CaMLA Working Papers

2018–02

# Investigating test-taking strategies during the completion of computer-delivered items from the Michigan English Test (MET): Evidence from eye tracking and cued retrospective reporting

Ruslan Suvorov
Center for Language & Technology
University of Hawai'i at Mānoa

# Investigating test-taking strategies during the completion of computer-delivered items from the Michigan English Test (MET): Evidence from eye tracking and cued retrospective reporting

## Author

**Ruslan Suvorov**
*Center for Language & Technology*
*University of Hawai'i at Mānoa*

## About the Author

**Ruslan Suvorov**
Ruslan Suvorov is a Language Technology Specialist at the Center for Language & Technology at the University of Hawai'i at Mānoa. His research interests lie at the intersection of language testing and assessment, computer-assisted language learning, and instructional technology and design. Ruslan has published in *Language Testing*, *CALICO Journal*, and *Canadian Journal of Applied Linguistics*, as well as in edited volumes, conference proceedings, encyclopedias, and research reports. He is a co-author of *Blended language program evaluation* (Palgrave Macmillan, 2016).

## Table of Contents

## Abstract

The overall goal of this study was to investigate test-taking strategies used by L2 learners while taking a computer-delivered test that was created using five types of items adapted from the paper-based Michigan English Test (MET). Using the convergence model of the data triangulation design (Creswell & Plano Clark, 2007), this study entailed gathering and analyzing eye-tracking data, verbal report data, and test performance data from 15 non-native speakers of English. The results of scanpath and verbal data analysis revealed a large variety of test-management and test-wiseness strategies used by the study participants when responding to 58 MET items. It was also found that while most test-taking strategies had been used across all item types, some of the strategies appeared to be applicable only to a specific item type. Furthermore, the results of the Wilcoxon Signed-Rank Test provided evidence that the participants' use of test-wiseness strategies introduced construct-irrelevant variance and had a statistically significant effect on the observed test scores. This study demonstrated the methodological value of utilizing eye tracking and verbal report methods for validation research in L2 assessment.

## Background

Test-taking strategies play a prominent role in the assessment of second language (L2) skills (Cohen, 2011). The past few decades have witnessed a surge of interest in research on test-taking strategies in L2 assessment, as evidenced by a large number of studies on this topic (e.g., Anderson, Bachman, Perkins, & Cohen, 1991; Cohen, 1998; Cohen & Upton, 2007; Kashkouli & Barati, 2013; Nevo, 1989; Phakiti, 2003; Sasaki, 2000; Storey, 1997; Yamashita, 2003). Understanding what strategies L2 test-takers use during language tests is particularly critical for validation research (Bachman, 1990; Brunfaut & McCray, 2015; Cohen, 2007b; Schmitt, Ng, & Garras, 2011; Weir, 2005; Wu & Stone, 2016) that traditionally has been product-oriented (i.e., focusing on assessment outcomes) and restricted to the use of statistical methods (O'Sullivan & Weir, 2011). Researchers have long recognized the importance of exploring test-taking strategies and argued that doing so can not only provide insights into the constructs measured by language tests (Anderson et al., 1991; Brunfaut & McCray, 2015; Storey, 1997) but is also necessary for determining whether the performance elicited through assessment tasks adequately represents L2 learners' proficiency in the target language rather than "behaviors employed for the sake of getting through the test" (Cohen, 2006, p. 325) that may introduce construct-irrelevant variance. Despite this recognition, the use of process-oriented approaches to validation that focus on the processes and strategies underlying L2 learners' responses to test items has been limited (Cohen, 2014; Wu & Stone, 2016). Thus, research is needed to investigate and better understand such response processes that represent one of the five sources of validity evidence outlined in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014).

Test-taking strategies can be defined as "those test-taking processes that the respondents have selected and of which they are conscious, at least to some degree" (Cohen, 1998, p. 92). In the context of L2 assessment, test-taking strategies are regarded as part of strategic competence that can produce construct-relevant variance or construct-irrelevant variance to test results (Bachman, 1990; Cohen, 2014). According to Nikolov (2006), test-taking strategies can be viewed from two perspectives. From one perspective, test-taking strategies are concerned with *what* L2 learners do when responding to test items; from another perspective, test-taking strategies relate to *why* L2 learners do what they do during the test. Cohen (2014) differentiates between two main types of test-taking strategies: (a) *test-management strategies* that are based on cognitive and linguistic processes relevant to the construct that test-takers consciously employ to answer test items meaningfully and responsibly, and (b) *test-wiseness strategies* that involve "using knowledge of testing formats and other peripheral information to obtain responses—very possibly the correct ones—on language tests without engaging the requisite L2 knowledge and performance ability" (p. 896). Consequently, test-management strategies contribute to construct-relevant variance, whereas test-wiseness strategies introduce variables that have an adverse effect on the validity of assessment outcomes.

Test developers must be fully aware of threats that test-takers' use of strategies may pose to the validity of test scores. Without this knowledge, test-takers may pass the test and answer many items correctly, not because of their ability to function in the target language but because of their effective use of test-wiseness strategies. Thus, the main threat posed by test-takers' successful use of test-wiseness strategies is inflated test scores that can be misinterpreted as evidence of their actual L2 proficiency and subsequently used by stakeholders to make misguided decisions about test-takers' placement, enrolment, hiring, or other purposes for which the test was designed. To identify such threats, Cohen (2014) suggests that test developers empirically examine how test-takers respond to a sample of test items and what strategies they employ in the process. Understanding these strategies can help test developers ensure that their test items are constructed in such a way that they "assess the respondents' requisite language skills, rather than their cleverness at circumventing an assessment of these skills" (Cohen, 2014, p. 894). While research on L2 test-taking strategies is itself relatively limited, studies looking specifically into test-wiseness strategies appear to be extremely scarce.

Existing investigations of test-taking strategies have addressed various item formats and language skills. Researchers have explored, for instance, how test-takers interact with cloze items (Sasaki, 2000; Storey, 1997; Yamashita, 2003), multiple-choice items (Cohen & Upton, 2007; Nevo, 1989; Phakiti, 2003; Yi'an, 1998), word associates format (Schmitt et al., 2011), and selected-response items such as multiple selection and drag-and-drop (Cohen & Upton, 2007). Although the bulk of research on test-taking strategies has been done in the context of L2 reading tests (Cohen & Upton, 2007; Nevo, 1989; Phakiti, 2003; Sasaki, 2000; Storey, 1997; Yamashita, 2003), some studies have also targeted L2 speaking (Barkaoui, Brooks, Swain, & Lapkin, 2013), writing (Plakans, 2009), and listening (Yi'an, 1998).

To explore test-taking strategies and processes used by L2 learners in the contexts of different item formats and language skills, researchers have traditionally employed concurrent or retrospective verbal reports, including think-aloud protocols (Anderson et al., 1991; Cohen & Upton, 2007; Plakans, 2009), retrospective protocols (Sasaki, 2000), questionnaires (Kashkouli & Barati, 2013), introspective interviews (Schmitt et al., 2011), and retrospective interviews (Phakiti, 2003; Plakans, 2009). While verbal reports can provide valuable insights into how L2 learners respond to test items and what test-taking strategies they use in the process (e.g., Anderson et al., 1991; Cohen, 2006, 2014; Cohen & Upton, 2007; Nikolov, 2006; Winke & Lim, 2014), this data collection method has two important limitations. In particular, verbal reports are prone to the threat of (a) *reactivity*, wherein respondents' verbalizations during a task may become an additional task that changes their thought processes; and (b) *veridicality*, wherein respondents' post-task verbalizations may be affected by forgetfulness, leading to a false or incomplete representation of their thought processes that have occurred during a task (Bowles, 2010). Moreover, the data furnished by verbal reports are subjective as they contain information about what test-takers *think* they do (and why they think they do it) rather than what they *actually* do in a particular L2 testing situation. Hence, these self-reported data must be supplemented with behavioral data collected via more objective measures, such as eye tracking, which can provide information about test-takers' actual engagement with L2 tasks.

