Studies in English Language Teaching
ISSN 2372-9740 (Print) ISSN 2329-311X (Online)
Vol. 4, No. 1, 2016
www.scholink.org/ojs/index.php/selt

Integration of Stimulated Recall, Self-Observation, and Retrospective Interview in the Collection of Strategy Data in Computer-Assisted Language Testing

Mohammed S. Assiri[1*]

[1] English Language Center, Institute of Public Administration, Riyadh, Saudi Arabia

[*] Mohammed S. Assiri, E-mail: assirim@ipa.edu.sa

Abstract

Research on language learning and use strategies has made extensive use of procedures that involved self-reporting and/or -revelation in data collection. However, scholarly reviews have pointed to certain flaws associated with such procedures especially whenever one procedure was used by itself. On one hand, strategies revealed through self-reporting (e.g., questionnaires) do not accurately represent the actual strategies used in response to language tasks. On the other, self-revelation (e.g., think-alouds) interferes with strategy use on language tasks as well as task performance. Drawing on empirical evidence, this paper proposes that the integration of three procedures of verbal reporting, namely stimulated recall, self-observation, and retrospective interview, in computer-assisted research can tremendously help capitalize on their strengths and control their weaknesses.

Keywords

stimulated recall, self-observation, retrospective interviews, test-taking strategies, computer-assisted language testing

1. Introduction

In L2 testing, researchers have generally been encouraged to use a mixed-method approach to get deep understanding of the interaction among constructs that reflect ability and performance, for example, strategy use, proficiency, and test performance (Phakiti, 2003; Yien, 2001). In this article, the integration of three methods of verbal reports (henceforth, SRSORI technique where SR = stimulated recall; SO = self-observation; RI = retrospective interview) in research design is illustrated using a study by the author (Assiri, 2011). Stimulated recall, self-observation, and retrospective interview represent types of verbal protocols that can reveal “what information [subjects] are attending to while performing their tasks, and by revealing this information, can provide an orderly picture of the exact way in which the tasks are being performed: the strategies employed, the inferences drawn from information…” (Ericsson & Simon, 1993, p. 220). What follows is a description of each one of these three methods:

1.1 Stimulated Recall

Stimulated recall is considered by scholars in the fields of language learning and teaching (e.g., Mackey & Gass, 2005) to be an inner-directed measure in which the learner is provided with a stimulus and engaged in reflections of the thought processes she had in mind while performing a language task. It is "an information processing approach whereby the use and access to memory structures is enhanced, if not guaranteed, by a prompt that aids the recall of information" (Gass & Mackey, 2000, p. 17). The prompt or stimulus can be either video or audio, or both, and serves to stimulate the learner's recollection of her mental thoughts during task performance. For instance, having just been in a conversation, learners are asked to listen to an audio recording of the conversation and use the recording to help them remember their mental thoughts during the conversation, and at the same time verbalize the thoughts and reactions they experienced while conversing.

Stimulated recall is used in strategy research for its potential to help researchers "to determine when and if particular cognitive processes, such as search, retrieval or decision making are employed" (Gass & Mackey, 2000, p. 21). One advantage to the use of stimulated recall over think-aloud is that any method-related reactive effects on task performance are avoided (Sanz, 2005). Another advantage is that during a stimulated recall session, the subject is totally focused on recalling and reporting of his thoughts (Kelly, 2009). This advantage stands in contrast to think-aloud where the subject performs the cognitive task and verbalizes his thoughts both at the same time, which Ellis (2001) referred to as "dual processing" (p. 37). Stimulated recalls help strategy researchers in particular to elicit "task-specific strategy descriptions with corroborating evidence of their use" (Chamot, 2005, p. 58).

1.2 Self-Observation

Self-observation can be defined as "the inspection of specific, not generalized, language behavior, either introspectively (i.e., within 20 seconds of the mental event) or retrospectively" (Cohen & Upton, 2006, p. 13). Accordingly, the learner verbalizes the line of thinking he went through while performing a given language task; for example, the learner may say "[w]hat I just did was to skim through the reading passage for possible surface matches between information in the text and that same information appearing in one of the alternative choices" (Cohen & Upton, 2006, p. 13). Therefore, self-observation can be said to represent a midway solution between self-reporting and self-revelation. The use of self-observation in strategy research gains support from the work of Pressley and Afflerbach (1995) in which they argued that strategic processes can be verbalized by having respondents take their time attending to more specific bits and pieces of their thoughts and by researchers' providing respondents with prompts to elicit more specific information and explanation (Desimone & Le Floch, 2004).

Similarly, Cohen (1996) noted that strategy researchers are advised to make the shift from using self-report questionnaires to self-observational procedures that enable them to tap into mental processes as close as possible to the moment these processes were generated. He added that the data collected through self-observation are more likely to "reflect accurately what learners actually do than might the response to a questionnaire item calling for a description of generalized behavior" (p. 13). Based on their research, MacLean and d'Anglejan (1986) argued that the use of self-observation as a retrospective technique furnishes a wealth of information about the kinds of resources learners draw on, be they cognitive or technical, during their performance of language tasks. These researchers also maintained that such information as that obtained through self-observation can serve as an exploratory tool to detect aspects of successful versus poor performance.

1.3 Retrospective Interview

The retrospective or post-task interview involves the use of verbal cues to have respondents recollect and report the thoughts they had in mind while performing a cognitive task, with some elaboration, description, or explanation. A number of scholars have championed the use of retrospective interviews following task performance to elicit descriptions and explanations of thought processes that could help address questions regarding strategy synchronization and compatibility, and the relationships between strategy use and test takers' characteristics (Nikolov, 2006) and test performance (Leighton, 2004; Phakiti, 2008). In studies where retrospective interviews were used after think-alouds (e.g., Cohen & Upton, 2006; Nikolov, 2006; Nguyen, 2008), the researchers seemed to agree about the usefulness of retrospective interviews in supplementing and checking data collected by means of think-alouds and facilitating discernment of the kinds of strategies the respondents used. Other researchers (e.g., Greene & Higgins, 1994; Pressley & Afflerbach, 1995; Yamashita, 2003) have highlighted the value of retrospective interviews that make use of premeditated prompts in tapping into processes not readily available for immediate recall or report.

