ĐÁNH GIÁ KHẢ NĂNG NGHE HIỂU CÁC THÔNG TIN TƯỜNG MINH VÀ KHẢ NĂNG SUY LUẬN TRONG KHI NGHE: CÁC CÂU HỎI TRẮC NGHIỆM CÓ THẬT SỰ HIỆU QUẢ?

Trần Thị Ngọc Diệp*[*], Phạm Ngọc Thạch[**]

Câu hỏi trắc nghiệm (TN) là một dạng thức câu hỏi rất phổ biến trong các bài thi Nghe. Tuy nhiên, có rất ít tài liệu đề cập tới tính hiệu quả của các câu hỏi TN trong việc đo lường khả năng nghe của người học. Nghiên cứu này đánh giá mức độ hiệu quả của các câu hỏi TN trong việc kiểm tra khả năng nghe hiểu các thông tin tường minh, và khả năng suy luận từ các thông tin nghe được (Buck, 2001). Nhóm nghiên cứu chúng tôi yêu cầu người nghe trình bày lại những suy nghĩ của mình trong quá trình họ vừa nghe vừa trả lời các câu hỏi TN. Đây là một nghiên cứu định tính dựa trên dữ liệu là phần trình bày của 10 sinh viên chuyên ngành tiếng Anh tại một trường dại học ở Việt Nam. Kết quả cho thấy tuy các câu hỏi TN đo lường tốt khả năng nghe thông tin tường minh, dạng thức này lại tạo điều kiện để người nghe sử dụng các chiến thuật làm bài thi khi trả lời các câu hỏi suy luận. Điều này đã làm hạn chế tính giá trị của các câu hỏi đánh giá khả năng suy luận trong khi nghe. Chúng tôi đã đưa ra những đề xuất cụ thể giúp cho người ra đề viết các câu hỏi TN nhằm đánh giá tốt hơn khả năng suy luận của người nghe. Những đề xuất này bao gồm việc chuẩn bị bài nghe, thiết kế tiểu mục và hình thức nghe hai lần. Bài viết cũng đề cập tới những khó khăn trong việc đánh giá khả năng suy luận của người nghe.

Từ khóa: *bài thi nghe, kỹ năng nghe, câu hỏi trắc nghiệm, tính giá trị.*

Multiple-choice questions (MCQs) are a widely accepted format used in second language (L2) listening tests. However, there is a lack of literature on the usefulness of this format in measuring the listening construct. This study investigates the extent to which MCQs engage two important components of the default listening construct: understanding explicit information and drawing inferences from the spoken input (Buck, 2001). Verbal reports were employed to explore the participants' cognitive processes as they answered multiple-choice test items. Taking a qualitative approach, this study draws on the think-aloud protocols provided by 10 English majors from a university in Vietnam. The results show that while the multiple-choice format successfully measured the ability to understand explicit information, it promoted the use of test-taking strategies that compromised cognitive validity of the test items that targeted the ability to draw inferences. Practical suggestions were given so that MCQs can be better prepared to engage the inference-making process. These included input text preparation, item design, and double-play format. The challenges in measuring L2 learners' ability to make inferences while listening were also discussed.

Keywords: *listening tests, listening construct, multiple-choice questions, validity.*

[*] **TS., Trường Đại học Victoria - Wellington (New Zealand); Khoa tiếng Anh, Trường Đại học Hà Nội**

[**] **TS., Trường Đại học Hà Nội**

Email: diepdhhn@gmail.com

ASSESSING THE ABILITY TO UNDERSTAND EXPLICIT INFORMATION AND DRAW INFERENCES IN L2 LISTENING: HOW USEFUL IS THE MULTIPLE-CHOICE FORMAT?

I. Introduction

Among the four language skills, listening is arguably the most difficult skill to assess (Field, 2013) because of its unique and often stressful nature. One of the challenges facing developers of any listening test is the unenviable choice of test formats. Testing textbooks often provide lists of testing methods that can be used to assess listening but mostly without discussing the usefulness of these methods in measuring the listening construct (Barta, 2009). So as to examine how effectively a certain task type engages the targeted listening construct, a number of studies have taken the mentalistic approach to investigate test-takers' cognitive processes. For example, Field (2009) found that while answering IELTS multiple-choice and gap-fill questions, test-takers primarily engaged in lexical processing and not in higher-level cognitive processes such as inferencing or building a structural representation of the listening input. These findings are enlightening and necessary for listening test developers. Nonetheless, studies that seek to unveil the relationship between task type and test-takers' thought processes are still rare in the literature on listening assessment. This study ventures into that under-researched area, focusing on the multiple-choice format, which has been widely used in tests of second language (L2) listening.