Although eye-tracking technology has existed for at least half a century, it has started to make inroads into the field of L2 assessment only recently. In the context of computer-based L2 testing, eye tracking has been used to study test-takers' cognitive processes during L2 reading tests (Bax, 2013; Brunfaut & McCray, 2015) and test-takers' use of visual information during a video-based L2 listening test (Suvorov, 2013, 2015). According to Godfroid and Schmidtke (2013), eye-tracking data can reveal only the visual aspects of the stimulus that participants looked at, but not what they were thinking about while looking at those visual aspects (also see Suvorov, 2015). To obviate this limitation, some researchers have attempted to triangulate the data elicited via eye-tracking methods and retrospective verbal reports in their research designs (e.g., Bax & Chan, 2016; Brunfaut & McCray, 2015; Godfroid & Schmidtke, 2013). When deployed in combination with verbal reports, eye tracking has been found to be capable of providing cogent evidence of the nature of processes and strategies employed by test-takers during the completion of computer-based L2 assessment tasks (Brunfaut & McCray, 2015; Suvorov, 2013), thereby demonstrating a strong potential of this method for validation research.

To capitalize on the promising initial outcomes of research that has successfully combined eye tracking and verbal report methods in the context of L2 assessment, the present study aims to leverage the potential of this emergent methodology to investigate strategies used by test-takers during the completion of computer-delivered items adopted from the paper-based MET. Specifically, this study intends to answer three main research questions:

1. What test-taking strategies do test-takers employ when completing computer-delivered items adapted from the MET?
2. What differences in test-taking strategies do test-takers demonstrate when completing different types of computer-delivered items adapted from the MET (i.e., discrete dialogue items, dialogic listening sets, monologic listening sets, discrete grammar items, and reading sets)?
3. To what extent do test-wiseness strategies introduce construct-irrelevant variance and affect scores for computer-delivered items adapted from the MET?

## Methodology

This exploratory study used the convergence model of the data triangulation design (Creswell & Plano Clark, 2007) to collect and analyze three types of data: eye-tracking data, verbal report data, and test performance data. Eye-tracking data comprised the recordings of participants' eye movements during their interaction with each test item. Verbal report data contained participants' verbalizations of their test-taking strategies elicited via cued retrospective reporting. Finally, test performance data consisted of scores for multiple-choice items taken from the Michigan English Test (MET). The use of the convergence model entailed separate collection and analysis of eye-tracking data and verbal report data that were subsequently converged at the interpretation stage.

## Participants

Participants in this study were 15 non-native speakers of English (seven male and eight female) who were students enrolled in advanced-level English-as-a-second-language (ESL) courses at the University of Hawaiʻi at Mānoa. The main reason for including only advanced-level participants in this study was to ensure that they were proficient enough to verbalize their thoughts and discuss their use of test-taking strategies in English. Out of 15 participants, seven were undergraduate students, seven were graduate students, and one participant who marked his/her student status as "other." Their age ranged from 18 to 32 years (*M* = 23, *SD* = 4.45). The study participants were native speakers of the following languages: Mandarin Chinese (n = 7), Japanese (n = 2), Korean (n = 2), Kiribati (n = 1), Spanish (n = 1), Thai (n = 1), and Ukrainian (n = 1). They had been learning English for an average of 11 years (*SD* = 4.58).

## Materials and Instruments

### Michigan English Test items

This study used 58 items from the Michigan English Test (MET), which is a standardized, multilevel English-as-a-foreign-language test designed "to measure general English language proficiency in social, educational, and workplace contexts" (Cambridge Michigan Language Assessments, 2016). There were five types of items used in the study: discrete dialogue (DD) items (k = 10), dialogic listening (DL) items (k = 6), monologic listening (ML) items (k = 8), discrete grammar (DG) items (k = 10), and reading set (RS) items (k = 24). According to Cambridge Michigan Language Assessments (2016), the construct underlying the three types of listening items (i.e., DD, DL, and ML) comprises three groups of listening abilities or subskills: global, local, and inferential. Global subskills include the ability to understand the main idea, identify the speaker's purpose, and synthesize ideas from different parts of the stimulus. Local subskills entail the ability to identify supporting details, understand vocabulary, synthesize details, and recognize restatement. Inferential subskills comprise the ability to understand rhetorical functions, make an inference, infer supporting details, and understand pragmatic implications. The construct measured by discrete grammar items consists of a variety of grammar skills that are expected to be within the range and control of learners at the upper beginner to lower advanced levels of the Common European Framework of Reference for Languages (CEFR). The construct underlying the reading set of MET items includes global, local, and inferential subskills similar to the subskills measured by the listening items.

Because the MET is a paper-based test, the items were adapted for computer delivery via the Quiz module in Moodle, a course management system. Although the underlying assumption was that test-takers would engage with computer-delivered MET items in a similar way to paper-based MET items, there were some differences in test administration procedures between the operational paper-based MET and the computer-delivered version of MET adapted for this study (see Table 1). These differences should be taken into consideration when making any generalization from this study to the operational MET.

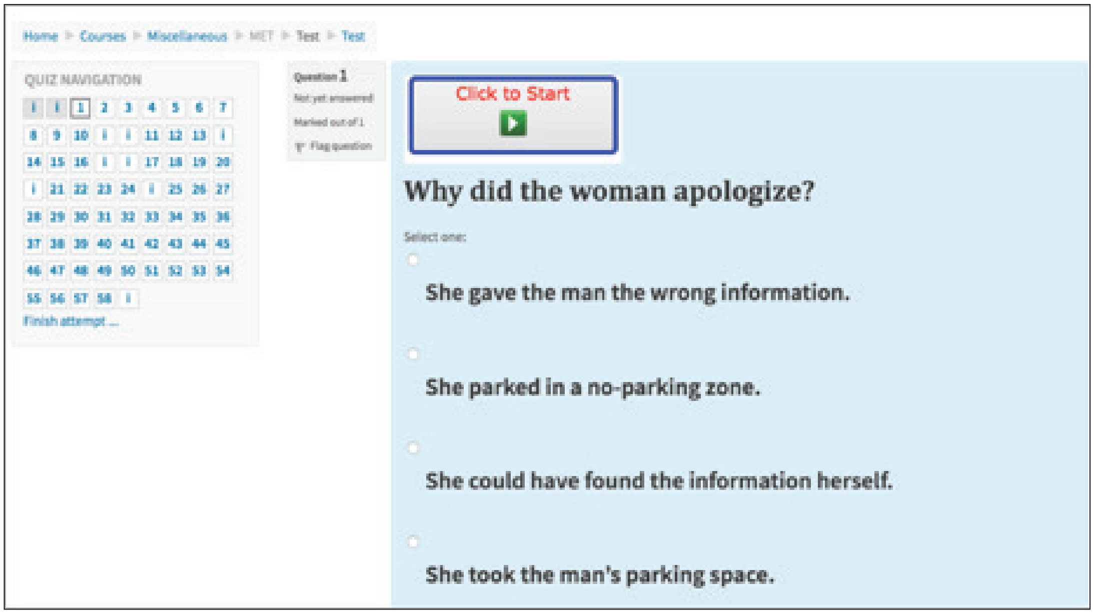
Instructions on how to complete the test were presented at the beginning of the test in audio and written formats. All items were multiple-choice items that comprised a question and four options. Responses to each item were automatically scored by Moodle as correct or incorrect. Each correct response was assigned a value of 1, and each incorrect response was assigned a value of 0. Participants were given up to an hour to answer all 58 items.

The 58 items adapted for this study were set up for computer delivery in a linear way. Test-takers were shown one item at a time. Each question and option in an item was displayed using 25-point font size. A large font size made it easier to identify which text elements the participants looked at even when there was some noise in the eye-tracking data. After answering one item, the participants could move to the next item by clicking on the "Next" button and were not allowed to go back and make changes to the items that they had already answered. The decision to present the items consecutively and unidirectionally was made because item revisits would have had a deleterious impact on the analysis and interpretation of participants' eye movements associated with each individual item.

For all the listening items (i.e., DD, DL, and ML), the participants could play an audio prompt by clicking on the Play button displayed next to the question. Because participants had to manually activate the Play button, they had an option to decide whether or not to preview a question before playing an audio prompt. Audio prompts could be played only once. All the reading items (RS) contained a text accompanied by 2–6 items. A concerted effort was made to fit each text and an associated item on the screen to preclude test-takers from scrolling the webpage while reading the text. Each reading item and the corresponding passage were displayed on a new, separate webpage. When an item tested comprehension across multiple passages, all passages related to that item and the item itself were shown together on a single, separate webpage as well. Figure 1 shows a sample listening item used in the study.

