ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 2, No. 4, pp. 850-858, July 2011

doi:10.4304/jltr.2.4.850-858

Test Taking Strategies: Implications for Test Validation

Mohammad Salehi
Sharif University of Technology, Iran
Email: salehy@sharif.ir

***Abstract*—To collect pieces of evidence for the construct validity of the reading section of a high-stakes test, test taking strategies of 40 test takers were analyzed via a checklist of strategies. The checklist consisted of 28 strategies tapping test takers' behaviors while taking some reading comprehension items. The goal was to see whether there was concordance between the type of strategies and the item types in the reading comprehension passages. For example, if the strategy of guessing is used on inference items, this jeopardizes the validity of the item because there is a mismatch between the intentions of the test makers and those of test takers (Cohen, 1984). Hopefully, it was found mostly the right strategies were used on the right item types. This speaks to the construct validity of the reading section of University of Tehran English Proficiency Test (the UTEPT) which was exposed to investigation.**

***Index Terms*—test taking strategies, construct validity, item type, frequency of strategy, item difficulty**

I. Introduction

The initial impetus for this study was drawn from the reading of Bachman (1990). In a chapter of the book, entitled "Validation", as a piece of evidence supporting construct validity, Bachman draws on Messick (1988), Cohen (1984) and Grotjahn (1986) to convince the readers that test taking processes provide evidence for construct validity of a test. Raising his dissatisfaction with correlational and experimental approaches, Bachman (1990) maintains, "A more critical limitation to correlational and experimental approaches to construct validation is that these examine only the products of the test taking process-the test scores- and provide no means for investigating the processes of test taking" (p. 269).

Cohen (1998) also points out that test taking strategies can be used for validation purposes "While there is nothing new in pointing out that certain instruments used in SLA research are lacking in validity, it is a relatively new undertaking to use data on test taking strategies to validate such tests" (p. 92).

Cohen's (1984) article is also revealing in that it considers the possibility of validation study by taking into account the fitness of presumptions of test makers and the actual processes in which the testees are involved. He maintains that "The main conclusion in [this study] is that a closer fit should be obtained between how the test constructors intend for their tests to be taken and respondents actually take them" (p. 70).

Based on insights gained from theoretical underpinnings as elaborated by Bachman (1990) and the studies of Cohen (1984), Nevo (1989), Anderson, Bachman, Perkins, and Cohen's (1991), and Storey (1997), among others, the decision was made to carry out this study.

II. The Review of the Related Literature

A. Preliminaries

The first scholar to draw attention to the feasibility of gathering evidence through verbal reports is Cohen (1984). Cohen argues that a mismatch between the intentions of test makers and the thought processes of testees will call into question the validity of a test. In other words, if an inference item is conceived to be a reference one by testees, this is a blow to the validity of a test. Kormos (1998) clarifies the difference between think-alouds, introspection and retrospection. For think alouds, researchers instruct the subjects to verbalize whatever that occurs to them while performing a task. For introspection, the subjects are not only asked to verbalize but also to justify their thought processes. Finally, retrospection is different in that the subjects are supposed to verbalize after they have performed the task. According to Kormos, (1998) the disadvantage is that the subjects need to transfer information from the long term memory to the short term memory which can jeopardize the accuracy of the verbalization.

Camps (2003) maintains that recent studies have shown the usefulness of think-aloud protocols in understanding learners' cognitive processes as they perform tasks designed to help them make form-meaning connections when processing input.

B. Theoretical Underpinnings of Protocol Analysis

Arguably, the best theoretical underpinning was provided by Cohen (1984). But there are also other scholars who are of the idea that protocol analysis or test taking strategies can speak to the validity of a test. One such scholar is Phakiti (2003) who calls for validation research on the relationship of strategic competence and language test performance.