The multiple-choice format is often chosen for listening tests not only because it is convenient but also because it helps to save cost (Yanagawa & Green, 2008). However, its usefulness in measuring listening ability has not been well explored. To investigate how well the multiple-choice format measures the targeted listening abilities, it is necessary to examine the thought processes that L2 listeners undergo as they listen to the spoken text and answer the given multiple-choice questions (MCQs). Verbal report is the most appropriate option since this method lends itself to the investigation of the thinking behind different types of performance (Ericsson & Simon, 1993). In this present study, this method is used to investigate how useful the multiple-choice format is in measuring the ability to understand explicit information and draw inferences, the two fundamental components of the default listening construct proposed by Buck (2001).

Literature review

Defining the listening construct

A common approach in defining the listening construct is to view it as a taxonomy of subskills. The simplest taxonomy is proposed by Carroll (1972) based on the two-stage view which divides listening into two processes:

- apprehending linguistic information (lower-level processing);
- relating the information to a broader context (macro-comprehension).

This early taxonomy is extended in Weir's (1993) framework for testing listening in which direct meaning comprehension and inferred meaning comprehension are the key components. Direct meaning comprehension includes but is not limited to:

- listening for gist;
- listening for specifics;
- determining speaker's attitude/intentions toward listener/topic where obvious from the text.

Inferred meaning comprehension, on the other hand, involves listening abilities such as:

- making inferences and deductions; evaluating content in terms of information clearly available from the text;
- relating utterances to the social and situational context in which they are made.

Buck (2001) emphasizes that "there are no hard-and-fast rules about what is an appropriate listening construct" (p. 112) because this decision depends largely on the purpose of the test, the target-language use situation, or the available resources. He, therefore, proposes a default listening construct that captures the key listening abilities and avoids most of the context-dependent elements. This default listening construct is defined as the ability to:

- process extended samples of realistic spoken language, automatically and in real time,
- understand the linguistic information that is unequivocally included in the test, and
- make whatever inferences are unambiguously implicated by the content of the passage.

(Buck, 2001, p. 114)

Although this description of the listening construct is rather simplistic, it fits in well with the purpose of this study which aims to examine the usefulness of MCQs in measuring a general, context-free listening construct. It should be acknowledged that in most of the English listening tests these days, the ability to process realistic samples of spoken language such as everyday life conversations or academic lectures is often expected from the test-takers regardless of the test format. For that reason, this study only investigates the extent to which MCQs engage the ability to understand explicit information and the ability to draw inferences from the spoken input.

Research on the use of MCQs in listening tests

The impacts of the multiple-choice format on test-takers' performance on L2 listening tests have been the focus of many research studies. Some of these studies sought to compare MCQs with other formats. For example, in a study that focused on assessing comprehension of authentic texts in French, Eykyn (1992) compared four task types (choose-the-picture, MCQs, Wh-questions, and vocabulary list) and found that beginning learners of French performed best with MCQs.

A different line of research explored the impacts of the ways in which MCQs are presented in a listening test. Yanagawa & Green (2008) studied the effects of the following three types of MCQs:

- Full question preview (FQP): Both item stem and response options are presented before listening.
- Answer option preview (AOP): Response options are shown before listening, but the questions are heard after the input text is played.
- Question stem preview (QSP): Only item stem is displayed before listening. Response options are heard after the input text is played.

They found that the group of students who answered MCQs under AOP condition performed worst. No significant difference was detected between the students who were in FQP and QSP conditions. In a similar study, Hemmati and Ghaderi (2014) investigated test-takers' performance under four conditions, namely FQP, AOP, QSP, and NP (no preview). It was found that there was a significant difference between NP and the other three variations of MCQs, which implies that the opportunity to preview MCQs can facilitate comprehension.

The previously reviewed studies explored the impacts of MCQs on merely test-takers' scores. However, the quantitative approach that they took cannot uncover the cognitive processes underlying test-takers' performance and how the multiple-choice format shapes such processes. Until recently, only a limited number of studies set out to explore the cognitive validity of multiple-choice listening test items. By means of immediate retrospective verbal reports, Wu (1998) found that previewing the questions and answer options seemed to help more advanced test-takers form anticipations of the incoming input and provide foci for listening. However, this was not the case for less able subjects. Wu (1998) also pointed out that the multiple-choice format allowed much uninformed guessing which sometimes led to the test-takers' selection of the right answer for the wrong reason. Therefore, the validity of MCQs in listening tests is left open to question.