Table 1: Differences in Test Administration Procedures Between Operational MET and Adapted MET

| Test Administration Procedures | Operational MET | Adapted MET |
|---|---|---|
| Test sections | Listening, Reading, and Grammar (optional: Speaking and Writing) | Listening, Reading, and Grammar |
| Total number of multiple-choice items | 135 | 58 |
| Number of Listening items | 60 | 24 |
| Number of Grammar items | 25 | 10 |
| Number of Reading items | 50 | 24 |
| Time allotted for the test | 2 hours and 15 minutes | 1 hour |
| Delivery format | Paper-and-pencil | Computer-delivered |
| Item previewing while listening | Possible | Possible for DD items only |
| Aural input in Listening items | Played once, controlled by the test administrator | Played once, controlled by the test taker |
| Item presentation | Multiple items visible at once in the test booklet | One item visible per webpage |
| Test navigation (i.e., skipping forward or returning to previously answered items) | Possible | Not possible |



Why did the woman apologize?

Select one:

○ She gave the man the wrong information.

○ She parked in a no-parking zone.

○ She could have found the information herself.

○ She took the man's parking space.

*Figure 1. A sample MET item delivered via Moodle.*

### Cued retrospective reporting

Cued retrospective reporting was used to collect verbal data from participants. Also known as eye-movement supported verbal retrospection (Hansen, 1991) or post-experience eye-tracked protocol (Ball, Eger, Stevens, & Dodd, 2006; Petrie & Harrison, 2009), cued retrospective reporting is a method for eliciting participants' verbalizations using their eye-movement recordings as a stimulus (Van Gog, Paas, Van Merriënboer, & Witte, 2005). In the present study, this method entailed showing participants recordings of their eye movements after each item type was completed and asking them to verbalize what test-taking strategies they used to respond to each individual item. To facilitate their verbalizations, the participants were asked the following guiding questions for each item:

1. What strategies did you use to answer this question?
2. Why did you choose this option?
3. Are you sure this is the correct answer or not? Please explain.

### Eye-tracking equipment and software

A remote Gazepoint GP3 Eye Tracker (60 Hz, 0.5–1 degree of visual angle accuracy, 50–80 cm operating distance) was used to gather eye-tracking data. The eye tracker was connected to a 24-inch display with a screen resolution of 1920x1080 pixels and a desktop computer (Intel Core i7, 3.60 GHz) using Windows 7 Professional 64-bit operating system. The second display was connected to the same computer and used by the researcher to monitor data collection. Gazepoint Control software was used to perform a 9-point eye calibration, whereas Gazepoint Analysis Professional Edition (version 3.1.0) was used to record and analyze eye-tracking data.

### Background questionnaire

An online questionnaire was created to collect background information about participants. This anonymous questionnaire comprised six questions that asked participants to provide information about their gender, age, home country, native language, current student status, and the number of years they had been studying English. The questionnaire was created and administered via Moodle 2.8.

## Procedure

Following the approval of the study by the Institutional Review Board, participants were recruited among students enrolled in the English Language Institute (ELI) at the University of Hawai'i at Mānoa. To recruit participants, I contacted instructors teaching advanced-level ESL courses in the ELI, asking them for permission to visit their classes (a total of seven classes). During each 10-minute visit, I introduced the study to students, answered their questions about the study, and invited those students who expressed interest to become participants to sign informed consent forms and choose a date and time when they would like to participate in the study. Each participant was offered a $30 Amazon gift card as a compensation for his/her time participating in the study.

Data were collected from one participant at a time, with each individual data collection session lasting for approximately two hours. All 2-hour data collection sessions took place in the Learner & User Xperience (LUX) Lab at the Center for Language & Technology at the University of Hawaiʻi at Mānoa. The LUX Lab contained a researcher station and a participant station that was equipped with the remote eye-tracking system described above.

Upon their arrival to the LUX lab, participants received instructions for the study and were given a headset and several sheets of paper for note-taking. Participants were seated in front of a 24-inch display with a remote eye tracker positioned underneath it. While their head movements were unrestricted, participants were instructed to sit straight, maintain the distance of approximately 60–70 cm from the display, and avoid making any substantial head or body movements during the test. Participants were first asked to log into Moodle and complete a background questionnaire. Next, Gazepoint Control software was used to perform a 9-point eye calibration for each participant in order to prepare the eye tracker for data collection. Upon completing the eye-calibration procedure, the participants were invited to start the test, during which their eye movements were recorded by the eye tracker using Gazepoint Analysis Professional Edition software.

After each of the five item types (i.e., DD, DL, ML, DG, and RS items), I paused the test to conduct cued retrospective reporting that entailed showing the participants their eye-movement recordings during that specific item type and asking them to describe test-taking strategies that they had used to answer each item. In line with Brunfaut and McCray (2015), who did stimulated recalls after each task type (seven items per one task type), in this study cued retrospective reports were gathered after each of the five item types rather than after each item or after the whole test for two main reasons. Firstly, because each set of items of a specific type was relatively short (except for the Reading set items), the participants could easily remember and explain what test-taking strategies they had just used. Secondly, pausing the test and asking participants to reflect on their test-taking strategies after each item type did not appear to distract participants from focusing on the test, interfere with their use of test-taking strategies, or have an adverse effect on their test performance (which would have been the case if the test had been paused after each individual item). Participants' verbalizations during cued retrospective reporting were audio-recorded using Audacity (version 2.1.2), which is software for audio-recording and editing, and Camtasia Studio (version 8.6.0), which is software for capturing screen audio and video.

This procedure had been piloted with two participants before the actual data collection began.

## Analysis

Generally speaking, researchers who gather and analyze eye-tracking data face three main challenges. First, eye tracking tends to generate large amounts of data with spatial and temporal characteristics that are complex for empirical analysis (Coco, 2009). For instance, an eye-tracker with the sampling rate of 60 Hz records a variety of eye-tracking metrics every 16 ms, which makes the identification and selection of the most relevant data for the study both challenging and imperative. Second, the existence of over 150 different eye-tracking measures (Holmqvist et al., 2011) makes it difficult for researchers to decide which measure or measures to use. More importantly, there are no generally recognized or established standard procedures for gathering, processing, and analyzing eye-tracking data (e.g., Brunfaut & McCray, 2015; Vansteenkiste, Cardon, Philippaerts, & Lenoir, 2015), which complicates any comparisons among eye-tracking studies or judgments regarding the quality of epistemological practices and their implications for future research. As pointed out by Winke and Lim (2014), because eye-tracking studies in the field of L2 assessment are only emerging, researchers have to "invent the wheel" (p. 20) when determining how to analyze eye-tracking data empirically.

Given these challenges and the nature of the research questions in this study, analyzing eye-movement data quantitatively by calculating eye-tracking metrics such as fixation duration, fixation count, or gaze duration for specific areas of interest (AOIs) within each item appeared to provide little useful information about participants' use of test-taking strategies. A more informative approach to analyzing and revealing test-taking strategies necessitated a visual investigation of participants' gaze patterns associated with each test item. One study that utilized such visual investigation was that of Ehmke and Wilson (2007), who explored a correlation between eye-movement patterns and usability problems by doing manual analysis of scanpaths and complementing it with the analysis of data from retrospective interviews. Similarly, scanpath analysis has also been suggested by Winke and Lim (2014) as a future direction for eye-tracking research in L2 assessment.

Following Ehmke and Wilson (2007) and Winke and Lim (2014), and given the qualitative nature of the research questions in this study, the complexity of the data (i.e., the length of the test, the number of test items, and the number of participants), and the challenge of identifying quantitative eye-tracking measures that would be both meaningful for answering the research questions and comparable across all participants and stimuli, I employed qualitative scanpath analysis of eye-tracking data via visual inspection. A scanpath provides a spatial and temporal sequence of a participant's eye-movements during the completion of a specific visual task (Coco, 2009). In line with Tzanidou, Minocha, and Petre (2005), I used visual inspection to analyze the scanpath data for each test item completed by each participant (i.e., a total of 58 items x 15 participants = 870 scanpath data sets). In particular, for each scanpath data set, I wrote a brief description of a scanpath, noting the element(s) of an item that each participant had looked at first, the direction of his/her eye gaze and the sequence of fixations, and the temporal characteristics of the participant's oculomotor behavior (e.g., whether the participant had spent a significant amount of time on a specific element of an item).

To analyze verbal data elicited via cued retrospective reporting, I first transcribed participants' verbalizations describing their use of test-taking strategies. Next, the verbal data were analyzed and coded for the types of test-management and test-wiseness strategies (as defined by Cohen, 2014) used by the participants to respond to each test item. In particular, I marked a strategy as a test-wiseness strategy if the following two conditions were met: (a) a participant provided no indication of knowing, even partially, the answer to the question and could not explain why he/she had chosen a specific option and why it was supposed to be the correct answer and (b) a participant was not sure whether the selected option was the correct answer. A strategy was also coded as a test-wiseness strategy if it clearly did not match the cognitive processing intended to be activated by a specific item. The results of verbal data analysis were subsequently converged with the results of scanpath analysis to provide answers to Research Questions 1 and 2. In cases when the verbal data contradicted the scanpath data, the latter were used to determine the strategy, as illustrated in Table 2.

