit MNSQ (between .86 and 1.29), Item 04 underfits in both versions when the infit ZSTD is taken into account. This underfit could be explained by the slightly different format of the stem (see Appendix A), where part of the stimulus is replayed ("What does the professor mean when he says *a variety of music will appeal to all musicians*?" [italics added to indicate replay from stimulus which is only heard and not printed in the booklet]). Moreover, the dialogic version of Item 02 demonstrated some overfit, as the infit ZSTD value was –2.4. However, content analysis did not reveal possible reasons for this overfit. It should also be noted that the infit MNSQ value for this item was within the acceptable range (.86).

To explore how the four options of each item version performed, an item option frequency analysis (Linacre, 2009, p. 254) was run through WINSTEPS (Table 6). Horizontally, Table 6 is divided into four parts, one for each item and its options (first and second columns), with the key indicated by an asterisk. Similarly to Table 5, item statistics are presented twice, first for the dialogic version and then for the monologic version. The frequency column shows the number of examinees that chose each option, which is also shown in percentages. The measure statistic in the third column of each version is a sample-dependent statistic that shows the average measure of examinees who responded to the item. The examinee measure for the correct option should be higher than the measure for any single distracter because more able examinees should choose the key, whereas less able examinees should choose the distracters. The measure statistics are accompanied by a standard error estimate. Finally, the measurement correlation is a correlation between the responses (1 for the key and 0 for the distracters) and the person measures. The key should demonstrate positive values, whereas the distracters should demonstrate negative values, or very low positive values.

TABLE 6
Item Option Statistics for Pair 1 DIA/MON

		Dialogic Version					*Monologic Version*				
Item	*Option*	*Frequency*	*%*	*Meas.*	*SE*	*Correl.*	*Frequency*	*%*	*Meas.*	*SE*	*Correl.*
01	1	32	12	−0.62	0.11	−0.27	9	7	−0.12	0.34	−0.09
	2	29	11	−0.35	0.21	−0.18	24	17	−0.90	0.20	−0.41
	3*	186	72	0.61	0.09	0.39	99	72	0.69	0.13	0.44
	4	10	4	−0.49	0.21	−0.13	6	4	−0.33	0.43	−0.10
02	1*	125	49	1.05	0.12	0.57	57	42	1.24	0.15	0.59
	2	17	7	−0.49	0.13	−0.17	14	10	−0.95	0.20	−0.31
	3	57	22	−0.38	0.11	−0.29	38	28	−0.05	0.18	−0.16
	4	57	22	−0.39	0.10	−0.29	28	20	−0.50	0.17	−0.30
03	1	26	10	−0.22	0.19	−0.14	21	15	−0.16	0.29	−0.15
	2	28	11	−0.63	0.11	−0.26	12	9	−0.10	0.38	−0.10
	3	38	15	−0.40	0.14	−0.23	13	9	−0.97	0.18	−0.31
	4*	162	64	0.73	0.10	0.43	92	67	0.66	0.13	0.36
04	1	21	8	−0.26	0.17	−0.13	5	4	0.31	0.60	0.00
	2	23	9	−0.23	0.18	−0.13	11	8	−0.40	0.37	−0.16
	3*	127	50	0.68	0.13	0.29	67	49	0.81	0.17	0.36
	4	85	33	0.04	0.12	−0.15	55	40	−0.15	0.15	−0.28

Note. DIA = dialogic; MON = monologic; Meas. = measure statistic; Correl. = measurement correlation; *= key.

Table 6 shows that the person measure value of the key of all items is higher than the measure value of any distracter. Correlations for the key are also positive and higher than the correlations for the distracters. Thus, these statistics indicate that the key and the distracters possess satisfactory measurement characteristics in both versions.

Comparison of Pair 2 Items

Table 7 presents the same statistical information for the second pair as Table 5 for Pair 1. Compared to the first pair, the difference in difficulty (0.78 logits) is more prominent, as the mean logit value for the dialogic version is –0.34 and for the monologic version is 0.44. Items 06, 07, and 08 demonstrate difficulty differences of more than half a logit (0.58, 1.39, and 0.84, respectively), with the monologic version being the most difficult. All item fit statistics are satisfactory, and only the dialogic version of Item 08 overfits according to the standardized infit statistic (the infit mean square value, however, is acceptable). A possible explanation for the overfit is interdependency with the previous item as they both refer to aspects of the price of shoes (see Appendix B). Overfit in the monologic version was probably not observed because of the additional item in that set, which, as explained earlier, is not part of the analysis. This item was placed between Items 07 and 08 and might have mitigated a potential interdependency.