Retrospective interviews can also be useful in eliciting metalinguistic information that may account for certain aspects of performance (Callies, 2009), steps of decision making during task performance (Robinson, 1992), and motives that spur language behaviors (Ross, 1997). Researchers can use retrospective interviews to check introspective reports for accuracy and to have respondents relate their metacognitive and metalinguistic knowledge to their reports (Taylor & Dionne, 2000). Retrospective interviews allow the researcher to ask any questions she may have, paraphrase and elucidate her questions, and use the informant's responses to formulate other questions (Chamot, 2005). In a study of language learning strategies, a retrospective interview using a stimulus has the capacity to "accurately reveal students' actual learning strategies because it is conducted immediately after a learning task" (Chamot, 2005, p. 113). As an example of how this can be accomplished, the researcher may videotape the learner's performance, play it back, and pause whenever the researcher wishes to have the learner explicate his thought processes. Retrospective interviews are conducted in a personal fashion. This helps the researcher provide the necessary guidance and coaching to the interviewee, attend to many cues that can provide a wealth of information about interviewee's performance, and collect data that are better reflective of the phenomenon under study than other forms of data collection (Hawley, 2003).

1.4 SRSORI in Previous Research

Taken on an individual basis, the three methods of stimulated recall, self-observation, and retrospective interview have proven helpful in the collection of strategy data in a number of studies. For example, Sasaki (2000) asked her respondents to provide verbal reports once they completed a cloze task so that she could determine how cultural schemata relate to the use of test-taking strategies and processes. Hudson and Park (2002) had their respondents think aloud their reactions in response to web-based tests of listening and reading comprehension and engage in retrospective interviews to describe their test-taking experiences in more details. In a study of metacognitive and cognitive strategy use on a reading test, Phakiti (2003) used retrospective interviews with two groups of higher and lower scorers to elicit detailed explanations of their responses to a strategy questionnaire.

Brown, Iwashita and McNamara (2005) had a group of raters verbalize their cognitive processes after rating samples of speeches to explore aspects of their rating performances. With the aim of determining whether multiple-choice questions make reading differ from testing to non-testing situations, Rubb, Ferne and Choi (2006) elicited think-aloud protocols from their respondents that basically involved their strategy use and response behaviors. Xu and Wu (2012) used a combination of think-aloud protocols, retrospective interviews, and questionnaires to find out what test-taking strategies their respondents would use while completing writing tasks involving picture prompts. Besides their usefulness as means that help researchers gain access to strategic thoughts, research evidence suggests that verbal reports help respondents become better aware of their strategy use (Assiri, 2011) and endorse their learning of language forms (Bowles, 2010).

1.5 Merits of SRSORI

Using two or more methods together helps us take full advantage of their strengths in a way that can counterbalance any weaknesses when a method is used by itself. The procedural integration of stimulated recall, self-observation, and retrospective interview (or SRSORI) into a research design stems from the fact that these three methods represent retrospective measures; that is, they are used following a cognitive task and they involve eliciting data about how the subject carried out the task. The integrated-method approach has support in the literature on qualitative research methodology. In this respect, many qualitative research specialists (e.g., Kelle, 2001; Konecki, 2008; Moran-Ellis et al., 2006) have commended this approach so that a researcher can come to grips with a multi-faceted phenomenon. According to these experts, through the use of mixed methods, the researcher can gain insights into the nature and intricacy of the phenomenon under study, become well-informed about the inner workings of the various elements of a multi-dimensional phenomenon, and increase the likelihood of achieving credible results and findings.

In a mixed-method approach, integration of methods can be made by merging these methods such that while each method keeps its own distinct design features, it complements the functions of the other methods and contributes perceptibly and meaningfully to the purpose of the whole procedure (Moran-Ellis et al., 2006). One concern linked specifically with the use of self-observation and retrospective interview is the gradual decline of memory as time passes since the completion of the cognitive task (Cohen, 1986). However, when these procedures are combined with stimulated recall that uses a stimulus or stimuli to augment the informant’s recollection of his thoughts during task performance, the problem of memory decline can be countered. Of course, the time lapse between task performance and verbal protocols, on the other hand, correlates inversely with retention and recollection of mental thoughts accompanying the performance of the task. Nevertheless, in the case of stimulated recall, the stronger the stimulus is and the sooner the recall occurs after task performance, the greater are the accurateness and completeness of the subject's recollections of the thoughts linked with task performance (Mackey & Gass, 2005; Sanz, 2005). Besides, researchers who used retrospective interviews (e.g., Polio, Gass, & Chapin, 2006) have called attention to the role stimuli can play in eliciting accurate and complete data.

2. SRSORI in a Test-Taking Strategy Study

The aforementioned merits of methodological integration apply to the use of the SRSORI technique by Assiri (2011) since all procedures combined would be expected to benefit from both the strength of the stimulus and the immediacy of the retrospection process. When using such procedures as those that draw upon memory and retention, researchers are advised to bring into play as many stimuli available during task performance as they can in order to cause the informant "to relive an original situation with great vividness and accuracy" (Bloom, 1954, p. 25). Generally, it was assumed that the use of the SRSORI technique in Assiri (2011) would benefit from a feature that characterizes self-observation, to be exact, the subject's reflections and descriptions of thoughts she had during task performance. Another merit was that subjects could reflect on ways of learning and problem-solving specific to their individual cases (see Cohen & Hosenfeld, 1981).