To answer Research Question 3, all three types of data were analyzed quantitatively. I first calculated descriptive statistics for all test scores (i.e., observed scores). Using the results of scanpath analysis and verbal data analysis, I tallied the number of test items that had been answered using test-wiseness strategies and divided them into the items that had been answered correctly and the items that had been answered incorrectly. To measure the extent to which the use of test-wiseness strategies helped the participants in this study respond to the MET items correctly, I calculated an "adjusted" score (Sadj) for each participant by subtracting the score for the items that had been answered correctly using test-wiseness strategies (Stw) from the observed score (Sobs): Sadj = Sobs – Stw. Finally, I ran a Wilcoxon Signed-Rank Test to determine whether there was a statistically significant difference between the observed scores and the adjusted scores, which was subsequently used to draw conclusions about the effect of test-wiseness strategies on scores for MET items and the extent to which the use of test-wiseness strategies contributed to construct-irrelevant variance.

Table 2: Analysis of Verbal Data and Scanpath Data for Strategy Determination (Participant 12)

| Item | Verbal Data | Scanpath Data | Strategy |
|---|---|---|---|
| DD 10 | Claims to have listened to the audio prompt first and then read the question and the options. | Clicked the Play button, read the question and all four options while listening to the audio prompt. | Reading the question and the response options while listening to the audio prompt. |

## Results

### Research Question 1

In answering Research Question 1 (i.e., What test-taking strategies do test-takers employ when completing computer-delivered items adapted from the MET?), the results of scanpath and verbal data analysis revealed that the participants had used a wide variety of test-taking strategies, comprising both test-management and test-wiseness strategies, as defined by Cohen (2014).

#### Test-management strategies

The test-management strategies identified in the data were classified into three main groups.

#### Group A: Strategies related to the order of viewing or interacting with item elements.

The first group of test-management strategies consisted of strategies that test-takers utilized before selecting the answers. Specifically, these strategies focused on the order in which the test-takers viewed or interacted with different elements of each item (i.e., the question, the response options, and the audio/text prompt).

TMA1. Reading the question and the response options while listening to the audio prompt (i.e., multitasking).

The main purpose of multitasking was to save time. Some participants who resorted to this strategy preferred to use it with shorter items because they were able to focus on both reading and listening, but avoided it when responding to longer and more difficult items. As explained by Participant 5,

> *When the audio is really easy, I want to read it at the same time. But if it's a little hard, I want to be careful to listen it first.* (Participant 5, DD Item 1)

In the meantime, for a number of test-takers, multitasking was distracting and caused them to miss some key information while listening to the audio prompt:

> *It's hard for me. I mean while I am paying attention to listening to the audio, I can't afford focus on reading the question and options. [It's] distracting.* (Participant 11, DD Item 1)

TMA2. Previewing the question and options before listening to the audio or reading the text prompt.

This strategy was utilized by test-takers to anticipate the topic of the audio or text prompt and determine what information they would need to be looking for in the audio or text in order to answer the question:

> *I can read the question first and then I can read the choices, so I know what I am expected to hear in the conversation and what can be relevant to the choice.* (Participant 8, DD Item 2)

For some test-takers, this strategy appeared to be particularly useful for longer and more difficult audio or text prompts:

*If I think it's really hard, if I know it's long and it's gonna be really hard, then I would look at the answers first to have a general idea of what it's talking about.* (Participant 4, DD Item 3)

TMA3. Previewing the question, listening to the audio prompt or reading the text prompt, then reading the response options.

This strategy was used by some test-takers to determine what information they would need to be looking for in the audio or text prompt without getting distracted by the response options:

> *I think it'll be better if I listen it first and then do it because it might get confused if I see the answers first.* (Participant 4, DD Item 1)

TMA4. Reading the text prompt first, then reading the question and the response options.

The participants adopted this strategy in order to familiarize themselves with the text prompt first before moving to the questions. Meanwhile, some of them acknowledged that reading the whole text without knowing the number and type of questions related to the text could be less efficient than looking for specific information relevant to the questions.

#### Group B: Strategies used for interacting with the text prompt.

Group B comprises test-management strategies that the participants used to interact with the text prompt.

TMB1. Reading, rather than skimming through, the entire text prompt.

This strategy was commonly used by the participants as it enabled them to get a good understanding of the entire text and answer the questions without a need to reread the text. While reading the entire text was more time-consuming than skimming through it, the participants who had carefully read the entire text generally did not need to reread it when answering the follow-up questions related to the same text. Participant

7, for example, who had initially relied on skimming through the text prompts, decided that reading the entire text prompt once would make it easier for him to respond to the items:

*I am gonna read this and try to understand the text properly and then it should be easier. Because now I know that the next question will be about this text. And I found it difficult before when there was the same text and I had to read it again. Better read it at first, at once, and then we can answer the questions.* (Participant 7, RS Item 47)

TMB2. Reading only the first paragraph or the first sentence in each paragraph.

By utilizing this strategy the test-takers were able to get the main idea of the text without needing to read it entirely:

*I think generally the first two or the last show the importance of every paragraph. So I think it made me more quickly to get the main idea of the readings.* (Participant 14, RS Item 41)

On the other hand, this strategy caused some test-takers to misinterpret the text and/or miss key information that was critical for answering the questions.

TMB3. Skimming through, rather than reading, the text prompt.

The skimming strategy was deployed in order to find specific information relevant to answering the question. It helped some test-takers to be more time-efficient and avoid reading the entire text, which some of them considered to be "a waste of time" (Participant 11, RS Item 35). Meanwhile, in some cases skimming through the text caused several participants to misunderstand the text and/or miss information that was critical for answering the questions.

TMB4. Re-reading the text prompt multiple times.

Some test-takers adopted this strategy because they were struggling to find information that was relevant to answering the question. In some case, this was due to the participants' not fully understanding the text prompt. In other cases, however, the participants simply could not locate a specific word or information in the text that the question was asking about.

TMB5. Searching for a keyword (taken either from the question or from the response options) in the text prompt and reading only the part of the text containing that keyword.

This strategy was used primarily to answer questions that asked for specific information from the text prompt (e.g., What does the word "X" in the last sentence mean?).

### Group C: Strategies used for interacting with the question and/or the response options and selecting the answer.

The strategies in Group C illustrate how the test-takers engaged with response options and selected the answers.

TMC1. Reading all the options carefully.

The participants adopted this strategy to ensure that they would not miss the correct option:

*Usually I read all of them… in case maybe some of the others are better than this one [option].* (Participant 3, DD Item 5)

TMC2. Re-reading the question and/or response options several times before selecting the answer.

Using this strategy helped some test-takers better understand the question and the response options and select the best answer. Doing so appeared to be particularly useful with longer options that put more cognitive load on the test-takers and were more difficult to process. Participant 15, for instance, tried to reread the options twice: one time to familiarize himself with all options and the second time to choose the best one. In some cases, however, multiple rereadings of response options led to overthinking and could be an indicator that the test-takers did not know or were not confident in their answer.

TMC3. Skimming through the question and/or the response options without reading them carefully.

Skimming helped the test-takers save time on the test. Participant 6, for instance, skimmed through the options by looking only at the parts that were different. For example, if every option started with "It will," he would skip it and read only the last part of each option. Similarly, Participant 9 deployed this strategy when he was confident in his answer and wanted to be more efficient:

*Because I kind of know that option B is the answer and I was hearing option B on the audio. So I was sure that it's option B and I clicked it first. And then I kind of skimmed through it to double-check other options.* (Participant 9, DL Item 12)

Meanwhile, the use of this strategy caused some participants to miss critical information in the question or the response options and choose the wrong answer:

> *Because I did not carefully read the question, so I assumed that chess has been a popular game since many centuries ago. So, yeah, I misread the question. I think the correct answer would be [x].* (Participant 1, DG Item 30)

Similarly, Participant 12, who used the skimming strategy when answering RS Item 57, also misread the question, thinking that it was asking about "monopoly," whereas in reality the question was about "Minneapolis." However, unlike Participant 1 in the example above, Participant 12 selected an option that, by coincidence, was the correct one.

TMC4. Reading one response option at a time and going back to the question or the text prompt to check if the option is correct.