The item option analysis for the second pair is presented in Table 8. The person measure value for the key of each version of the three items is in all cases higher than the measure value for the distracters. Moreover, correlations for the key of all items are positive and higher than the measurement correlations for the distracters. These statistics suggest satisfactory measurement characteristics for the item options in both versions.

Statistical analysis was followed by content analysis of the part of the stimuli relevant to the three items that demonstrated difficulty differences of more than half a logit. The answer to Item 06 (“Which shoes are on sale?”) in the dialogic version is partly based on a turn by the interlocutor who is not the presenter in the monologic version (the customer) and a turn by the employee who is also the speaker in the monologic version. The customer mentions to the employee that he does not know which shoes are on sale on a display table. The employee explains that only the shoes placed under the sale sign hanging over the table are on sale, “not the ones on the left side of the display table” (verbatim in both versions as shown in Appendix B). In both versions, the sale

TABLE 7
Item Statistics for Pair 2 DIA/MON

	N		*F.V.*		*Logit*		*SE*		*Infit MNSQ*		*Infit ZSTD*	
Item	*DIA*	*MON*	*DIA*	*MON*	*DIA*	*MON*	*DIA*	*MON*	*DIA*	*MON*	*DIA*	*MON*
05	138	98	0.72	0.62	−0.98	−0.67	0.21	0.24	0.91	0.98	−0.9	−0.2
06	138	99	0.45	0.29	0.57	1.15	0.20	0.26	0.94	1.18	−0.6	1.3
07	138	99	0.73	0.43	−1.07	0.32	0.22	0.23	0.98	1.16	−0.1	1.5
08	138	99	0.53	0.32	0.12	0.96	0.20	0.25	0.78	1.07	−2.5	0.6

Note. DIA = dialogic; MON = monologic; F.V. = facility value; MNSQ = mean square statistic; ZSTD = transformed, standardized fit statistic.

TABLE 8
Item Option Statistics for Pair 2 DIA/MON

		Dialogic Version					*Monologic Version*				
Item	*Option*	*Frequency*	*%*	*Meas.*	*SE*	*Correl.*	*Frequency*	*%*	*Meas.*	*SE*	*Correl.*
05	1	17	12	−0.93	0.22	−0.35	15	15	−0.40	0.22	−0.15
	2	18	13	−0.57	0.20	−0.25	19	19	−0.99	0.14	−0.40
	3*	99	72	0.76	0.13	0.52	61	62	0.53	0.17	0.48
	4	4	3	−1.34	0.13	−0.21	3	3	−0.93	0.38	−0.14
06	1	22	16	−0.12	0.26	−0.14	19	19	−0.53	0.19	−0.21
	2	44	32	−0.42	0.16	−0.37	40	40	−0.24	0.18	−0.17
	3	10	7	−0.76	0.24	−0.22	11	11	−0.04	0.47	−0.02
	4*	62	45	1.17	0.15	0.57	29	29	0.80	0.24	0.39
07	1*	101	73	0.68	0.13	0.44	43	43	0.60	0.21	0.39
	2	14	10	−0.66	0.33	−0.24	22	22	−0.29	0.26	−0.13
	3	12	9	−0.78	0.36	−0.25	20	20	−0.63	0.18	−0.26
	4	11	8	−0.58	0.22	−0.20	14	14	−0.27	0.23	−0.10
08	1	29	21	−0.54	0.18	−0.33	42	42	−0.19	0.17	−0.15
	2*	73	53	1.16	0.14	0.66	32	32	0.84	0.25	0.44
	3	16	12	−0.47	0.18	−0.21	12	12	−0.77	0.23	−0.23
	4	20	14	−0.89	0.14	−0.37	13	13	−0.50	0.22	−0.16

Note. DIA = dialogic; MON = monologic; Meas. = measure statistic; Correl. = measurement correlation; *= key.

sign (part of the key) is mentioned twice, but the table (part of one distracter) is mentioned twice in the monologue as opposed to three times in the dialogue (two by the customer and one by the employee). This additional reference to the table in the dialogue, combined with the employee's explanation to the customer might have helped examinees avoid the distracter.