The kind of stimulus used as part of the stimulated recall process could further the subject's ability to elaborate on and account for her thoughts at every single point of task performance. Retrospective interviews using probes, on the other hand, would allow the researcher to access the kinds of metacognitive strategies and processes that the informant relied on during task performance (see Taylor & Dionne, 2000). Moreover, because retrospective interviews would be conducted directly after task performance, they could reveal more complete and accurate information on problem-solving techniques and steps as used by the test takers (see Ericsson & Simon, 1993; Pressley & Afflerbach; 1995). It is worth noting here that drawing on previous research, precautions were taken in Assiri's (2011) study against any pitfalls associated with the use of retrospective measures, and recommendations for optimum use of these measures were followed.

2.1 Method

The use of SRSORI in Assiri (2011) followed each participant's response to each of two computerized-reading sets in two sessions, one at a time. The participant was asked to make use of a stimulus, in the form of a playback of screen recording of his or her response to each reading set, to help him or her recall thoughts he or she had in mind while responding to the reading set. He or she was instructed to describe and explain these thoughts as much as he or she could. At the same time, the researcher provided necessary prompts and asked opportune questions to elicit more details. The two

SRSORI sessions for each participant were mainly in Arabic, and were both audio-recorded in a digital format. Figure 1 below depicts how the use of the three procedures followed a sequence and at the same time how they converged in the production of strategy data in the SRSORI technique.

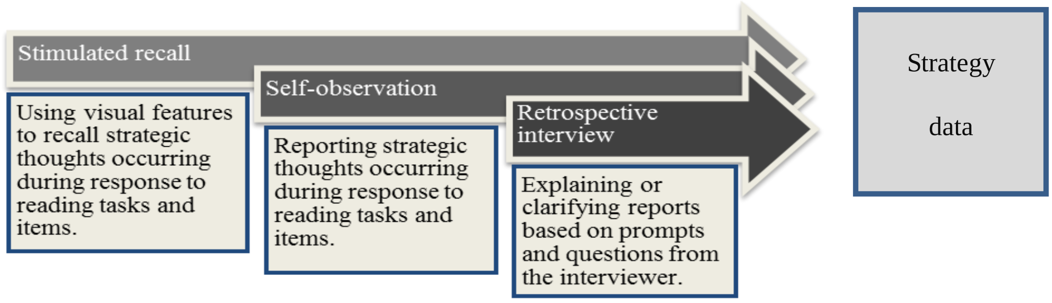

Stimulated recall
Using visual features to recall strategic thoughts occurring during response to reading tasks and items.
Self-observation
Reporting strategic thoughts occurring during response to reading tasks and items.
Retrospective interview
Explaining or clarifying reports based on prompts and questions from the interviewer.
Strategy data

Figure 1. Sequence and Convergence of SRSORI Procedures in Producing Strategy Data

2.1.1 Instruments

The two reading sets had the same tasks that appear on the actual TOEFL-iBT reading section. They were part of one of two authentic tests in The Official Guide to the TOEFL Test (ETS, 2009). The guide has a companion CD that has the two tests administered in a computerized form and so it uses automated scoring with the objective sections of the tests, including the reading section. Each reading set consisted of a 600-700 word text followed by 12-13 question items. One reading set was chosen to be from a field that belongs to the arts, whereas the other was from a science field so as to balance out the effects of text content of the reading sets across two potential groups of participants according to their fields of study.

Each reading set includes test items designed to evaluate reading for Basic Comprehension, Inferencing, and Reading to Learn skills. Basic Comprehension (BC) has five item types: Vocabulary (BC-v), factual information (BC-fi), negative fact (BC-nf), pronoun reference (BC-pr), and sentence simplification (BC-ss). The Inferencing (I) three item types are basic inference (I-bi), insert text (I-it), and rhetorical purpose (I-rp). The Reading to Learn (R2L) tasks are prose summary (R2L-ps) and schematic table (R2L-st). Table 1 below outlines descriptive information about the reading sets, including the topic of each set, type of text genre, total number of question items, item types, and total score. Each item is worth one point except for the last item on both reading sets, which receives partial credit (0-2 points for item 13 on reading set 1 and 0-3 points for item 12 on reading set 2).

As stated by the ETS (2009), the two tests in the TOEFL-iBT guide and on the CD companion are actual TOEFL-iBT tests used in previous administrations. The ETS (2009) argues that the two tests can provide the test taker with "an estimate of how [he or she] would perform on the actual exam" (p. 3). The TOEFL iBT has established validity and reliability (See also ETS, 2011a, 2008, 2011b), and uses unbiased objective scoring (ETS, 2009). As for authenticity and face validity (cf., Alderson, 2005), the kinds of testing tasks used on the TOEFL iBT mirror the tasks students are expected to perform in academic settings (ETS, 2009).