This strategy was adopted extensively for answering DG items that comprised a question with a blank and four options that the participants had to choose from in order to fill in the blank. Inserting one option at a time in the blank helped the test-takers decide whether the sentence was grammatically correct and made sense semantically. Additionally, some participants utilized this strategy for answering some RS items:

> *I think I looked back and forth… I looked at the answers and I looked at the passage just to double-check, to match the answers with the content of the passage.* (Participant 1, RS Item 36)

In some cases, however, reading one option at a time appeared to be time-consuming, inefficient, and even confusing:

> *I think it's a very bad habit to have because sometimes I confuse myself even more… Sometimes I jump back and forth because I want to double-check. But then sometimes I confuse myself even more.* (Participant 1, RS Item 40)

TMC5. Reading the response options only until the one that is perceived as correct, selecting that option, and skipping the remaining option(s).

Several participants employed this test-taking strategy to save time during the test. Typically, this strategy was used when the test-takers were absolutely sure that the selected response option was the correct one and, therefore, there was no need to read other options:

> *If I am very confident, I will just choose it. If I am not, I will prefer to read all of them and choose the best.* (Participant 5, DL Item 11)

Other participants, however, adopted this strategy not because of their confidence that the selected option was the correct answer but because of their intention to save time on the test. Participant 6, for example, used this strategy several times in order to speed through the test, even though he recognized that this was a bad test-taking strategy for him that led him to wrong answers:

> *Yeah, like I save time always. I always, even in exams, I always go fast and then I have the problem…* (Participant 6, DD Item 2)

TMC6. Selecting a response option while listening and skipping the rest of the audio prompt to move to the next question.

Similar to the previous two strategies, this test-management strategy was utilized to save time during the test. It was used specifically for answering DD items, each of which was associated with one short audio prompt. The participant deployed this strategy after hearing the part of the audio prompt that contained information necessary for answering the question and felt there was no need to continue listening to the audio.

TMC7. Reading the question and the response options, then choosing the answer without consulting the text prompt after its initial reading.

In the reading section, this test-management strategy was used when the text prompt was followed by several questions and the participants had already read the text to answer previous questions. The participants deployed this strategy when they wanted to be more time-efficient and appeared to know the answer. As described by Participant 2,

> *I read through the options first because I read the passage in previous question. So, I thought I could solve it by just reading the options.* (Participant 2, RS Item 36)

TMC8. Selecting a response option based on a keyword from the audio or text prompt.

This test-management strategy was employed primarily by those participants who did not fully understand the audio or text prompt and relied on micro-level information (i.e., information at the word level rather than sentence or text level). It was also used to answer questions that asked for specific information from the text prompt (e.g., What does the word "X" in the first sentence refer to?).

TMC9. Selecting a response option based on the inferences drawn from the audio or text prompt.

This strategy was used by the test-takers who were able to understand the audio or text prompt at the macro level (i.e., at the level of a paragraph or entire text). It was also deployed to respond to the questions that asked about the main idea of the text.

TMC10. Switching the answer after having selected a response option.

Some participants switched their answers because they had either found additional information in the prompt or changed their mind for reasons other than guessing:

> *Because when I read other options, I thought, 'Oh, this may be better or may be potential right answer.' So I decided to re-read and double-check.* (Participant 3, RS Item 49)

In some cases (especially when accompanied by a lengthy fixation on the response options or looking back and forth between the options), the test-takers resorted to this strategy because they were not confident in their answer and hesitated when deciding which option to choose.

TMC11. Eliminating other response options when selecting the answer.

As a test-management strategy, elimination was used by some test-takers not to find the correct answer, but to double-check and confirm that the selected answer was correct. For DG items, this strategy was also utilized to eliminate the options that were grammatically incorrect.

All test-management strategies identified in the data are summarized in Table 3.

## Test-wiseness strategies

There were several test-wiseness strategies identified in the data. This type of a test-taking strategy was used by all participants to respond to some items without knowing the answer or relying on the requisite L2 knowledge assessed by the item.

TW1. Selecting a response option by making a random guess.

Guessing, defined in this study as a random choice of an option, was the most commonly used test-wiseness strategy. The participants deployed this strategy when they did not know the answer to the question, did not care about the item, or were too tired to think and focus. Participant 6, for instance, guessed the answer to DG Item 30 because he felt tired and could not focus:

> *Probably "for" have more sense, but I choose "by" because I don't want to think more.* (Participant 6, DG Item 30)

TW2. Selecting a response option out of a vague sense that the other options are incorrect without understanding the option or the question.

This test-wiseness strategy was also quite common in this study. It was utilized when the participants did not know the answer and were trying to choose an option that sounded better than others without being able to explain why it was better or without understanding the option or the question. In the following example, the participant chose the "better option" without

Table 3: Test-management Strategies

| Strategy Code | Strategy Description |
|---|---|
| *Group A* | *Strategies related to the order of viewing or interacting with item elements.* |
| TMA1 | Reading the question and the response options while listening to the audio prompt (i.e., multitasking). |
| TMA2 | Previewing the question and options before listening to the audio or reading the text prompt. |
| TMA3 | Previewing the question, listening to the audio prompt or reading the text prompt, then reading the response options. |
| TMA4 | Reading the text prompt first, then reading the question and the response options. |
| *Group B* | *Strategies used for interacting with the text prompt.* |
| TMB1 | Reading, rather than skimming through, the entire text prompt. |
| TMB2 | Reading only the first paragraph or the first sentence in each paragraph. |
| TMB3 | Skimming through, rather than reading, the text prompt. |
| TMB4 | Re-reading the text prompt multiple times. |
| TMB5 | Searching for a keyword (taken either from the question or from the response options) in the text prompt and reading only the part of the text containing that keyword. |

Table 3: Test-management Strategies

| *Group C* | *Strategies used for interacting with the question and/or the response options and selecting the answer.* |
|---|---|
| TMC1 | Reading all the options carefully. |
| TMC2 | Re-reading the question and/or response options several times before selecting the answer. |
| TMC3 | Skimming through the question and/or the response options without reading them carefully. |
| TMC4 | Reading one response option at a time and going back to the question or the text prompt to check if the option is correct. |
| TMC5 | Reading the response options only until the one that is perceived as correct, selecting that option, and skipping the remaining option(s). |
| TMC6 | Selecting a response option while listening and skipping the rest of the audio prompt to move to the next question. |
| TMC7 | Reading the question and the response options, then choosing the answer without consulting the text prompt after its initial reading. |
| TMC8 | Selecting a response option based on a keyword from the audio or text prompt. |
| TMC9 | Selecting a response option based on the inferences drawn from the audio or text prompt. |
| TMC10 | Switching the answer after having selected a response option. |
| TMC11 | Eliminating other response options when selecting the answer. |

understanding the question that asked about the reason why prescription lenses were mentioned in the passage:

> *I just choose the better option from these, I think… I don't know what the 'prescription lenses' means, so I do not understand that [question]… I am sure because the other options are not good.* (Participant 3, RS Item 36)

TW3. Selecting a response option because it contains a word or phrase from the audio or text prompt that is not relevant to the information being tested by the question.

Some participants relied on this strategy to select an option because it contained a random word or phrase from the audio or text prompt. When answering RS Item 52, for instance, Participant 2 selected the response option "how the entrepreneurial growth happens in the area" simply because she had seen the word "area" in the text prompt without understanding or being able to explain how this word was relevant to the question.

TW4. Selecting a response option that looks different from others.

Choosing a response option that appears different from other options was another test-wiseness strategy adopted by some test-takers. The following example from the verbal data illustrates the use of this strategy:

> *This one I wasn't sure at all. I missed that part, so I just guessed 'seashells'… because these are just not interesting [options].* (Participant 7, ML Item 18)

TW5. Selecting a response option because of clues from other items.

Information from previously answered items was utilized by some test-takers to respond to the following item. When trying to answer ML Item 19, for example, Participant 5 selected the option "the natural world" because it was related to the option "seashells" from ML Item 18. Even though these items asked completely different questions, this participant was hopeful that she would have a higher chance of answering both questions correctly by choosing semantically related options because both items were associated with the same audio prompt.

TW6. Selecting a response option by using background knowledge.

In a couple of cases, the participants relied on their background knowledge to respond to the question. In the following example, Participant 2 explained how she used this strategy to answer a question that asked about the meaning of a specific word from the text prompt. Instead of using the context to infer the meaning of the word, she utilized her background knowledge to choose a response option:

> *Maybe this is not relevant, but I remember that when you ride a car there is a mirror on the side of you and it says 'Present objects may be not be close than what you see'… Yeah, I think I remember that, so I chose the 'present'.* (Participant 2, RS Item 42)

TW7. Selecting a response option because of the speaker's tone of voice (i.e., a cue in the audio prompt).