The key to Item 07 ("What did the woman offer to ask the manager?") was presented in direct speech in the dialogue ("I can ask the manager if he'd give you a special discount") as opposed to reported speech in the monologue ("and if he liked them, we could ask the manager about giving him a special discount"). Similarly, the key to Item 08 ("What does the woman mean when she says *they're a lot less pricey*" [italics added to indicate replay from stimulus]) is presented in direct speech in the dialogue ("They're so expensive") as opposed to reported speech in the monologue ("He thought they were kinda expensive"). Moreover, the customer provides additional input in the dialogic version that is not present in the monologic version—for example, that he needs to find shoes suitable for a business trip and that he liked the expensive pair of shoes. Therefore, it could be argued that the additional input in the dialogic format, in combination with the direct speech, might have resulted in lowering the difficulty of these items in the dialogic version.

Comparison of Pair 3 Items

Table 9 presents the same statistical information for the third pair as Table 5 for Pair 1 and Table 7 for Pair 2. Unlike the other two pairs, the dialogic version of the third pair were more difficult with a mean logit value of 0.52, compared to 0.43 in the monologic version. Difficulty differences of more than half a logit are observed for Items 09, 11, and 12 (0.68, 0.63, and 0.60,

TABLE 9
Item Statistics for Pair 3 DIA/MON

	N		*F.V.*		*Logit*		*SE*		*Infit MNSQ*		*Infit ZSTD*	
Item	*DIA*	*MON*	*DIA*	*MON*	*DIA*	*MON*	*DIA*	*MON*	*DIA*	*MON*	*DIA*	*MON*
09	99	256	0.34	0.52	0.84	0.16	0.25	0.14	1.02	0.86	0.2	−2.4
10	99	255	0.37	0.36	0.66	0.97	0.24	0.15	1.10	1.07	0.9	1.0
11	98	255	0.49	0.42	0.02	0.65	0.23	0.15	1.12	1.31	1.2	4.3
12	99	256	0.39	0.56	0.54	−0.06	0.24	0.14	0.88	0.91	−1.1	−1.7

Note. DIA = dialogic; MON = monologic; F.V. = facility value; MNSQ = mean square statistic; ZSTD = transformed, standardized fit statistic.

TABLE 10
Item Option Statistics for Pair 3 DIA3/MON

		Dialogic Version					*Monologic Version*				
Item	*Option*	*Frequency*	*%*	*Meas.*	*SE*	*Correl.*	*Frequency*	*%*	*Meas.*	*SE*	*Correl.*
09	1*	34	34	0.90	0.22	0.49	132	52	1.01	0.11	0.57
	2	18	18	−0.36	0.25	−0.14	53	21	−0.43	0.11	−0.30
	3	28	28	−0.74	0.16	−0.38	30	12	−0.79	0.12	−0.32
	4	19	19	−0.01	0.26	−0.02	41	16	−0.19	0.12	−0.17
10	1	11	11	−0.82	0.24	−0.24	65	25	−0.12	0.12	−0.19
	2*	37	37	0.77	0.21	0.45	93	36	1.05	0.14	0.44
	3	22	22	−0.20	0.29	−0.10	33	13	−0.17	0.18	−0.14
	4	29	29	−0.41	0.16	−0.22	64	25	−0.10	0.13	−0.18
11	1	7	7	−1.10	0.22	−0.24	21	8	−0.74	0.17	−0.25
	2	25	26	−0.40	0.19	−0.19	69	27	−0.22	0.09	−0.25
	3*	48	49	0.59	0.20	0.43	108	42	0.73	0.14	0.28
	4	18	18	−0.43	0.24	−0.17	57	22	0.56	0.15	0.10
12	1	13	13	−0.48	0.27	−0.16	21	8	−0.32	0.21	−0.15
	2*	39	39	0.96	0.21	0.59	143	56	0.91	0.11	0.53
	3	31	31	−0.57	0.16	−0.32	60	23	−0.53	0.10	−0.37
	4	16	16	−0.64	0.16	−0.23	32	13	−0.37	0.12	−0.20

Note. DIA = dialogic; MON = monologic; Meas. = measure statistic; Correl. = measurement correlation; *= key.

respectively). The dialogic version of Items 09 and 12 were more difficult, whereas the opposite happened with Item 11 (as well as Item 10, which is not discussed because the difference was smaller, 0.31 logits). All item fit statistics are satisfactory, with the exception of the standardized fit for Item 09 and both fit statistics for Item 11 in the monologic version (fit in the dialogic version was acceptable). Table 10 provides more information about the performance of the keys and distracters for these items.