Another study that casts doubt on the usefulness of MCQs is the one that investigated the cognitive validity of lecture-based questions in the IELTS Listening paper by Field (2009). The participants' responses to the MCQs revealed that this format promoted a great level of test-wise strategy use. Field (2009) noted that while answering the multiple-choice items, the participants were engaged in "a process of checking information against pre-established cues rather than a more ecological one of receiving, interpreting, and organizing it" (p. 35).

From the cognitive point of view, Wu's (1998) and Field's (2009) studies shed light on the undesirable impacts of MCQs when used in listening tests. While the multiple-choice format offers a number of practical advantages and seems easier for test-takers than other formats, it has been found to change the nature of listening and promote test-taking strategies which are irrelevant to the listening construct. Given the popular use of MCQs in listening tests, much more research is needed to further investigate the cognitive validity and unwanted impacts of this format. Studies of this kind would be beneficial for test developers and language teachers; however, they are relatively rare. This present research aims to address the issue by examining the extent to which MCQs engage the two fundamental components of the default listening construct (Buck, 2001).

Research questions

This study seeks the answers to the following research questions:

1. To what extent does the multiple-choice format engage the ability to understand explicit information?

2. To what extent does the multiple-choice format engage the ability to draw inferences?

By means of verbal reports, the study examines how well the listening construct is represented by MCQs, providing practical recommendations for the use of this format in L2 listening tests. Here and throughout, the test items that target the ability to understand explicit information are termed EI items. The ones that target the ability to draw inferences are called DI items.

Methodology

Material and participants

The listening test used for this study was a component of a high-stakes English proficiency test designed by a team of experts from a university in Vietnam. It targeted Vietnamese learners of English aged 18 and above, and was intended for use solely within country. The test consisted of three parts with 35 multiple-choice questions. Test-takers were allowed to see both the questions and the response options prior to listening. The recording was played once only.

This study recruited a total number of 10 participants who were English majors at a university in Vietnam. All of them were female students whose ages ranged from 19 to 21. To protect their identities, a pseudo name was given to each participant. It should be noted that the participants were not selected according to any set of criteria. Since the nature of the study was exploratory, any subject could be appreciated for the insights they could provide about what was going through their minds as they were tackling test tasks. In this case, it is more useful to have participants who were interested in the study and willing to give in-depth information of their cognitive processes (Charters, 2003). This best describes the 10 participants who voluntarily chose to be part of this project.

Procedure

A team of experts consisting of five reviewers was first invited to review and categorize test questions into EI items and DI items. The reviewers were Vietnamese teachers of English from five different universities in Vietnam. At the time of the study, they all had more than five years of teaching experience and held Master's Degrees of Teaching English to Speakers of Other Languages (TESOL). In the reviewing session, the reviewers were first asked to give their own judgement on each test item. Their judgments were based on the pre-determined descriptions for EI and DI items. In the particular context of this listening test, EI and DI items are described as follows:

- EI items target the ability to comprehend the linguistic information in the input text (Caroll, 1972).
- DI items target the ability to use the acoustic, lexical, contextual information contained in the listening input, and background knowledge to (a) extract necessary information which was not explicitly given to the listeners, (b) understand the implied meaning of the input text (Buck, 2001; Gui, 2015).

The reviewers' decisions on whether an item was EI or DI were then shared within the group. Whenever there was disagreement, they kept discussing until a consensus was reached. The outcome of the discussion was two separate lists, one for EI items and the other for DI items.

To collect verbal reports for this study, each participant was invited to a one-on-one meeting with the primary researcher. In this meeting, the participant was asked to listen to the recording and answer test questions. Whenever an answer was chosen, the primary researcher paused the recording and asked the participant to report the thinking behind her decision. This approach known as immediate retrospective verbal report helps to minimize the effect of memory. Wu (1998) stresses that when verbal reports are collected this way, they provide information that closely reflects the participants' actual thought processes. The language used in all the verbal reports was Vietnamese. Once collected, these reports were transcribed and translated into English by the primary researcher. Since the investigation of cognitive processes engaged by EI and DI items was exploratory, a pre-defined set of codes was not created. Instead, the codes arose directly from the verbal reports. Apart from the primary researcher, another coder was invited to code one entire verbal report. The exact agreement between coders was 82.9% which indicated acceptable inter-coder reliability.

Findings

Among 35 MCQs of the test, the item reviewers agreed on a list of 12 items that measured the ability to understand explicit information (EI items). There was a consensus that the remaining 23 items were designed to engage the ability to draw inferences (DI items). These EI and DI items are listed in Table 1.