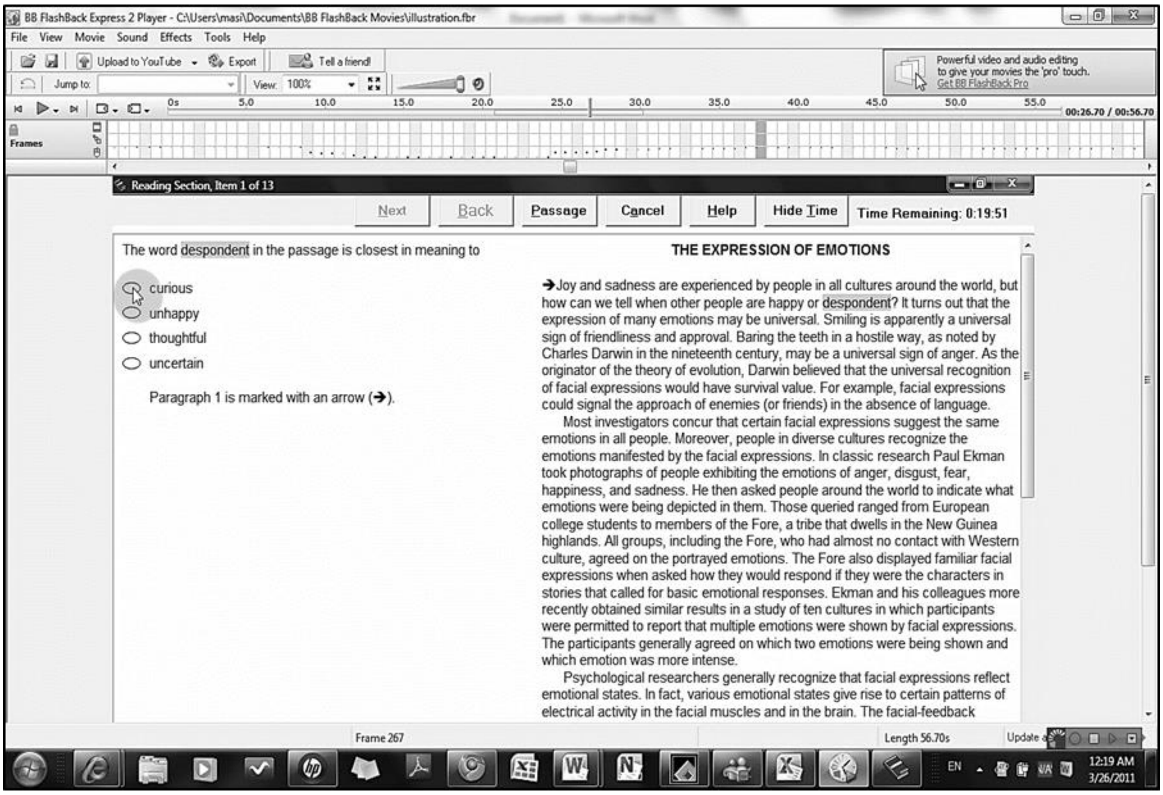
Table 1. Descriptive Information about the Reading Sets

	Topic of set	Type of genre/field	Number of items	Item types	Total score
Reading set#1	Nineteenth-Century Politics in the United States	historical, arts	13	1. BC-*v* 2. BC-*fi* 3. I-*rp* 4. BC-*fi* 5. BC-*fi* 6. BC-*v* 7. BC-*fi* 8. BC-*v* 9. I-*bi* 10. BC-*nf* 11. BC-*ss* 12. I-*it* 13. R2L-*ps*	14
Reading set#2	Geology and Landscape	expository, sciences	12	1. BC-*fi* 2. BC-*v* 3. I-*bi* 4. BC-*v* 5. BC-*fi* 6. I-*rp* 7. BC-*v* 8. BC-*pr* 9. BC-*ss* 10. BC-*fi* 11. I-*it* 12. R2L-*st*	14

The scores the participants received on the reading sets were used to determine which participants fit in the high- and low-scoring groups. Further, based on score reports, participants' successful answers to the question items on the reading sets were used to identify effective aspects of test-taking strategy use in relation to task and item types.

2.1.2 Pilot Study

Before collecting data for the main study, a pilot study was conducted with five volunteers from the target population of the study in order to check the adequacy of, and make any necessary refinements in, research procedures and materials. As a result, the instructions and directions that were part of the orientation and training of participants were made clearer and more elaborate. The modeling and the training steps of verbal report were improved in two ways. First, the modeling was made expressive of more details so as to make it clear to the participant that while producing verbal reports, he or she is expected to provide as much detailed reports as possible. Such reports should include even mental and physical behaviors that participants may consider unworthy of mentioning, because these behaviors might still be valuable to the research.


BB FlashBack Express 2 Player - C:\Users\masi\Documents\BB FlashBack Movies\illustration.fbr
File View Movie Sound Effects Tools Help
Upload to YouTube Export Tell a friend!
Jump to: View: 100%
Powerful video and audio editing to give your movies the 'pro' touch. Get BB FlashBack Pro
0s 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0 45.0 50.0 55.0
00:26.70 / 00:56.70
Frames
Reading Section, Item 1 of 13
Next Back Passage Cancel Help Hide Time
Time Remaining: 0:19:51
Frame 267
Length 56.70s
Update
EN
12:19 AM
3/26/2011

The word despondent in the passage is closest in meaning to

○ curious
○ unhappy
○ thoughtful
○ uncertain

Paragraph 1 is marked with an arrow (➔).

THE EXPRESSION OF EMOTIONS

➔Joy and sadness are experienced by people in all cultures around the world, but how can we tell when other people are happy or despondent? It turns out that the expression of many emotions may be universal. Smiling is apparently a universal sign of friendliness and approval. Baring the teeth in a hostile way, as noted by Charles Darwin in the nineteenth century, may be a universal sign of anger. As the originator of the theory of evolution, Darwin believed that the universal recognition of facial expressions would have survival value. For example, facial expressions could signal the approach of enemies (or friends) in the absence of language.

Most investigators concur that certain facial expressions suggest the same emotions in all people. Moreover, people in diverse cultures recognize the emotions manifested by the facial expressions. In classic research Paul Ekman took photographs of people exhibiting the emotions of anger, disgust, fear, happiness, and sadness. He then asked people around the world to indicate what emotions were being depicted in them. Those queried ranged from European college students to members of the Fore, a tribe that dwells in the New Guinea highlands. All groups, including the Fore, who had almost no contact with Western culture, agreed on the portrayed emotions. The Fore also displayed familiar facial expressions when asked how they would respond if they were the characters in stories that called for basic emotional responses. Ekman and his colleagues more recently obtained similar results in a study of ten cultures in which participants were permitted to report that multiple emotions were shown by facial expressions. The participants generally agreed on which two emotions were being shown and which emotion was more intense.