As can be seen in Table 10, the person measure value for the key is in all cases higher than the measure value for the distracters. Moreover, correlations for the key of all items were positive and higher than the measurement correlations for the distracters. These statistics suggest that the key and the distracters performed in the intended manner in both versions. It should be noted, however, that the fourth option of the monologic version of Item 11 demonstrated a

person measurement value that was relatively close to the value of the key and that the correlation, although lower than the correlation of the key, was positive (the only case among all item distracters discussed in this study). This is further addressed next as part of the content analysis that was performed regarding Items 09, 11, and 12 (see also Appendix C).

Content analysis related to Item 09 ("What is the main point of Jerry Hampton's book"), revealed that the monologic stimulus, a radio report about a book, was more structured and detailed than the dialogic version, a conversation between the radio reporter and her friend about the same book. Moreover, lexical items that appear in the key ("Running shoes are not for runners") were mentioned more often in the monologue. For example, the word "shoes" is mentioned five times in the monologue but only once in the dialogue, and the statement "I've suffered a lot of running injuries myself over the years—all while wearing expensive shoes" is made only in the monologue.

Contrary to Item 09, the dialogic version of Item 11 ("What does the woman think will happen in the future?") was easier than the monologic version. The key "researchers will test Hampton's theory" is a paraphrase of the stimulus utterance "I think that some real scientists will do some experiments soon," which appears verbatim in both the dialogic and monologic versions. The difference in difficulty was probably observed because Option 4 of this item ("All marathoners will run barefoot") functioned differently in the two versions. In the monologic version, it attracted examinees of high ability, as shown by the positive person measurement value (0.56) and correlation (0.10) in Table 10. These figures were negative in the dialogic version (–0.43 and –0.17, respectively). This differential functioning of Option 4 could have been observed because high-ability learners in the monologic version inferred that it was a plausible key: The author of the book recommends that all people run barefoot, and if research confirms the benefits of running barefoot, this is a possibility in the future.

Item 12 for the monologic version of the stimulus ("What does the speaker mean when she says *in the book he's clear about his lack of medical credentials*?" [italics added to indicate replay from stimulus]) might have been easier than the dialogic version for various reasons. As shown in Appendix C, the radio presenter stresses in the monologic version that Mr. Hampton is not a doctor ("But I must say, Mr. Hampton is not a doctor. In the book he's clear about his lack of medical credentials"), which contains strong lexical overlap with the key ("Hampton is not a qualified doctor"). However, in the dialogic version the presenter's friend uses a tag question to ask if Mr. Hampton is a doctor and the presenter responds with a pronoun referent ("No, in the book he's clear about his lack of medical credentials"). The emphatic point in the monologue and the use of negation in the tag question and the pronoun referent in the response might have also helped examinees choose the key. This suggests that the interlocutor's contribution in a dialogue might not always offer additional information that is helpful to a listener when the conversation is overheard and there is no chance to interact. Likewise, monologues that are well structured and detailed may be easier to process for listeners in some cases.

DISCUSSION AND CONCLUSIONS

Regarding the first research question (What is the relative difficulty of the same listening comprehension items when discourse in the stimulus varies from monologic to dialogic?), the statistical analyses revealed a consistent pattern only with Pair 2, as three out of four items were easier by

more than 0.50 logits when administered with the dialogic version of the stimulus. The pattern of relative difficulty for the other two pairs of items was mixed. For Pair 1, none of the items for either version of the stimuli differed by more than half a logit, and the mean logit value of the items indicated that the dialogic and monologic versions were of similar difficulty. Pair 3 demonstrated a mixed trend. Of the items with differences above 0.50 logits, two were easier when administered with the monologic version of the stimulus and one item was easier with the dialogic version.

Statistical analyses conducted for the purposes of answering the second research question (Do the identical options of these items demonstrate satisfactory measurement characteristics when discourse in the stimulus varies from monologic to dialogic?) indicate that the items and their options performed satisfactorily with both versions of the stimuli. The only item that all statistical indices flagged as potentially problematic was Item 11 in the monologic version of the third pair. Distracter analyses further indicated that some high-ability examinees selected a specific distracter. However, the measure value for these examinees remained lower than the measure value of the examinees that chose the key, which is encouraging because more able examinees should choose the key, whereas less able examinees should choose the distracters.

To interpret the statistical findings, items with differences of 0.50 logits and above were analyzed qualitatively. Content analyses of the second pair indicated that three of the items might have been easier with the dialogic version because of the additional information provided by both speakers, combined with the use of direct speech as opposed to reported speech in the monologic version. Regarding the third pair, two items were easier with the monologic version of the stimulus than the dialogic version (09 and 12). Content analysis showed that the monologue was more structured than the dialogue, contained additional explicit statements, and had more lexical overlap with the language used in the correct responses. However, Item 11 of the same pair was easier when used with the dialogic version of the stimulus. Although the key was a paraphrase of a sentence that was identical in both versions of the stimulus, one of the distracters functioned differently in the two versions and was attractive for examinees of high ability who took the monologic version. Content analysis pointed to a possible double key by inference.