Among the MET items used in this study, there was one item (i.e., DD Item 10) that asked about the feelings of the speaker. With the speaker sounding clearly annoyed during the conversation in the audio prompt, some participants (e.g., Participant 14) used their interpretation of the speaker's tone of voice—rather than the content of the conversation—to select the response option "annoyed." The use of this test-wiseness strategy thus illustrated a potential issue with the design of this specific test item.

All test-wiseness strategies identified in the data are summarized in Table 4.

Table 4: Test-wiseness Strategies

| Strategy Code | Strategy Description |
|---|---|
| TW1 | Selecting a response option by making a random guess. |
| TW2 | Selecting a response option out of a vague sense that the other options are incorrect without understanding the option or the question. |
| TW3 | Selecting a response option because it contains a word or phrase from the audio or text prompt that is not relevant to the information being tested by the question. |
| TW4 | Selecting a response option that looks different from others. |
| TW5 | Selecting a response option because of clues from other items. |
| TW6 | Selecting a response option by using background knowledge. |
| TW7 | Selecting a response option because of the speaker's tone of voice (i.e., a cue in the audio prompt). |

## Research Question 2

To answer Research Question 2 (i.e., What differences in test-taking strategies do test-takers demonstrate when completing different types of computer-delivered items adapted from the MET?), the results of scanpath and verbal data analysis were used. These results revealed that while most test-taking strategies had been used across all item types, some of the strategies appeared to be applicable only to a specific item type. Table 5 shows the types of test-management strategies, whereas Table 6 illustrates the types of test-wiseness strategies for each item type.

Table 5: Test-management Strategies by Item Type

| Strategy | Item Type | | | | |
|---|---|---|---|---|---|
| | DD | DL | ML | DG | RS |
| TMA1 | x | | | | |
| TMA2 | x | | | | x |
| TMA3 | x | | | | x |
| TMA4 | | | | | x |
| TMB1 | | | | | x |
| TMB2 | | | | | x |
| TMB3 | | | | | x |
| TMB4 | | | | | x |
| TMB5 | | | | | x |
| TMC1 | x | x | x | x | x |
| TMC2 | x | x | x | x | x |
| TMC3 | x | x | x | x | x |
| TMC4 | | | | x | x |
| TMC5 | x | x | x | x | x |
| TMC6 | x | | | | |
| TMC7 | | | | | x |
| TMC8 | x | x | x | | x |
| TMC9 | x | x | x | | x |
| TMC10 | x | x | x | x | x |
| TMC11 | x | x | x | x | x |

Table 6: Test-wiseness Strategies by Item Type

| Strategy | Item Type | | | | |
|---|---|---|---|---|---|
| | DD | DL | ML | DG | RS |
| TW1 | x | x | x | x | x |
| TW2 | x | x | x | x | x |
| TW3 | x | x | x | | x |
| TW4 | x | x | x | x | x |
| TW5 | | | x | | |
| TW6 | | | | | x |
| TW7 | x | | | | |

As can be seen from the list of strategies listed in both tables, some strategies were skill-specific and applied only to a certain item type (such as the test-management strategies TMB1-TMB5 that the participants used to interact with text prompts when completing RS items). Furthermore, some strategies applied only to some listening item types (i.e., DD) but not to others (i.e., DL and ML) due to the peculiarities of the test design: Test-takers were not able to preview questions and options for DL and ML items because audio prompts and questions with options were hosted on separate pages. In other words, when completing DL and ML items, test-takers had to finish playing an audio prompt first before they could move on to the next page that contained a question with response options related to that specific audio prompt. In addition, due to a relatively small sample size, certain test-wiseness strategies such as TW6 (i.e., the use of background knowledge) were found for only one specific item type (i.e., ML items), although theoretically this strategy could have been used by participants to answer any item type.

## Research Question 3

To answer Research Question 3 (i.e., To what extent do test-wiseness strategies introduce construct-irrelevant variance and affect scores for computer-delivered items adapted from the MET?), I first calculated descriptive statistics for 58 MET items completed by 15 participants. The results revealed that the mean (*M*) was 48 and standard deviation (*SD*) was 5.49. The absolute values for skewness (-0.99) and kurtosis (1.91) were less than two, which, according to Bachman (2004), indicated a reasonably normal distribution.

Next, I counted the number of items that each participant answered, either correctly or incorrectly, by using test-wiseness strategies (see Table 7).

Table 7: Number of MET Items Answered Using Test-wiseness Strategies (n = 15, k = 58)

| Participant | Correctly Answered Items | | Incorrectly Answered Items | | Total | |
|---|---|---|---|---|---|---|
| | Number | % | Number | % | Number | % |
| Participant 1 | 4 | 6.9 | 3 | 5.17 | 7 | 12.07 |
| Participant 2 | 8 | 13.79 | 6 | 10.35 | 14 | 24.14 |
| Participant 3 | 3 | 5.17 | 2 | 3.45 | 5 | 8.62 |
| Participant 4 | 2 | 3.45 | 8 | 13.79 | 10 | 17.24 |
| Participant 5 | 7 | 12.07 | 3 | 5.17 | 10 | 17.24 |
| Participant 6 | 8 | 13.79 | 9 | 15.52 | 17 | 29.31 |
| Participant 7 | 5 | 8.62 | 8 | 13.79 | 13 | 22.41 |
| Participant 8 | 2 | 3.45 | 3 | 5.17 | 5 | 8.62 |
| Participant 9 | 2 | 3.45 | 7 | 12.07 | 9 | 15.52 |
| Participant 10 | 3 | 5.17 | 6 | 10.35 | 9 | 15.52 |
| Participant 11 | 4 | 6.9 | 2 | 3.45 | 6 | 10.35 |
| Participant 12 | 5 | 8.62 | 8 | 13.79 | 13 | 22.41 |
| Participant 13 | 10 | 17.24 | 17 | 29.31 | 27 | 46.55 |
| Participant 14 | 1 | 1.72 | 3 | 5.17 | 4 | 6.9 |
| Participant 15 | 3 | 5.17 | 1 | 1.72 | 4 | 6.9 |

As shown in Table 7, all participants used test-wiseness strategies to varying degrees: While Participants 14 and 15 used test-wiseness strategies to answer four items, Participant 13 relied on these strategies significantly more and used them to respond to 27 items. The results in the table also indicate that the use of test-wiseness strategies helped all participants answer correctly anywhere from one item (Participant 14) to ten items (Participant 13) without deploying the requisite L2 knowledge.

Observed scores and adjusted scores for each participant are reported in Table 8. As mentioned in the methodology section, adjusted scores were calculated by subtracting the number of items answered correctly with the help of test-wiseness strategies from the observed scores for each participant.

Table 8: Participants' Observed and Adjusted Test Scores for MET Items (n = 15, k = 58)

| Participant | Observed Score | Adjusted Score |
|---|---|---|
| Participant 1 | 51 | 47 |
| Participant 2 | 45 | 37 |
| Participant 3 | 52 | 49 |
| Participant 4 | 47 | 45 |
| Participant 5 | 54 | 47 |
| Participant 6 | 45 | 37 |
| Participant 7 | 47 | 42 |
| Participant 8 | 50 | 48 |
| Participant 9 | 49 | 47 |
| Participant 10 | 46 | 43 |
| Participant 11 | 54 | 50 |
| Participant 12 | 42 | 37 |
| Participant 13 | 34 | 24 |
| Participant 14 | 48 | 47 |
| Participant 15 | 56 | 53 |

The graph in Figure 2 provides a visual summary of the observed and adjusted scores for MET items for each participant.

The results of the Wilcoxon Signed-Rank Test indicated that observed scores were statistically significantly higher than adjusted scores, $Z = -3.41$, $p < .01$. These findings suggest that the participants' use of test-wiseness strategies introduced construct-irrelevant variance that had a statistically significant effect on the observed test scores for the 58 items adapted from the MET for this study.