Psychological researchers generally recognize that facial expressions reflect emotional states. In fact, various emotional states give rise to certain patterns of electrical activity in the facial muscles and in the brain. The facial-feedback

Figure 2. A Snapshot of Screen Recording

Second, the practice step was timed in order to simulate the experience of responding to the reading sets under time allotment. Another reason was that practicing with no time limit might cause participants to use strategies they would not use in real-testing conditions. The use of the screen recording feature as a stimulus in the SRSORI procedure gained support from all of the volunteers in the pilot study (For illustration, Figure 2 above shows a snapshot of the sample reading set in the display mode of the screen-recording program).

Other improvements made on the basis of the pilot study included providing participants with a tutorial on how to respond to the reading set using a sample set having the same format as those sets used in the actual data collection sessions. The tutorial was necessary because some of the volunteers in the pilot phase of this research expressed their forgetfulness of how some of the task items looked and how to select the answer to certain task items. It was also seen necessary to provide a demonstration of how to use the cursor in the sample clip to express thoughts linked with its movements during the response activity. In addition, the pilot phase helped the researcher develop the habit of providing selective prompts during the SRSORI procedure. These improvements can be said to have enhanced both validity and reliability of data collection materials and procedures.

2.1.3 Data Collection

The materials that were crucial elements of the data collection process were made available and ready, including a laptop computer, the TOEFL-iBT guide CD companion, the screen recorder program, and a SONY digital recorder. The screen recorder program, named BB FlashBack Express (Blueberry Software, 2009), records any online activity, or activity that appears on a computer screen, and produces a movie of the whole activity. This program was used in this research as a stimulus, and proved to be a significant feature of the SRSORI technique. Interviews were scheduled with participants to be conducted at a time convenient to them. The data collection setting was selected such that it provided an atmosphere as quiet and conducive to optimal test-taking and interviewing as possible.

The researcher's first priority was to establish rapport with the participant through informal conversation in order to facilitate subsequent communication. The researcher used a prepared step-by-step guide as to what to do from the start to the end of each data collection session. The participant and the researcher sat side-by-side at a table on which the laptop computer, the audio-recorder, and the other materials were placed. Arabic was used as a medium of communication in all data collection procedures in order to ensure that all participants understood what they were expected to do and to collect data not affected by any language deficiencies in using English. Nonetheless, a mix of Arabic and English was permitted so long as this did not interfere with providing accurate and complete reports of strategy use on the reading tasks.

After completing the setup of the research materials, the participant was requested to complete a participant background questionnaire. Next, the participant was oriented to the research procedures and steps. The participant was requested to deal with the reading sets in the same manner he or she would with actual reading sets on the TOEFL iBT. In order to motivate participants to respond to the reading sets to the best of their abilities, to achieve scores as high as they could, and to make this experience of consequential value to them, each participant was informed that he or she would be rewarded a sum of money from $10 to $20 depending on his or her total score on the reading sets.

Then, the participant was introduced to verbal report, and given directions on how to verbalize his or her thoughts that accompanied the performance of a given task. The participant was directed that when producing his or her verbal reports, he or she should express any thoughts he or she had in mind while responding to the reading tasks. The participant was informed that he or she could use Arabic, or a mix of Arabic and English only if this did not interfere with his or her providing accurate and complete verbal reports. The participant was also informed that the researcher would provide some prompts or ask questions to elicit more information or clarification from him or her as part of the verbal report.

Next, verbal report was modeled by the researcher using a grammar exercise instead of a reading task so as not to lead or influence the participant towards reporting certain reading-test-taking strategies, or to ward off the so-called halo effect. The participant was then asked to practice verbal report, using a reading mini-test composed of a short text followed by a multiple-choice question item. Before practice, the participant was reminded that he or she was expected to express any thoughts he or she had in mind while responding to the reading mini-test, using Arabic or a mix of Arabic and English as he or she wished, and that the researcher might provide some prompts or ask questions to elicit more information. The practice step was timed in order for the participant to work under roughly the same conditions he or she would experience when responding to the reading sets. Upon completing this task, the participant was requested to verbalize whatever he or she thought about to get to the answer to the mini-test. The researcher provided the participant with prompts and questions whenever further clarification or explanation was needed.

At the end of the practice session, the participant was provided with feedback on his or her verbal report, and offered recommendations as necessary. The training was also aimed to familiarize the participant with the researcher and the setting. Assuming that the reading mini-test could serve as a stimulus, it can be noted that each participant was also trained in stimulated recall (see Mackey & Gass, 2005). At the end of the training session, the participant was asked if he or she had any questions about verbal report, and his or her questions, if any, were answered. Arabic was used during the training session at the three stages of giving instructions, modeling, and practice.

Afterwards, the researcher ran a sample reading set (entitled The Expression of Emotions) to demonstrate for the participant how to go through the reading tasks, starting from the instructions to pressing the Finished Button to end the test-taking session. The tutorial involved going through the instructions; displaying how to use the interactive features of the interface, including a demonstration of how to move to the next question and how to go back to a previous one, how to go to the passage in a full view, and how to select answers to the question items, and demonstrating what to do after completing the reading set. It was brought to the respondent’s attention that he or she should try to complete the reading set before the time expires. Although participants were already familiar with the formats of the task and item types on the test, it was assumed that providing them with such a tutorial would help in two ways. First, it would refresh their memories of task and item formats. And second, it would minimize the effect of any differences among them that are due to time length since they last took the test.

Using a short clip of the screen recording of the sample reading set, a demonstration was provided of how to use the arrow on the screen to express thoughts linked with certain movements of the cursor during the response activity. These thoughts were basically about technical matters (e.g., run the program, press Continue Button, choose an answer, etc.) in order not to cue the participant in certain test-taking strategies. The participant was informed that he or she would see a screen recording of his or her response to each reading set, and that he or she would use this recording as a stimulus to help him or her recall the thoughts he or she had in mind during test-taking. At the end of the tutorial, the participant was asked if he or she had any questions and if he or she was ready to respond to the first reading set.