Table 1. List of EI and DI items

EI items	Items 1, 2, 4, 6, 7, 13, 15, 23, 24, 25, 26, 27
DI items	Items 3, 5, 8, 9, 10, 11, 12, 14, 16, 17, 18, 19, 20, 21, 22, 28, 29, 30, 31, 32, 33, 34, 35

This section reports on the participants' thought processes as they answered the test items. These processes are illustrated by examples from the test and the verbal data. For each example, the question and the response options are provided first with the correct answer in bold. The audio script is then presented. Excerpts from the participants' verbal reports are numbered and italicized.

Responses to EI items

For most of the EI items, the participants' responses showed that their comprehension of the input at a literal level was simply enough to answer the questions. Only Items 13 and 24 were found to activate both input comprehension and strategy use. Table 2 provides a summary of the reported cognitive processes and strategies.

Table 2. Cognitive processes and strategies reported for EI items

Cognitive processes	No. of instances (percentage)
Comprehension of the input at literal level	109 (87.9%)
Elimination strategy to support a decision	15 (12.1%)

As seen from Table 2, the participants' focus of attention was primarily on the auditory input. A total number of 109 instances indicated a common process in which the participants demonstrated their comprehension by correctly retelling or paraphrasing the input information. They confidently chose the correct answers without re-reading the questions or considering other response options. The responses to Item 1 are typical examples as follows.

Item 1

How does the man travel to Liverpool?

A. By train

B. By bus

C. By car

D. By plane

Auditory input:

How does the man travel to Liverpool?

- Excuse me, I'd like to go to Liverpool on Friday.
- Well, you can go by bus or train.
- Is the train expensive?

- Yes, the bus is much cheaper. It's only 20 pounds.

- Right, that's better for me. Can I have a ticket, please?

- Yes, certainly.

Below are the two responses to Item 1 characterised by input understanding:

(1) Chi: *I choose B, by bus. In the conversation, the woman gave the man two options, train and bus. She also said that train was more expensive and bus was much cheaper. So, the man chose to go by bus because he said it was better for him.*

(2) Trang: *My answer is B, by bus because when the woman gave the man two options, he asked whether the train was expensive. She said yes and added that bus was much cheaper. He said it was a better choice. So, he chose to go by bus and after that he bought the ticket.*

Chi and Trang paraphrased the input conversation in different ways but the overall meaning remained the same. Their good comprehension of the input justified the validity of this item. A similar thinking process was reported for Items 2, 4, 7, 15, 25, 26 and 27. Thus, it became apparent that these items successfully engaged the ability to understand explicit information.

In terms of strategy use, there were 15 instances of the elimination strategy and no evidence of other test-specific behaviours (Table 2). Most of the instances for the elimination strategy were found in the participants' responses to Item 13.

Item 13

When Nick and Mel were younger,

A. they studied music at school.

B. their father took them to live concert.

C. they did not like music.

D. their mother encouraged them to play music.

Auditory input:

- I'm talking to Nick Parker, the singer with the band called Krispy. Nick, your sister, Mel plays guitar in the band, too, doesn't she?

- Yeah, Mel is a year younger than me. We've been playing and singing together since we were eight, nine. Dad is a guitarist. He took us to hear great bands playing live. Mel and I put on shows at school which was a lot of fun. Mom thought we were good but she didn't want us to get too serious about our music because of the hard lives professional musicians have.

It is clear that both the correct answer and distractors shared certain lexical items with the input text. That might help to explain why some subjects hesitated and turned to the elimination strategy for support. For example:

(3) Thu: *He said they played music when they were 9. He mentioned his father and live concert but I'm not so sure. He* *didn't mention whether they studied music at school. "They did not like music" does not seem right. He mentioned their mother. She thought they were good but she didn't want them to be serious about it. I'll choose B.*

Thu understood the input well enough to disconfirm the three distractors but she was uncertain about the targeted information. Thus, her successful performance on this item was not fully attributed to input comprehension. The multiple-choice format and the written response options did indeed play a crucial part in helping her achieve the correct answers.

Thus far, the answer to Research Question 1 has been mostly straightforward and positive. With 87.9% of the responses (Table 2) indicating that comprehension of input at a literal level was sufficient to answer EI items, the multiple-choice format could be said to achieve cognitive validity since it effectively engaged the ability to understand explicit information. The use of elimination strategy to support the decision-making process was limited to only 15 instances in total.

Responses to DI items

The cognitive processes underlying the answers to each DI item were markedly more complex than for the EI items. Table 3 provides a tally of different cognitive processes and strategies emerging from the verbal data.