Six items were discussed in this article because their difficulty changed by more than 0.50 logits in the two versions. Because four of these items were easier in the dialogic version, it could be argued that the findings of this study mostly confirm the hypothesis that dialogic input is easier for examinees to comprehend than monologic input. However, the fact that two items were easier in the monologic version also provides some support to the hypothesis that it might be easier to comprehend some monologues when they contain more structured and explicit information. Because these hypotheses were formed by researchers outside the language testing context, it is important to remember that when analyzing data from the administration of a language test such as the one in this study, additional factors might affect results. For example, the fact that the monologues and dialogues in this study were scripted and intentionally parallel might have played a role in the lack of a consistent pattern of differences in difficulty because examinees were listening to very similar input. This holds true for the first pair in particular, for which no item had a difficulty of more than 0.50 logits between the two versions. Identical or almost identical wording in the stimuli of the first pair was used more extensively than in the other two pairs.

Although domain coverage and construct representation are essential for the test development process, achieving the desired difficulty is important in many contexts, for example, when multiple cut scores are reported to score users in relation to multiple proficiency levels (this is the case

with the test used in this study). The findings of this study suggest that item writers may be able to manipulate aspects of dialogic or monologic stimuli to achieve a desired level of difficulty. Although this was not purposely done in this study, writers could, for example, increase item facility value by adding additional information to support the content in a stimulus, by conveying information through the use of direct speech, or by adding explicit or direct information to monologues. It should be pointed out, however, that even if these aspects of dialogic and monologic discourse are controlled effectively, the numerous factors that affect listening comprehension will probably do so irrespective of the format of the stimulus. For example, in this study we found that lexical overlap between the options and the stimulus was a factor affecting item difficulty regardless of whether the stimulus was a monologue or a dialogue.

Naturally, the findings of this article should be interpreted with caution due to some limitations. The data were collected from the administration of a single test, containing scripted stimuli, as opposed, for example, to the unscripted conversations in Read's (2002) study, which were found to be challenging for the listeners. The use of scripted dialogues and monologues is generally not considered to be as representative of the TLU domain as using unscripted authentic recordings. Therefore, although the use of both dialogic and monologic input might increase the face validity of the test (there are two types of spoken texts), the construct validity may be impacted.

Because the items discussed in this article had to adhere to the overall format of the test, monologic stimuli were slightly longer than dialogic stimuli and contained an additional question. Also, due to the additional length and item, memory could have affected examinee performance, given that human memory capacity places limitations on comprehension (Wu, 1998, p. 23). Furthermore, examinees were of a specific language background and responded to selected-response items only. As with the format of the test (which resulted in an additional item for the monologic version), we were also unable to assign test forms to these examinees. Therefore, we recognize that the higher error associated with the logit value of the items in Form C is probably due to the smaller number of test takers of that form and might make the comparison of item difficulty problematic. Future studies adopting similar methodology will need to ensure larger test taker numbers across all forms of the test.

Despite our attempts to make the two stimuli versions parallel and control for a number of variables, it should be acknowledged that there were still some differences between the monologic and dialogic versions of each pair for example in terms of information density, redundancy, and discourse markers, and also across the three pairs. As discussed earlier, this is the case in particular with Pair 3, the monologic version of which is more structured and detailed than the dialogic version. Moreover, the lexical items in the key appeared more often in the monologue than in the dialogue. Therefore, the higher facility value of the items of the monologic version should be attributed not to the nature of the input only but to many confounding factors.

It should also be noted that these examinees employed their listening comprehension ability as passive overhearers of discourse, not as active participants in a face-to-face interaction. Although language users frequently find themselves in "noncollaborative" listening situations (Buck, 2001, p. 98), the fact that examinees were only overhearers suggests underrepresentation of the TLU domain (Wagner, 2010, p. 510) and possibly underestimation of how examinees understand aural input (Schober & Clark, 1989). It is also possible that overhearing some stimuli is relatively far removed from the realm of possible situations for overhearing everyday language. This could be the reason why the monologic version of Pair 2 was more difficult than the dialogic version. Examinees who missed the narrator's lead-in sentence, "Listen to an employee speaking at a

staff meeting," might have found it difficult to grasp the scenario and identify the audience. The dialogic version of this pair (interchange between a shoe store employee and customer) is more likely to be within the realm of experience of most test takers.