Once the participant expressed his or her readiness to respond to the first reading set, the researcher ran the screen recorder, and immediately the participant was instructed to start responding to Reading Set 1. The reading set was timed for 20 minutes and the time remaining showed on the window the participant was interacting with. The time allotted for the participant to complete a reading set was similar to that of the actual TOEFL iBT. Consequently, this presumably made the participants live the experience of actual test-taking, and so their strategy use in the research setting would be considerably similar to that in an actual TOEFL-iBT setting (see Anderson et al., 1991). Next, the researcher checked to make sure that both the TOEFL-iBT program and the screen recorder were running well, and then he left the participant to work on his or her own and sat at another nearby table, yet not directly visible to the participant.

Upon the participant’s completion of the reading set, the researcher stopped the screen recorder and saved the recording as a video file. A few participants were not able to complete one or the other of the given reading sets before the time expired; thus, these participants were permitted to complete the reading set even after the time expiration. Otherwise, ending the testing session by time expiration would have led to excluding the last item type, namely the R2L task item, from the collection and analysis of data from these participants. In all of these cases, the participants were able to press the Finished button in just one to two minutes after time expiration. Because the TOEFL-iBT guide software is intended to serve practice purposes, it allows the user to work on reading sets even after the time expires.

Upon the respondent’s completion of the reading set, the researcher informed him or her that the verbal report procedure would be started. The researcher operated the audio-recorder and played the screen recording. Then, the participant was engaged in stimulated recall combined with self-observation using the screen recording as a stimulus, while the researcher was listening attentively, providing the participant with specific and purposeful prompts and/or questions to elicit further clarification or explanation. Such prompts and questions made use of visual cues to enable the participant to remember many more details. To avoid interrupting the respondent’s recall and self-observation, prompts were generally provided whenever the respondent paused. In addition, the questions were often posed when the respondent showed that he or she had made a complete verbalization of thoughts linked with the response to a given question item. Guidance was also offered when it was needed, for example, to have the participant verbalize his or her mental thoughts as they occurred during the test-taking activity when he or she started describing what he or she usually does in response to a given task or item.

To control for the effect of individual differences among respondents in the manner in which they reported their strategy use, the researcher offered prompts and questions to respondents in more of an informal, relaxed, and individualized manner in order to ensure more effective and complete reporting on the respondent's part (see Patton, 2004). Next, the respondent was offered a break in between two data collection sessions. Afterwards, the participant was asked if he or she was ready to respond to the second reading set. Upon the participant's agreement, the researcher ran the screen recorder and immediately instructed the participant to start Reading Set 2. Then, the same procedures as those followed in the first data collection session were followed exactly in the second session.

At the end of the whole data collection procedure, the participant was provided with his or her scores on the reading sets, using a feature in the TOEFL-iBT guide CD that shows scores and correct versus incorrect answers. Besides, this feature furnishes the user with key answers along with explanations for these answers. Based on his or her total scores on the two reading sets, each participant was rewarded a sum of money from $10 to $20 depending on his or her total score—out of 28—on both reading sets. The participant was then debriefed about his or her performance and thanked for participating in the research.

2.1.4 Data Analysis

After all collected data were transcribed, verbal reports associated with certain items were selected for coding based on the results of item analyses. A coding scheme was constructed for each item type by means of inductive coding of transcribed data. Combining both quantitative figures and qualitative accounts of the data was deemed ideal to address the three research questions of this study. On the one hand, quantitative figures in terms of frequencies would indicate patterns of strategy use; on the other, qualitative accounts would describe in more detail how strategies relate to such factors as item type, scoring level, and the choice of the right answer (cf., Creswell, 2009).

1) Quantitative Analysis

Data coding led to the development of a strategy list for each item type. Occurrences of strategies in the transcripts were tallied in order to calculate how frequently strategies were used with the given item types. The reason why occurrences of strategies were quantified was three-fold: First, to find out tendencies among participants with respect to what strategies they used frequently with which item types; second, to detect potential differences between top and low scorers regarding which group tended to use which strategies more frequently with the given item types than the other; and third, to explore aspects of effective strategy use through examination of clusters of frequent strategies that seemed to have resulted in the provision of the right answers to item types. Calculating frequencies of strategy use for each item type proved useful as each item called for the use of a set of strategies specific to its type.

Because the selected item types had among them pairs of BC-v, BC-fi, BC-ss, I-bi, I-rp, and I-it item types and single items representing the other item types, the former set of item types would presumably show higher strategy frequencies than the latter. To deal with this issue, raw totals of strategies were converted into type/token ratios by dividing the number of occurrences of the strategy by the number of items of a type selected for data analysis. For example, on both of BC-v items across the two reading sets, 18 participants reported that they read the sentence containing the target word and the surrounding context (or strategy BC-v4), so the type/token ratio for this strategy equals the result of dividing 18 by 2, which is 9. Certain strategies were grouped together since they tended to occur together with a given item type.

2) Qualitative Analysis

This part of data analysis focused on offering qualitative accounts of strategy use by participants across item types, potential differences between high- and low-scoring participants in their strategy use, and aspects of strategy use that could be related to the choice of the right answer. That is, there were three goals of this analysis: first, to describe strategy use among participants as they were trying to answer each item type on the reading sets; second, to explain potential differences between high- and low-scoring participants in their strategy use; and third, to inductively identify and categorize themes that seemed to characterize effective strategy use. It was assumed that these qualitative accounts would reflect a clear and accurate picture of the nature of strategies that participants tended to use with the item types, potential discrepancies between the high and low scorers among participants in their strategy use, and aspects of effective strategy use that seemed to have helped test takers perform well in response to the task and item types.

3. Findings

The findings of Assiri (2011) are summarized here with reference to strategy use by task and item type, high- versus low-scorers' strategy use across task items, and aspects of effective strategy use across item types.