Table 3. Cognitive processes and strategies reported for DI items

Cognitive processes and strategies	No. of instances (%)
Drawing inferences from the spoken input	185 (55.2%)
Drawing inferences from key words	38 (11.3%)
Drawing inferences using background knowledge	18 (5.4%)
Drawing inferences using personal opinions	9 (2.7%)
Lexical matching	28 (8.4%)
Elimination strategy to support a decision	39 (11.6%)
Elimination strategy based solely on the written text	5 (1.5%)
Random guessing	13 (3.9%)

The data from Table 3 suggest a noticeable degree of mismatch between the processes targeted by DI items and those reported by the participants. The total number of 185 instances which indicated that the listeners drew inferences from their comprehension of the spoken input only accounted for 55.2% of all recorded instances for DI items. While this cognitive process remained dominant in the participants' answers to DI items, what was striking was the number of other processes and test-taking strategies involved in the responses to these items. There was evidence that the participants' inference-making process was aided by factors other than the auditory input. When the subjects' comprehension of input was not adequate to lead to an answer, they reportedly clung to a key word or phrase identified from the stream of speech. In certain cases, their personal opinions and background knowledge were also brought into play. It should be noted that the dependence on key words, background knowledge, or personal opinions can be seen as strategic behaviours which help to compensate for gaps in understanding (Field, 2009; 2013). These behaviours reflect L2 listeners' struggle in a real-world setting, and thus should not be taken as threat to cognitive validity (Field, 2013).

The behaviours that raised concern were the three test-taking strategies which reflected the participants' dependence on the written text. As shown in Table 3, the use of lexical matching, elimination strategy, and random guessing accounted for 25.4% of all instances reported for DI items. Lexical matching was a common strategy shared across participants who lost track of the spoken input and failed to establish the text's overall meaning. These participants admittedly chose the response options that matched the only words or phrases they could hear.

As evidenced by the verbal data, the participants most often utilized the elimination strategy to help them make the final decisions. In this case, the listeners acquired only a partial understanding of the auditory input and from there, they went on to make inferences that helped them eliminate certain response options. They relied heavily on the written input without which the final answer might not have been determined. Another behaviour that also raised concern about validity was the one in which the subjects totally depended on the written text, drawing inferences in light of the test questions and response options. However, with only five occurrences, it was not a major problem among this group of participants.

In most of the responses to DI items, the processes and strategies mentioned above were intertwined in many different ways. The cognitive demand of these items appeared to push the listeners to activate both their linguistic and non-linguistic knowledge. Under the real-time constraint, the participants utilized multiple strategies to tackle the test questions. The participants' answers to Items 10 and 19 provide insights into the complexity of responses to DI items.

Item 10

What do Wendy and her mum disagree about?

A. Whether or not Wendy covered her eyes.

B. Whether or not they will see the film again.

C. Whether or not the film was P.G rated.

D. Whether or not the film was frightening.

Auditory input:

Adrian: You just didn't like it because you were frightened.

Wendy: No, I wasn't. What are you talking about?

Adrian: I saw you covering your eyes.

Wendy: I wasn't covering my eyes. I was laughing.

Mrs. Turner: Well, I couldn't look sometimes. I mean it was only a PG film but some of the scenes were well pretty.

Wendy: You mean hilarious. Well, at least it wasn't too long.

In response to Item 10, although some of the participants were able to pick up clues from the spoken input, the location of such information was often assisted by the given response options. Much of the processing involved the participants drawing inferences from several cues located in the spoken input and then eliminating options that did not match their inferences. For example:

*(4) **Vy:** Wendy and the boy were arguing about whether Wendy was afraid and covering her eyes, So, I will exclude A and D because it was not with her mom. And her mother mentioned P.G rating but I'm not sure whether they were arguing about it. But they did not mention whether or not they will see the film again. So, I choose C, whether or not the film was P.G rated because it sounds better.*

*(5) **An**: I think A and C are possible because they mentioned whether Wendy covered her eyes and P.G rated. I am in favour of A since they talked more about it. They did not talk a lot about P.G rated and it did not sound like they were arguing with each other about it.*

It should be pointed out that not all the participants were able to distinguish between the three speakers in the input conversation. Unlike Vy, An did not seem aware that Adrian, Wendy, and her mother were all involved in this short discussion. With more than two voices in the auditory input, it is more challenging for the listeners to recognize the speakers and remember what each of them is saying (Green, 2017, p. 18). Not surprisingly, some other participants reportedly lost track of the input and turned to lexical matching to figure out the answer.