Due to the limitations of this study, stimulus and item format, as well as examinee language background and overhearer role, warrant further investigation. Future qualitative research, such as examinee verbal protocols (Buck, 1991; Wu, 1998) could offer additional insights into how aspects of dialogic and monologic format are processed by examinees and how they affect item difficulty.

The study of the relative difficulty of dialogic and monologic input is a complex issue due to the numerous, well-documented variables that affect listening comprehension (Brindley, 1998; Brindley & Slatyer, 2002; Buck & Tatsuoka, 1998; Freedle & Kostin, 1999; Kostin, 2004; Nissan et al., 1996), even when similar dialogic or monologic stimuli are used, such as university lectures (Chiang & Dunkel, 1992). Although it could be argued that the monologic–dialogic distinction should be investigated in conjunction with such variables, it is important in many contexts, as Read (2002, p. 117) pointed out, to adjust the level of difficulty of the input text to make it suitable to evaluate examinees' proficiency level. Moreover, to address the need for valid inferences in relation to a specific TLU domain, the practice of using both dialogic and monologic stimuli is likely to remain unchanged in the majority of testing contexts. For these reasons, and despite the complexity of controlling for other variables affecting item difficulty, further exploration of the relative difficulty of these two input types remains important as a primary research focus.

ACKNOWLEDGMENTS

This paper was submitted for publication when the first author was employed by Cambridge Michigan Language Assessments, the publisher of the Michigan English Test. Any opinions expressed in this publication are those of the first author and not necessarily of Educational Testing Service, his current affiliation. The research reported in this article started when all three authors were employed by the English Language Institute of the University of Michigan, the original publisher of the Michigan English Test. We thank the three anonymous reviewers for their constructive feedback and advice on an earlier version of this article.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, *18*, 171–191.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*, 369–394.

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, *8*, 67–91.

Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*, 119–157.

Chiang, C. S., & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly*, *26*, 345–374

Cohen, B. H. (2008). *Explaining psychological statistics* (3rd ed.). Hoboken, NJ: Wiley.

Coulthard, M., & Montgomery, M. (1981). Developing a description of spoken discourse. In M. M. Coulthard (Ed.), *Studies in discourse analysis* (pp. 1–50). Boston, MA: Routledge & Kegan Paul.

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

Enyedy, N., & Hoadley, C. M. (2006). From dialogue to monologue and back: Middle spaces in computer-mediated learning. *Computer-Supported Collaborative Learning*, *2006*, 413–439.

Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice*. Cambridge, UK: Cambridge University Press.

Fox Tree, J. E. (1999). Listening in on monologues and dialogues. *Discourse Processes*, *27*, 35–53.

Fox Tree, J. E., & Mayer, S. A. (2008). Overhearing single and multiple perspectives. *Discourse Processes*, *45*, 160–179.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? the case for the construct validity of TOEFL's minitalks. *Language Testing*, *16*, 2–32.

Kostin, I. (2004). *Exploring Item characteristics that are related to the difficulty of TOEFL dialogue items* (No. RR–79). Princeton, NJ: Educational Testing Service.

Linacre, J. M. (2009). WINSTEPS Rasch measurement computer program version 3.68.2. Chicago, IL: Winsteps.com.

Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.

McNamara, T. (1996). *Measuring second language performance*. Harlow, UK: Longman.

Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension*. Princeton, NJ: Educational Testing Service.

Papageorgiou, S. (2010).*Setting cut scores on the Common European Framework of Reference for the Michigan English Test* (Tech. Rep.). Ann Arbor: University of Michigan. Retrieved from http://www.cambridgemichigan.org/sites/default/files/resources/MET_StandardSetting.pdf

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Read, J. (2002). The use of interactive input in EAP listening assessment. *Journal of English for Academic Purposes*, *1*, 105–119.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*, 211–232.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing*, *8*, 23–40.

Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73–92). Maple Grove, MN: JAM Press.

Tannen, D. (1982). The oral/literate continuum in discourse. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 1–16). Norwood, NJ: Ablex.

Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*, 493–513.

Wu, Y. (1998). What do tests of listening comprehension test? A retrospective study of EFL test-takers performing a multiple-choice task. *Language Testing*, *15*, 21–44.

APPENDIX A

Pair 1, Dialogic Version Script

Narrator: Listen to a conversation between a music professor and a student.