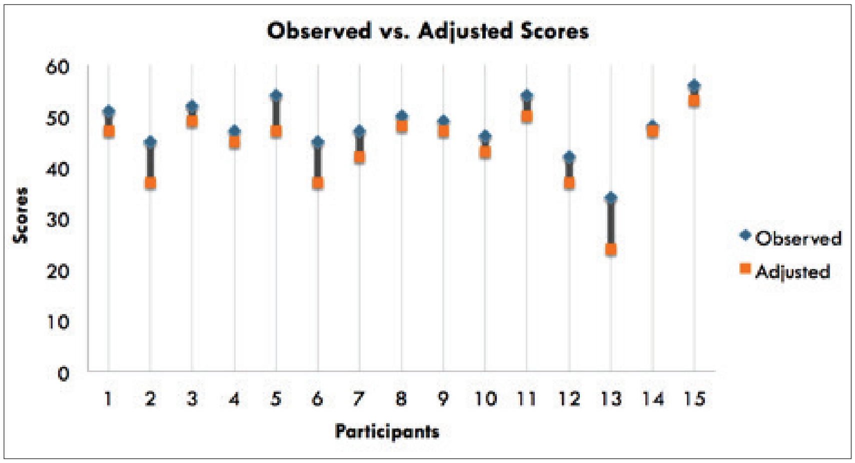


*Figure 2. Observed vs. adjusted scores for 15 participants.*

## Discussion

This study has demonstrated that L2 test-takers tend to use a large variety of test-taking strategies, including both test-management and test-wiseness strategies. The findings have also revealed individual differences among participants regarding their use of test-taking strategies. For example, while Participant 6 tended to skim through the questions and response options and tried to be time efficient, Participant 7 read all options very carefully and appeared to be more concerned with accuracy than efficiency. Furthermore, while some participants were consistent in their use of strategies, others changed them throughout the test. For example, Participant 15 read the question, the text, and the response options to answer the first few items in the reading section, but later switched to previewing the question and the response options before reading the text:

> *First, it was really strange how in general to approach this type of test. But then I realized that I read the question, then I read options, and I eliminate that just don't make much sense, if there are some. And then I focus on the few options that make sense and try to find proof in the text.* (Participant 15, RS Item 38)

The results of this study are generally in line with Nikolov (2006), who found two trends in the data: (a) some participants relied on the same strategies for each item throughout the whole test, and (b) participants used a combination of strategies rather than one particular strategy when answering the test items. The finding that L2 learners tend to use clusters of test-taking strategies rather than apply one strategy in isolation is further corroborated by Cohen (2007a, pp. 35–36).

In this study, the test-takers' choice and use of strategies appeared to be directly affected by several test design characteristics. One such characteristic was the length of response options and the complexity of prompts. In particular, because longer options required more time to read and process, the participants avoided strategies that were more cognitively demanding (e.g., reading the question and response options while listening to the audio prompt) and instead opted for the strategies that enabled them to process information sequentially rather than simultaneously (e.g., playing the audio first, then reading the question and the response options).

Another test design characteristic was time constraints. For example, Participant 7 considered multitasking to be a bad strategy, but he nevertheless used it for answering some discrete dialogue items because of his concern that he might run out of time and be unable to finish the test:

> *If I know I have time, I would read this [options] carefully and then I would listen. I would prefer that.* (Participant 7, DD Item 10)

The fact that the test-takers could see only one RS item at a time on the computer screen had an impact on the participants' choice of test-taking strategies for interacting with text prompts in the Reading section. Without knowing how many items were associated with each specific reading prompt, Participant 15, for example, struggled to decide whether it was worth reading the whole text or skimming through it:

> *The thing is because you never know whether the next question will be related to this text, you don't want to spend much time on this text. So I think it actually would be helpful to say, 'OK, the next five questions will be about this text.' So, you will be like, 'OK, I will have five questions about this text, so I can spend some time on reading it.' So I was thinking that next question may be another text, just two questions, for example. So you will like, 'why I spend time reading it? I just better find the answers'.* (Participant 15, RS Item 40)

The participants' use of test-taking strategies during the reading items also appeared to be affected by the design of the text prompts. Some reading items asked about the meaning of specific words in the text and referred to those words by referencing the paragraph and sentence in which the words occurred (e.g., In the fourth sentence of paragraph 3, what does the word "X" refer to?). However, because the words were not visually marked in the text, some participants had a hard time finding them there. In a few cases, participants simply ended up guessing the meaning of the words without being able to locate them in the text. Had the words been highlighted or marked in some way, the participants would have used a different set of strategies when responding to those items.

Furthermore, the findings revealed that the choice of test-taking strategies was influenced by the type of questions. For instance, to answer questions about the main idea of the passage, some participants utilized the strategy of scanning the text, but to answer more specific questions, they read the whole text or a specific portion of the text more carefully:

> *For this one, it asks for the main idea, right? So, whenever they ask for the main idea, I scan the text and look at the introduction or the heading. But if they ask for details like words, I need to read it carefully.* (Participant 8, RS Item 49)

The results also suggest that the choice of test-taking strategies might depend on the stakes of the language test. In the case of a high-stakes exam like TOEFL, some participants may be more likely to use strategies that emphasize and promote accuracy over efficiency (such as read all the options carefully rather than skim through them, even if they are confident in their answer). As Participant 11 explained,

> *This is just experiment, right? But when I have like TOEFL examination, I always keep in my mind to read all choices because—I am sorry—that's more important for me. But now, I mean…* (Participant 11, DD Item 4)
>
> *If I take TOEFL examination, I think I read everything—questions, options—first and then press the button [to listen]. Yeah, before listening I would read everything. But… should I do that [in this test]?* (Participant 11, DD Item 5)

With respect to eye-tracking data, its inclusion and analysis in this study appeared to provide more accurate, detailed, and nuanced information (including spatial and temporal characteristics) regarding the participants' use of strategies related to their interaction with the elements of each item (i.e., the audio/text prompt, the question, and the response options). It also revealed information about the extent of participants' confidence in answering individual items. For instance, if the eye-tracking data showed that the participant had spent a long period of time rereading the response options, it was a relatively clear indication that the participant had most likely struggled when answering that test item.

When converging the results from scanpath analysis and verbal data analysis, I came across a number of cases in which there was a mismatch between what the participants claimed they had done and what the eye-tracking data showed they had done. One of the most common mismatches was related to the participants' interactions with item elements. For instance, when describing his test-taking strategies for answering RS Item 56, Participant 3 reported that he had first read the question and response options, then scanned the text to identify information necessary for answering the question, and, finally, selected the answer. However, the recording of this participant's eye movements revealed that he first scanned the text and then proceeded to read the response options and select the answer. Similarly, Participant 7 claimed that he had first listened to the audio prompt and then read the question and response options when responding to DD Item 2; however, according to the eye-tracking data, he was reading the question and options while listening to the audio. There were also a few instances when participants reported choosing a specific option and provided the reasons for their choice, but the eye-tracking data revealed that, in fact, a different response option had been selected (e.g., Participant 6, RS Item 41). When such mismatches were detected, I chose to rely on the eye-tracking data because it offered more direct and compelling evidence of what had actually happened.

On the other hand, eye-tracking data were misleading at times and could be interpreted only with the help of verbal data. For instance, when answering DG Item 30, Participant 4 spent a lot of time looking at a specific word at the beginning of the question, which would normally be an indication that the participant is cognitively focused on that word. However, when asked why she had focused on that word, the participant responded that she had been "just staring at the word" and acknowledged that she tends to do that sometimes. Similarly, Participant 11 claimed that while listening to the audio prompt in DD Item 3, he had not focused mentally on the question, even though the eye-movement recording suggested that his eye gaze had been directed towards the question during the audio.

## Conclusion

The findings of this study have several important implications for future work. First, test developers should keep in mind that the test design has the potential to encourage or discourage the use of certain test-taking strategies and that those strategies can have an impact on test-takers' performance and, subsequently, the construct measured by the test. For example, if test-takers can control the audio and are given access to the audio play controls and the questions at the same time, they are enabled to use the strategy of previewing the question and options before playing the audio. However, if test-takers are not allowed to access the questions until after the audio has been played, they are precluded from using the previewing strategy. Whether test designs should allow and prevent test-takers from using particular test-taking strategies should be informed by whether those strategies are relevant to the construct measured by the tasks that test-takers are asked to complete. In light of this principle, if a test contains questions that ask, for example, about the meaning of specific words from a text prompt, it is critical for those words to be highlighted in the text so that test-takers can easily find them. Similarly, test designers should indicate how many test items are related to each specific listening or reading prompt (as is done in the operational paper-and-pencil MET listening and reading sections) so that test-takers can decide which strategy to use (e.g., whether it is worth reading the whole text or just skimming through it).