(6) ***Nga****: I don't understand much but I choose D, whether or not the film was frightening because I could hear the word "frightened" in the conversation.*

(7) ***Mai****: If I have to choose, I'll choose C because I heard P.G.*

The responses to Item 10 raised concern for validity since they demonstrated the listeners' dependence on the written text. Under the influence of the item wording, the focus of the participants' attention was mainly at word level. The number of speakers in the input text was shown to contribute to the items' high cognitive demand.

Item 19

Roger regards his early days in business as

A. frustrating

B. demanding

C. irrelevant

D. boring

Auditory input:

- And what kind of success did you have in the early days?

- You could say it was a bit like taking a roller coaster ride and wondering when you're going to come flying off at breakneck speed. Everything was a challenge, finance, production, marketing.

In Item 19, the absence of lexical overlap between the response options and the input text made it more difficult for the listeners to locate the necessary information. With very little guidance from the written text, some listeners constructed their own hypotheses, relying on a particular key word. For example:

(8) Chi: *He compared his early days in business with a roller coaster. It goes up and down, so it can never be boring or irrelevant. Demanding is more suitable than frustrating because ... roller coaster goes up and down. Like there are many demands that he has to fulfil. So, I choose demanding.*

(9) Vy: *Roger compared his early days to a roller coaster ride. He was spinning. He didn't seem to know which direction to take. I think he was frustrated. I choose A.*

Both Chi and Vy picked up "roller coaster" from the spoken input and treated the word in a spotlight fashion. Interestingly, they each interpreted this word through their own lens, which resulted in different inferences. While Chi associated "roller coaster" with an "up and down" movement, Vy believed that this word implied a "spinning" movement. The hypotheses they constructed using their personal opinions led them to make different decisions on the final answers. These examples have served to highlight the fact that the inference-making process may go in unexpected directions once the participants' personal beliefs are activated while listening.

The use of personal opinions was also reported by participants who were entirely reliant upon the written input. These participants sought to eliminate the response options which, in their opinions, were incorrect. They ended up choosing their final answer simply because it was the option that did not get eliminated. For instance:

(10) Thu: *I couldn't hear anything for Item 19. I think he achieved some success in his early days but it was not significant. So, frustrating and demanding are not suitable. Not boring because if he was successful, how can it be boring? So, I think it's C, irrelevant.*

The evidence of participants' dependence on the written input suggested that the validity of Item 19 was highly questionable. Without the lexical overlap between the auditory input and the response options, the participants got lost quite easily and became reliant upon factors other than the spoken text.

In answering Research Question 2, the verbal data showed that the multiple-choice format did help to bring out the listeners' ability to draw inferences from the spoken input but at the same time, promoted the use of test-taking strategies (Table 3). These findings suggest that the usefulness of the multiple-choice format in measuring the ability to draw inferences be reconsidered. The format itself was found to shape the participants' thinking processes in ways that did not reflect the targeted construct. There was also evidence showing that the subjects' performance on DI items was influenced by the input text and the design of test items.

Discussion

The insights obtained from the verbal protocols made it possible to gauge the extent to which the multiple-choice format engaged two essential components of the default listening construct: understanding explicit information and drawing inferences from the spoken input. While the format was shown to successfully measure the former, it was found to be less useful in assessing the latter. Certain issues emerging from the analysis of the verbal data are deemed beneficial for developers of L2 listening tests.

Undesirable processes promoted by the multiple-choice format

One of the undesirable processes observed in the responses to both EI and DI items was the use of item wording to identify information from the input text and to decide on a particular answer (Excerpts (3), (4), (5), (6) and (7)). In many cases, the participants were listening to check whether the information from the auditory input matched one of the response options. This finding is consistent with that by Field (2009) who also found that, when answering MCQs, the listeners' thinking operated in the direction: written lexical input – spoken lexical input. As a consequence, the listening process was influenced and guided by the expectations that listeners have from the questions and the response options. This result accords with Weir's (1993) concern that the nature of the listening process is likely to be changed when the questions are provided prior to listening. Moreover, evidence from the verbal data lends further support to Yanagawa and Green's (2008) observation that the opportunity to preview response options may promote the use of lexical matching strategy.

That being said, it is worth pointing out that the participants did not always use item wording to direct their listening process. In their responses to most of the EI items, the subjects naturally let their comprehension lead the way to the correct answers (Excerpts (1) and (2)). This evidence suggests that when the listeners understood the input well enough, their listening process was not influenced by the item wording. Therefore, the opportunity to preview the questions and response options is not always a threat to validity. It became more of a concern when the listeners' comprehension of input was inadequate for them to arrive at an answer.