Woman: Hi, Professor. Do you have a few minutes? I missed our last class, and I was wondering if you could review what you discussed about band directors choosing songs for concerts.

Man: Sure. Well, when selecting music, of course a director should keep the tastes of the audience in mind, but it's also important to consider the musicians themselves; that

is, what type of music will keep them interested. So, there are two main concerns: difficulty and diversity. First, in terms of difficulty, it's good to select a range: some songs should be easy to play and others more challenging. Just be sure the difficult songs aren't too difficult.

Woman: Oh, because the harder they are, the more time musicians have to spend practicing?

Man: Right. Once, I was directing a community band, and I picked a song with a very complex trumpet part. The trouble was, during practice, the other band members had to wait around while the trumpets played their part of the music over and over, until it sounded right. Before I knew it, there were only five minutes of rehearsal left!

Woman: So the director has to remind musicians to practice more challenging parts at home, huh?

Man: Exactly. And then, when considering the diversity of music for a concert, to keep musicians engaged, choose pieces that vary in both style and pace. Remember, not all players like the same kind of music, and you want to include something for everyone. A variety of music will appeal to all musicians.

Pair 1, Monologic Version Script

Narrator: Listen to a professor speaking to a music class.

Man: I know that some of you hope to find jobs as directors of concert bands, and one of the things you'll need to do as a band leader is to select the music – that is, the songs – to include in each concert your band plays. Of course, a director should keep the tastes of the audience in mind, but it's also important to consider the musicians themselves; that is, what type of music will keep band members interested. So, there are two main concerns: difficulty and diversity. First, you should select pieces that include a range of difficulty levels. Some songs should be fairly easy to play and others should be more challenging. However, it's important to make sure that even the more challenging pieces are not beyond the musicians' level of ability. Remember: the harder the music is, the more the musicians will have to practice. Once, I was directing a community band, and there was a really complex trumpet part. The trouble was, during practice, the other band members had to wait while the trumpets played their part of the music over and over, until it sounded right. Before I knew it, there were only five minutes of rehearsal left! So you have to remind musicians to practice the more difficult parts at home. Another thing to remember is variety: include pieces that vary in both style and pace. A variety of music will appeal to all musicians. Some prefer classical music; some prefer more modern music. You want to include something that appeals to band members' varied tastes. Let's practice this, OK? I'd like you to work with a small group and make a list of songs that a high school band might play for a concert, keeping in mind the information we've just covered.

Items Used With Both Versions

01 What is the professor mainly discussing?

a. how to select instruments for a concert
b. how to compose concert music

c. how to choose music for a concert
d. how to pick a concert to attend

02 Why does the professor mention trumpet players?

a. to describe a problem he had
b. to explain why slow songs are better
c. to compare them to other musicians
d. to show the type of music audiences prefer

03 What advice is given about challenging music?

a. It should not be included in a concert.
b. Players only need five minutes of practice time.
c. Only advanced students should perform it.
d. Musicians should practice it on their own.

04 What does the professor mean when he says:
[audio only: A variety of music will appeal to all musicians]

a. Certain music styles should be avoided.
b. Playing several styles will confuse people.
c. Playing different styles will keep people interested.
d. Musicians usually request a specific music style.

APPENDIX B

Pair 2, Dialogic Version Script

Narrator: Listen to a conversation in a store.
Woman: May I help you?
Man: Yes, thanks. I was wondering if this pair of shoes is on sale. There's a sale sign hanging over part of the display table, so I'm not sure which shoes on the table are on sale.
Woman: I've had at least ten customers ask me about this. We'll have to fix the display. Only the shoes that are under the sale sign are being discounted this week. Not the ones on the left side of the table.
Man: Oh, that's too bad, I really like this pair, but they're so expensive . . .
Woman: Yeah, it's a good brand. Did you see the black shoes that are on sale? They're a lot less pricey.
Man: Uhh, I did. Those are too casual though. I need something suitable for a business trip.
Woman: Well, if you want . . . you can try the dressier pair on. And if you really like the way they fit, I can ask the manager if he'd give you a special discount. The store's good about trying to make customers happy. Want to give them a try?
Man: Sure, thanks.

Pair 2, Monologic Version Script

Narrator: Listen to an employee speaking at a staff meeting.