3.1 Strategy Use by Task and Item Type

Strategies used by test takers were classified into two groups: common strategies that were used across a variety of item types, and specific strategies that exhibited a high level of item dependency. Both levels of familiarity with item formats and language proficiency determined the extent to which test takers made use of test-management versus test-wiseness strategies. While test takers used strategies most of which were test-management strategies, they resorted to test-wiseness strategies with items they found challenging. Test-wiseness strategies involved systematic use of a variety of technical and textual elements. Strategy use among test takers had three facets. First, it was purposeful in that test takers applied certain strategies for different goals. Second, it was multi-form because certain strategies were used by test takers with some variation across several item types. And third, it was resourceful since strategies were often applied using different means that were largely determined by item formats in addition to test takers' reading ability and test-taking skill.

3.2 High-Versus Low-Scorers' Strategy Use across Task Items

Strategies of high scorers among test takers were dependent upon item format across most of the item types, which points out high scorers' awareness that individual item types call for distinct strategies. High scorers made use of strategies that combined aspects of both a high level of understanding of textual information and a superior skill of test-management. Top scorers were also disposed to use strategies involving confirmation of the selected answers and check of potential answers, which indicates a high level of strategic monitoring. They consulted background knowledge in such a manner that helped them relate their understanding of item content to real-world facts and experiences. On the other hand, low scorers adhered to certain strategies regardless of item formats and demands, and most of these strategies involved the use of clues and guessing. They made use of certain strategies to compensate for deficient comprehension skills, such as rereading the text or part of it. Although both high- and low-scoring groups used such strategies as elimination of options, reading of the options before the text, and use of clues, they differed in their intentions behind the use of these strategies.

3.3 Aspects of Effective Strategy Use across Item Types

The manner in which test takers sequenced strategies determined the extent to which their strategy use was effective. Source strategies in strategy sequences possessed both a high level of compatibility with item formats and flexibility of accepting other strategies as attached strategies. Aspects of effective strategy use among test takers can be summarized as follows: a) certain attached strategies were modified forms of source strategies, b) certain attached strategies endorsed the functions of source strategies, c) source strategies facilitated the functions of attached and subsequent strategies, and d) attached strategies were synchronized with source strategies. There were other aspects of effective strategy use that were prompted by behaviors of specific strategies in strategy sequences. First, the strategy of option elimination was either synchronized with or used after source strategies in order to reduce the options available for consideration.

In three-strategy sequences, the strategy of option elimination assumed either the second position to support the function of a subsequent strategy with BC-nf, or the third position to draw on an antecedent strategy with I-rp. Second, the strategy of answer confirmation tailed strategy sequences in which it was derived from either a source strategy or an attached strategy, depending on which strategy was critical to the item response. And third, the strategy of deciding on an option tailed strategy sequences in which it served to resolve a state of hesitation between two or more options, as determined by application of an antecedent strategy. It was also found, based on an additional analysis, that the ability to use effective strategy sequences made an important difference between high- and low-scoring groups in favor of the former.

4. Concluding Remarks

The findings of Assiri's (2011) study adequately addressed its research goals with respect to how strategy use among 25 Arab ESL learners interacted with item format and performance on the TOEFL-iBT reading tasks. The findings generally broke new ground for exploration of strategy use on the TOEFL-iBT reading section, among respondents who considerably shared linguistic and cultural backgrounds. They manifestly contributed to the knowledge and research bases of language testing and assessment, specifically proficiency-oriented testing. Such a contribution pertained to three converging areas of strategy use among Arab ESL learners on the reading section of the TOEFL iBT. First, test takers used strategies depending on item format and difficulty in ways that allowed them to achieve different goals, adapted certain strategies to various item types, and applied strategies using several textual and technical means.

Second, high-test performance and scoring were characterized by superior skills of both comprehension and test-management as well as high levels of strategic awareness and monitoring. Conversely, low-test performance and scoring were associated with poor skills of comprehension and excessive use of test-wiseness. Third, test takers sequence strategies such that certain strategies derived from other strategies, endorsed or facilitated the functions of other strategies, or acted in sync with other strategies. Such strategy sequences were linked with other aspects of effective strategy use that included reducing the options available for consideration, confirming the choice of the right answer, and deciding on an option among two or more options.

It is interesting to see how the procedural integration of stimulated recall, self-observation, and retrospective interview as used by this study was so fruitful in addressing the research goals and questions formulated at the outset. The merits of using the SRSORI technique in this study were as follows. First, it produced data comparable to data collected through a think-aloud procedure, but without any interference with task performance. Second, it warded off the problem of boring the subject by having him first perform the cognitive task and think-aloud his mental thoughts, and then sit for a retrospective interview. Such a problem could have affected the quality and quantity of the data reported. Third, it had the potential to reveal rich information about individual differences among test takers in their performance of the reading-testing tasks. Gass and Mackey (2000) made the same last comment about combining stimulated recall and retrospective interview. In fact, the richness of the data collected by means of the SRSORI technique in the study can certainly meet any needs for further exploration of strategy use on the TOEFL iBT reading tasks. The use of SRSORI as a means of data collection can definitely be extended to and tried in the study of strategy use on tests of other language skills or modalities.

Acknowledgment

I would like to express my sincere gratitude to Dr. Gene Halleck, Professor of English at Oklahoma State University, for her comments and suggestions on earlier versions of this paper.

References

Alderson, J. (2005). *Diagnosing foreign language proficiency: The interface between assessment and learning*. New York, NY: Continuum.

Anderson, N. (1991). Individual differences in strategy use in second language reading and testing. *Modern Language Journal*, *75*(4), 460-472.

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

Assiri, M. (2011). *Test-taking strategy use on the reading section of the TOEFL iBT: A study of Arab ESL learners* (Doctoral dissertation). Available on ProQuest dissertation and theses database.

Bloom, B. (1954). The thought processes of students in discussion. In S. French (Ed.), *Accent on teaching: Experiments in general education* (1st ed., pp. 23-46). New York, NY: Harper & Brothers.