Strategy use

Field (2013) points out that while listening, test-takers may be engaged in two types of strategies which are communication strategies and test-wise strategies. Communication strategies such as using background knowledge or drawing inferences from key words are used to compensate for gaps in understanding and can be viewed as an important part of L2 listening proficiency (Field, 2013). Meanwhile test-wise strategies are adopted to exploit loopholes in the format (Field, 2009). This section discusses the test-wise strategies which were identified from the verbal data and often associated with the multiple-choice format.

Two noteworthy trends presented themselves regarding the participants' use of test-taking strategies while answering the test items. First, evidence of strategy use was mainly found in responses to DI items. In many cases, the auditory input was so challenging that the participants could only process it at a very local level. Therefore, they showed themselves unable to make any plausible inferences. Instead, they turned to the written clues, using the word matching or elimination strategy (Excerpts (6) and (7)). Apart from that, some DI items imposed a heavy cognitive load which involved listening to the input, reading the response options, drawing inferences, and making quick decisions almost at the same time. Given the stressful nature of the single-play mode, participants reportedly sought help from the written input and eventually arrived at the answers after the process of elimination (Excerpts (4) and (5)).

Second, the elimination strategy was reported with the highest frequency compared to lexical matching and random guessing. This study elicited evidence of different ways in which the elimination strategy was utilized to answer both EI and DI items. In case the participants had only a partial understanding of the input, the strategy was used to aid the inference-making and decision-making process (Excerpts (4) and (5)). This approach limited the test validity since the participants were shown to rely heavily on the written response options. The correct answer was sometimes achieved simply as a result of elimination. It could be argued that if a different format had been used, a candidate who was unable to make inferences from the listening input would have had no or a very low chance of coming up with the right answer.

In general, elimination can be argued to be on the fringe of the listening construct. L2 listeners might, at times, need to choose between the hypotheses they construct about what they hear. However, the multiple-choice format brings this strategy forward and encourages the listeners to use it more frequently than they are likely to in real life. As this study unveiled, in responses to both EI items and DI items, the elimination strategy accounted for a relatively substantial percentage of all reported cognitive processes (12.1% for EI items and 13.1% for DI items). The finding suggests that the multiple-choice format may well have elicited the elimination behaviour more than needed.

Assessing the ability to draw inferences

As this study reveals, the use of MCQs might not be ideal in measuring the ability to draw inferences from the spoken input.. Nevertheless, it is only fair to acknowledge that inference-making, as a higher-level cognitive skill in listening (Rost, 1994), is very difficult to assess in general. Oftentimes, inferring while listening is challenging because, unlike readers who can refer back to the passage and take time to make an inference, listeners have to process the input information as the recording goes, trying to retain, synthesize or relate the information to their prior world knowledge. While it is often considered a core element of the listening construct, there has been very limited research conducted on inference-making in L2 listening. Perhaps, for this reason, the question as to how this crucial listening ability can be effectively assessed has been left unanswered. This study partly addressed the issue by providing evidence that MCQs, to a certain extent, are useful in engaging the ability to draw inferences; however, they gave rise to the use of test-taking strategies. Other variants such as the number of speakers in the input text or the reading load of an MCQ did contribute to increasing the cognitive demand of some DI items, causing unnecessary difficulty for test-takers.

Recommendations

In light of the verbal protocols, a number of suggestions can be given to make MCQs a better measure of the inferential ability. As far as the input text is concerned, attention should be given to the number of voices in a conversation. In Item 10, the participants' reports indicated that the inclusion of three speakers in the input conversation contributed to increasing the cognitive load of the item (Excerpts (4) and (5)). It might have been the test developer's intention to assess the test-takers' ability to synthesize, evaluate, and select the necessary information to make an appropriate inference. However, since human attention is limited, it is unrealistic to expect candidates to demonstrate high-level thinking while reading four relatively long alternatives in a situation where no second chance is granted. It is advisable to have two speakers in an input conversation, ideally one male and female speaker so that listeners can easily distinguish the speakers' voices (Weir, 2005).

The lexical overlap between the input text and the response options was shown to promote shallow processing of the input rather than high-level thinking and inference making (Excerpts (6) and (7). However, not providing any lexical overlap was not ideal, either. The participants' responses to Item 19 indicated that without some guidance from the response options, they easily lost track of the input and failed to make any plausible inferences (Excerpts (8), (9) and (10)). It could be argued that under test condition, a certain degree of lexical overlap would still be necessary despite its threat to validity. However, more research is needed to shed further light on how much lexical overlap should be provided in the response options to better engage the ability to draw inferences.