Woman: I know that we have a lot of things to talk about in our meeting today, but I wanted to bring up a problem we're having in the shoe department. Like, just yesterday, I had ten customers ask me which shoes are on sale. I think it's because of the sale sign, it's hanging down over part of the display table, you know, only the shoes that are under the sign are being discounted this week. Not the ones on the left side of the table. One guy was asking about a pair of formal black shoes he found on the display. He thought they were kinda expensive, so I showed him in the other black shoes, the ones that are on sale. They're a lot less pricey. But he didn't want those because they're too casual. So I handled it the way we learned in training, you know, always try to make the customers happy—I told him to try on the ones he liked and if he liked them, we could ask the manager about giving him a special discount. Turns out we didn't need to do that because he didn't like the way they fit. Anyway, I was thinking that if we're going to keep that hanging sale sign, we should only put sale items on the table beneath it. It would save us some trouble and make the customers happier, too.

Items Used With Both Versions

05 What is the store's problem?

a. The staff doesn't know which shoes are discounted.
b. The store doesn't sell formal shoes.
c. The customers are confused about a sales display.
d. The manager is unavailable to answer questions.

06 Which shoes are on sale?

a. only the black shoes
b. all the shoes on the table
c. all the formal shoes
d. the shoes under the sign

07 What did the woman offer to ask the manager?

a. whether she can offer a special price
b. when the formal shoes go on sale
c. when the customer can talk to him
d. whether other formal shoes are available

08 What does the woman mean when she says:
[audio only: They're a lot less pricey]

a. The shoes are more attractive.
b. The shoes are more affordable.
c. The shoes are more formal.
d. The shoes are more comfortable.

APPENDIX C

Pair 3, Dialogic Version Script

Narrator: Listen to a conversation between two friends.

Woman: Tony, have you heard about a book called "The Runner in You"?

Man: I think somebody at school was talking about it. But remind me, what's it about?

Woman: It's about the writer, Jerry Hampton, a guy who started running about fifteen years ago. But he kept getting injured. Then one day he came across a magazine article about the Tarahumara tribe in Mexico.

Man: Oh, I've heard about them. They have those incredible 50-mile races that they run in just a pair of old sandals. Those guys are great runners.

Woman: Exactly. And when Hampton started reading more about them, he learned that the Tarahumara rarely suffer the kinds of injuries that you and your friends that run always get – even though the Tarahumara run a lot more. Hampton's saying that when we put on shoes, we interfere with the body's natural design – which is perfect for running.

Man: That sounds plausible. But this guy Hampton's not a doctor, is he?

Woman: No, in the book he's clear about his lack of medical credentials. But I think that some real scientists will do some experiments soon—and in the meantime he's advocating that everyone start running barefoot.

Pair 3, Monologic Version Script

Narrator: Listen to part of a radio report about a book.

Woman: Today I want to talk about a fascinating book that I just read. It's called "The Runner in You". It's about the writer, Jerry Hampton, a guy who started running about fifteen years ago. But he kept getting injured. Then one day he came across a magazine article about the Tarahumara tribe in Mexico, a group of people who regularly run 50-mile races for fun. In handmade sandals! He was amazed and decided to do some investigating. He soon learned that the Tarahumara runners were not suffering from the kinds of foot problems that people who wear expensive shoes to "protect" their bodies were. That is, people like himself who wear special shoes to protect their feet from the effects of running. So Hampton decided to start running barefoot. He's now run two marathons without suffering from any of the injuries that bothered him when he wore expensive running shoes. So his hypothesis is that the shoes are causing the problems, that the natural structure of our feet and bodies enables us to run long distances without harm. This makes sense to me. I've suffered a lot of running injuries myself over the years—all while wearing expensive shoes. But I must say, Mr. Hampton is not a doctor. In the book he's clear about his lack of medical credentials. But I think that some real scientists will do some experiments soon—and in the meantime he's advocating that everyone start running barefoot.

Items Used With Both Versions

09 What is the main point of Jerry Hampton's book?

a. Running shoes are not good for runners.
b. Running barefoot is dangerous.
c. Runners need to see their doctors more often.
d. Everyone can run a marathon.

10 What does Jerry Hampton have in common with other runners?

a. He thinks more expensive shoes are safer.
b. He has gotten injured while running.
c. He was inspired to run by the Tarahumara.
d. He started running without talking to a doctor.

11 What does the woman think will happen in the future?

a. More Tarahumara will run in marathons.
b. More regulations will be created for marathons.
c. Researchers will test Hampton's theory.
d. All marathoners will run barefoot.

12 What does the speaker mean when she says:
[audio only: In the book he's clear about his lack of medical credentials]

a. Hampton is a better runner than most doctors.
b. Hampton is not a qualified doctor.
c. Hampton will write a book about sports medicine.
d. Hampton no longer agrees with his own theory.