It is also noteworthy that the choice and use of test-taking strategies might be affected by the level of test-takers' proficiency in the target language. Although all participants in this study were recruited from advanced-level ESL classes, Participant 13 stood out as someone whose English language proficiency was noticeably lower than that of other participants. Meanwhile, compared to other participants in this study, Participant 13 demonstrated the most extensive use of test-wiseness strategies during the test, which may suggest a possible interaction between proficiency level and strategy use. Had this study included participants from different levels of English language proficiency rather than only advanced-level L2 learners, the findings might have evinced other different types of test-taking strategies or different degrees of test-takers' reliance on test-wiseness strategies as opposed to test-management strategies. Further research is needed to corroborate this speculation.

This study also confirms the value of combining eye tracking and cued retrospective reporting to investigate test-taking strategies. It demonstrates that eye tracking can provide compelling evidence about temporal and spatial characteristics of test-takers' interaction with each test item. Unlike verbal data, eye-tracking data can reveal, for instance, which specific elements of each item the test-taker looked at and how long that interaction lasted. On the other hand, eye-tracking data can provide evidence about only the *whats* and *hows* of participants' oculomotor behavior (namely, what visual elements they looked at, as well as how and how long they looked at them); however, these data are not capable of explaining why participants engaged in a particular oculomotor behavior and what they were thinking about in the process. That is why complementing eye-tracking data with verbal report data allows for a more complete and in-depth understanding of the strategies used by test-takers to respond to individual test items.

This study has several limitations. The first limitation is that the experimental setting in this study was low-stakes for the participants, so the strategies that they used to answer the MET items might have been different from the strategies they would have used in a high-stakes, real-life testing context. This limitation consequently reduces the generalizability of the findings. Second, cued retrospective reporting was carried out in English rather than in participants' native languages both because of the practicality issue and because all participants came from advanced-level ESL courses. Given that Participant 13 had a noticeably lower level of overall English language proficiency than the other participants, asking this participant to verbalize her use of test-taking strategies in her native language might have resulted in a more comprehensive and detailed set of verbal data. Another limitation is that the scanpath analysis method was exploratory and conducted by one person. Having a second "coder" to assign test-taking strategies used by participants to answer each test item would have allowed for calculating inter-coder reliability and provided additional validity evidence for the assigned strategies. Finally, the data analysis carried out to answer Research Question 2 did not account for the use of strategy clusters to answer some items. Future research is therefore needed to tease apart the effect of strategies on test scores when test-takers use clusters of strategies.

Given that the multiple-choice format appeared to be conducive to the use of some test-wiseness strategies such as guessing and elimination of response options out of a vague sense that they might be incorrect, future research should investigate what test-wiseness strategies are employed for answering other item formats and the extent to which their use contributes to construct-irrelevant variance. In addition to the use of qualitative scanpath analysis, eye-tracking studies that employ quantitative measures are also needed to explore test-taking strategies in L2 assessment.

## Acknowledgements

I would like to acknowledge the support from the Center for Language & Technology at the University of Hawai'i at Mānoa. I am also very grateful to Cambridge Michigan Language Assessments (CaMLA) for funding this study through the Spaan Research Grant Program 2016.

## References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, *8*(1), 41–66.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Ball, L. J., Eger, N., Stevens, R., & Dodd, J. (2006). Applying the post-experience eye-tracked protocol (PEEP) method in usability testing. *Interfaces*, *67*, 15–19.

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tests. *Applied Linguistics*, *34*(3), 304–324.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, *30*(4), 441–465.

Bax, S., & Chan, S. H. C. (2016). Researching the cognitive validity of GEPT High-Intermediate and Advanced Reading: An eye tracking and stimulated recall study. *LTTC–GEPT Research Reports*, *RG–07*. Retrieved from https://www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG07.pdf

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York, NY: Routledge.

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study*. (ARAGs Research Reports – Online. Vol. 1, No. 1). London, UK: British Council. Retrieved from http://www.britishcouncil.org/sites/default/files/brunfaut-and-mccray-report_final.pdf

Cambridge Michigan Language Assessments (2016). *The Michigan English Test*. Retrieved from http://cambridgemichigan.org/institutions/products-services/tests/proficiency-certification/met/

Coco, M. I. (2009). The statistical challenge of scan-path analysis. In *Proceedings of the 2nd Conference on Human System Interactions* (pp. 369–372). Catania, Italy.

Cohen, A. D. (1998). Strategies and processes in test taking and SLA. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 90–111). Cambridge, UK: Cambridge University Press.

Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, *3*(4), 307–331.

Cohen, A. D. (2007a). Coming to terms with language learner strategies: Surveying the experts. In A. D. Cohen & E. Macaro (Eds.), *Language learner strategies: 30 years of research and practice* (pp. 29–45). Oxford, UK: Oxford University Press.

Cohen, A. D. (2007b). The coming of age for research on test-taking strategies. In J. Fox, M. Weshe, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 89–111). Ottawa, Canada: Ottawa University Press.

Cohen, A. D. (2011). *Strategies in learning and using a second language* (2nd ed.). New York, NY: Routledge.

Cohen, A. D. (2014). Using test-wiseness strategy research in task development. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 893–905). Malden, MA: Wiley-Blackwell.

Cohen, A. D., & Upton, T. A. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL®. *Language Testing*, *24*(2), 209–250.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage Publications.

Ehmke, C., & Wilson, C. (2007). Identifying web usability problems from eye-tracking data. In *Proceedings of the 21st British HCI Group Annual Conference on HCI* (pp. 119–128). Swinton, UK: UK British Computer Society.

Godfroid, A., & Schmidtke, J. (2013). What do eye movements tell us about awareness? A triangulation of eye-movement data, verbal reports, and vocabulary learning scores. In J. M. Bergsleithner, S. N. Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 183–205). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Hansen, J. P. (1991). The use of eye mark recordings to support verbal retrospection in software testing. *Acta Psychologica*, *76*(1), 31–49.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford, UK: Oxford University Press.

Kashkouli, Z., & Barati, H. (2013). Type of test-taking strategies and task-based reading assessment: A case in Iranian EFL learners. *Procedia – Social and Behavioral Sciences*, *70*, 1580–1589.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6*(2), 199–215.

Nikolov, M. (2006). Test-taking strategies of 12- and 13-year-old Hungarian learners of EFL: Why whales have migraines. *Language Learning*, *56*(1), 1–51.

O Sullivan, B., & Weir, C.J. (2011). Test development and validation. In B. O Sullivan (Ed.), *Language testing: Theories and practices* (pp. 13–32). Basingstoke, UK: Palgrave Macmillan.

Petrie, H., & Harrison, C. (2009). Measuring users' emotional reactions to websites. In *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3847–3852). Boston, MA: ACM.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing*, *20*(1), 26–56.

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561–587.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. *Language Testing*, *17*(1), 85–114.

Schmitt, N., Ng, J. W. C., & Garras, J. (2011). The Word Associates Format: Validation evidence. *Language Testing*, *28*(1), 105–126.

Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing*, *14*(2), 214–231.

Suvorov, R. (2013). *Interacting with visuals in L2 listening tests: An eye-tracking study*. (Unpublished doctoral dissertation). Iowa State University, Ames, Iowa, USA.

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing*, *32*(4), 463–483.

Tzanidou, E., Minocha, S., & Petre, M. (2005). Applying eye tracking for usability evaluations of e-commerce sites. In *Proceedings of the Workshop on 'Commercial Uses of Eye Tracking' Held at the 19th British HCI Group Annual Conference*. Retrieved from https://pdfs.semanticscholar.org/70ff/d9a9ded9a06d5426609bf882253d791ef1fc.pdf

Van Gog, T., Paas, F., Van Merriënboer, J. J. G., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied*, *11*(4), 237–244.

Vansteenkiste, P., Cardon, G., Philippaerts, R., & Lenoir, M. (2015). Measuring dwell time percentage from head-mounted eye-tracking data – comparison of a frame-by-frame and a fixation-by-fixation analysis. *Ergonomics*, *58*(5), 712–721.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, UK: Palgrave Macmillan.

Winke, P., & Lim, H. (2014). The effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation. *IELTS Research Reports Online Series*, 3. Retrieved from https://www.ielts.org/~/media/research-reports/ielts_online_rr_2014-3.ashx

Wu, A. D., & Stone, J. E. (2015). Validation through understanding test-taking strategies: An illustration with the CELPIP-General Reading Pilot Test using structural equation modeling. *Journal of Psychoeducational Assessment*, *34*(4), 362–379.

Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing*, *20*(3), 267–293.

Yi'an, W. (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.