Attention should also be paid to reducing the reading load of MCQs by giving shorter response options (Field, 2009) and perhaps fewer alternatives for each question. The literature related to MCQs, in general, shows many studies in favour of three-option MCQs because this format improved content coverage while not affecting the psychometric quality of a test (Rodriguez, 2005).

Making inferences while listening is challenging for L2 listeners because of their limited cognitive ability and the lack of time. It can therefore be argued that with the double-play format, listeners will have more time to be engaged in higher-level cognitive processes like inferencing. Field (2009) maintains that when listening to the input for the first time, test-takers are often able to locate the necessary information. However, it is the second listening that gives them the chance to review, synthesize information and make plausible inferences.

Limitations and further research

The data obtained from retrospective verbal reports provided rich and useful insights into the test-takers' cognitive processes. However, this method is time-consuming, labour-intensive, and therefore, suitable for the study of a relatively small group of participants. The small sample size of only 10 subjects certainly limited the generalizability of the study's findings. It should also be stressed that the participants were invited to take a listening test in a non-test condition. Thus, their performances were not affected by factors like time pressure or test anxiety as in a real test event. The non-test condition was a trade-off in the design of this study because the insights into the listeners' cognitive processes were prioritized at the expense of authenticity.

This paper explores the relationship between the multiple-choice format and the default listening construct with a firm belief that the test method should lend itself to the targeted construct (Haladyna & Rodriguez, 2013). The challenges in assessing the ability to draw inferences while listening call for further research. It would be valuable to use introspective methods to investigate the extent to which the inferential ability is engaged by other popular formats such as gap filling, multiple-matching, or short answer questions. Green (2017) reminds us that choosing the most appropriate test method to measure a particular construct is not always obvious. She adds that, to some extent, it becomes easier with experience. However, we would argue that the decision on the test format should be primarily informed by research. There might not be an ideal format to measure the ability to draw inferences while listening. However, studies of this kind will enable item writers to make informed choices of test formats and critically consider various aspects of item design to better engage this elusive listening ability.

REFERENCES

1. Aryadoust, V. (2018). Taxonomies of listening skills. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1–8). John Wiley & Sons, Inc.

2. Barta, E. (2009). Analysis of listening comprehension assessment task. In G. Szabo, J. Horvath, & M. Nikolov (Eds.), *UPRT 2009: Empirical studies in English applied linguistics* (pp. 65-80). Linguia Franca Csoport.

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

4. Carroll, J. B. (1972). Defining language comprehension: Some speculations. In R. O. Freedle & J. B. Carroll (Eds.), *Language comprehension and the acquisition of knowledge* (pp. 1-29). Winston & Sons.

5. Charters, E. (2003). The use of think-aloud methods in qualitative research: An introduction to think-aloud methods. *Brock Education: A Journal of Educational Research and Practice*, *12*(2), 68-82.

6. Elliott, M., & Wilson, J. (2013). Context validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 152-241). Cambridge University Press.

7. Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. The MIT Press.

8. Eykyn, L.B. (1992). *The effects of listening guides on the comprehension of authentic texts by novice learners of language*

[Unpublished doctoral dissertation]. University of South Carolina.

9. Field, J. (2009). The cognitive validity of the lecture-based question in the IELTS listening paper. *IELTS research reports, 9*, 17–65.

10. Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77-151). Cambridge University Press.

11. Green, R. (2017). *Designing listening test*. Spinger.

12. Gui, J. (2015). Inference-making and linguistic skills in listening comprehension: An observation of French students learning Chinese. *Electronic Journal of Foreign Language Teaching, 12*(1), 318-331. https://e-flt.nus.edu.sg/wp-content/uploads/2020/09/v12s12015/guo.pdf

13. Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

14. Hemmati, F., & Ghaderi, E. (2014). The effect of four formats of multiple-choice questions on the listening comprehension of EFL learners. *Procedia - Social and Behavioral Sciences, 98*, 637–644. http://dx.doi.org/10.1016/j.sbspro.2014.03.462

15. Rodriguez, M. C. (2005). Three-options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

16. Rost, M. (1994). On-line summaries as representations of lecture understanding. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 219-238). Cambridge University Press.

17. Weir, C. J. (1993). *Understanding and developing language tests*. Prentice Hall.

18. Weir, C. J. (2005). *Language testing and validation*. Palgrave McMillan.

19. 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. https://doi.org/10.1177/026553229801500102

20. 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. http://dx.doi.org/10.1016/j.system.2007.12.003

(Ngày nhận bài: 14/9/2022; ngày duyệt đăng: 19/10/2022)