Blueberry Software. (2009). *BB FlashBack Express* (Version 2.7.1) [Computer software]. Birmingham, UK: Author.

Bowles, M. (2010). Concurrent verbal reports in second language acquisition research. *Annual Review of Applied Linguistics*, *30*, 111-127.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tasks* (TOEFL Monograph No. MS-29). Princeton, NJ: ETS.

Callies, M. (2009). *Information highlighting in advanced learner English: The syntax-pragmatics interface in second language acquisition*. Amsterdam: John Benjamins Publications.

Chamot, A. (2005). The role of learning strategies in second language acquisition. In K. Pitt (Ed.), *Debates in ESOL teaching and learning: Culture, communities and classrooms* (pp. 55-64). Harlow, UK: Longman.

Cohen, A. D. (1986). Mentalistic measures in reading strategy research: Some recent findings. *The ESP Journal*, *5*(2), 131-145.

Cohen, A. D. (1996). Verbal reports as a source of insights into second language learner strategies. *Applied Language Learning*, *7*, 5-24.

Cohen, A. D., & Hosenfeld, C. (1981). Some uses of mentalistic data in second language research. *Language Learning*, *31*, 285-313.

Cohen, A. D., & Upton, T. (2006). *Strategies in responding to the New TOEFL reading tasks* (TOEFL Monograph Series Report No. 33). Princeton, NJ: ETS.

Desimone, L., & Le Floch, K. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis, 26*(1), 1-22.

Ellis, R. (2001). *Form-focused instruction and second language learning*. Oxford, UK: Blackwell.

Ericsson, K., & Simon, H. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: Massachusetts Institute of Technology Press.

ETS. (2008). *TOEFL research—Ensuring test quality*. Princeton, NJ: Author.

ETS. (2009). *The official guide to the TOEFL test*. New York, NY: McGraw-Hill.

ETS. (2011a). *Reliability and comparability of TOEFL iBT scores*. Princeton, NJ: Author.

ETS. (2011b). *Validity evidence supporting the interpretation and use of TOEFL iBT scores*. Princeton, NJ: Author.

Gass, S., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Greene, S., & Higgins, L. (1994). "Once upon a time": The use of retrospective accounts in building theory in composition. In P. Smagorinsky (Ed.), *Speaking about writing* (pp. 115-140). Thousand Oaks, CA: Sage Publications.

Hawley, P. (2003). *Being bright is not enough: The unwritten rules of doctoral study* (2nd ed.). Springfield, IL: Charles C. Thomas Publishers.

Hudson, T., & Park, S. (2002). *Validity issues for selected versus constructed response Internet-based language tests*. Paper presented at the American Association of Applied Linguistics conference, Arlington, Virginia.

Kelle, U. (2001). Sociological explanations between micro and macro and the integration of qualitative and quantitative methods. *Forum: Qualitative Social Research*, *2*(1), 95-117.

Kelly, D. (2009). *Methods for evaluating interactive information retrieval systems with users*. Hanover, MA: Now Publishers.

Konecki, K. (2008). Triangulation and dealing with the realness of qualitative research. *Qualitative Sociology Review*, *4*(3), 7-28.

Leighton, J. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6-15.

Mackey, A., & Gass, S. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacLean, M., & d'Anglejan, A. (1986). Rational cloze and retrospection: Insights into first and second language reading comprehension. *The Canadian Modern Language Review*, *42*(4), 814-826.

Moran-Ellis, J., Alexander, V., Cronin, A., Dickinson, M., Fielding, J., & Sleney, J. et al. (2006). Triangulation and integration: Processes, claims and implications. *Qualitative Research*, *6*(1), 45-59.

Nguyen, T. N. H. (2008). *An investigation into the validity of two EFL (English as a Foreign Language)* *listening tests: IELTS and TOEFL iBT* (PhD thesis). Department of Linguistics and Applied Linguistics, The University of Melbourne.

Nikolov, M. (2006). Test-taking strategies of 12-13-year-old Hungarian learners of EFL: Why whales have migraine. *Language Learning*, *57*(1), 1-51.

Patton, M. (2004). *Qualitative research and evaluation methods*. Thousand Oaks, CA: Sage Publications.

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.

Phakiti, A. (2008). Strategic competence as a fourth-order factor model: A structural equation modeling approach. *Language Assessment Quarterly*, *5*(1), 20-42.

Polio, C., Gass, S., & Chapin, L. (2006). Using stimulated recall to investigate native speaker perceptions in native-nonnative speaker interaction. *Studies in Second Language Acquisition*, *28*(2), 237-267.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Purpura, J. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test takers: A structural equation modeling approach. *Language Testing*, *15*(3), 333-379.

Robinson, M. (1992). Introspective methodology in interlanguage pragmatics research. In G. Kasper (Ed.), *Pragmatics of Japanese as native and target language* (pp. 27-82). Honolulu, HI: University of Hawaii Press.

Ross, S. (1997). Listener inference on a second language test. In G. Kasper, & E. Kellerman (Eds.), *Advances in communication strategies research* (pp. 216-237). Harlow, UK: Longman.

Rubb, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*(4), 414-474.

Sanz, C. (2005). *Mind and context in adult second language acquisition: Methods, theory, and practice*. Washington, DC: Georgetown University Press.

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.

Taylor, K., & Dionne, J.-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, *92*(3), 413-425.

Wijgh, I. F. (1995). A communicative test in analysis: Strategies in reading authentic texts. In A. Cumming, & R. Berwick (Eds.), *Validation in language testing* (pp. 154-170). Clevedon: Multilingual Matters Ltd.

Xu, Y., & Wu, Z. (2012). Test-taking strategies for a high-stakes writing test: An exploratory study of

12 Chinese EFL learners. *Assessing Writing*, *17*(3), 174-190.

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

Yien, L. (2001). Effective test-taking strategies on English tests: Implications from Taiwanese students. *Hong Kong Journal of Applied Linguistics*, *6*(2), 22-43.