C. Theoretical Basis of Introspective Validation

Perhaps the best theoretical underpinning has been provided by Grotjahn (1986). As a way of setting the stage for the introspective validation, behavioral validation and logical task analysis, he starts by making a critique of the correlational validation of language tests. One reason he brings is that if we correlate a test with another test to see if they show any correlation as a sign of validity assumes that the criterion measure is already valid which may not necessarily be the case. In the words of Grotjahn, "the potential circularity of this approach should be obvious" (p. 161). Another problem is that correlational analyses do not tell us anything about the mental processes while taking tests. The third problem is that "The results of factor analyses depend heavily for instance on the number and type of variables included in a study or on the specific factor analytic technique used" (p. 162).

Grotjahn (ibid) makes the three closely related predictions in terms of the use of introspection methodology:

1- The more analyzed knowledge is, the more assessable it will be via introspection.

2- Foreign or second language knowledge (including skills) is more analyzed than first language knowledge and thus more likely to be accessible via introspection.

3- Learned knowledge is more analyzed than acquired knowledge and thus more likely to be accessible via introspection.

D. Problems with Introspection Validation

Grotjahn (1986) mentions the following problems with the introspective validation. The first one is the controversy surrounding the dubious validity of introspective data. He maintains that the problem can be alleviated in the following ways. One is using logical task analysis. The other one is using supplementary observational data such as "psycho-physiological" reactions such as eye movement. Finally, there are other ways such as communicative and behavioral validation.

Another problem is that introspection methodology produces verbal data that are subject to misinterpretation. Therefore, attempts should be made to correctly and properly interpret the data that come out. The nature of using verbal data is that they "…are reports of something and thus have representational function; their *representational validity* [emphasis original] must therefore also be ensured" (p. 169).

There are two terms that need clarification: communicative validation and behavioral validation. What Grotjahn (ibid) means by communicative validation is that it "attempts to ensure validity by examining to what extent the subject agrees with the interpretation of his or her utterances" (p. 169). The concept of "behavioral validation has been suggested as a means for such an examination in terms of falsification methodology" (p. 169).

E. Limitations of Protocol Analysis

Cohen (1994) mentions the following points as mentioned by critics:

1- Much cognitive processing is inaccessible because it is unconscious.

2- Students may be forced to come up with a verbal response that is not closely related to their actual thought processes.

3- Respondents may provide socially appropriate responses. In other words, their responses may be an edited version of what happens in their thought processes.

4- It is entirely possible that verbal reports may have an "intrusive effect". Their trains of thoughts may be interrupted by the verbalization process and it severely distorts the process.

5- The results obtained may be colored by the verbalization skills of the respondents. This is especially important if the respondents are supposed to do the reports in their L2. Additionally, as Cohen puts it, "respondents may use different terms to describe similar processes or the same terms for different processes" (p. 681).

6- Spoken and written data may be different in nature and may not be compatible.

7- There may also be problems if the testees read in L2 and report in L1. A great deal of information may be lost. The translation of processes into another language may give an inaccurate version of the original and authentic processes.

F. Research in Protocol Analysis

In this section, a few studies which have used protocol analysis will be referred to. One study was done by Anderson et al. (1991) who used a triangulation approach to construct validity. They investigated the relationship between three sources of data: think-aloud protocols, item content and item performance. The chi-square analysis showed that there was a statistically significant association between the reported strategies and the three different item types. Their study provided the incentive for the current study. The authors moved from a product- oriented approach to validation to what is termed a process-oriented approach. They mention a few scholars in the field to set the stage for the study (e.g., Cohen, 1984; Nevo, 1989). The authors contend that so far there have been studies directed at protocol analyses and item analysis/content analysis. But no study has ever brought the three major sources of information together. They claim that theirs is one such undertaking. Twenty eight out of 65 testees were chosen for the study. The entire study

rested on a test of reading comprehension. After a time lapse of a month, the protocols were collected from the testees. A checklist of strategies as constructed by Nevo (1989) was used as the starting point for the analysis. The testees verbalized the strategies they used retrospectively. The strategies they used were assigned to one of the strategies used in the checklist. Additional strategies used by the testees were also added to the list. Out of 47 strategies used, the ones having high frequency were analyzed and the others were excluded. All in all, 17 strategies were studied in depth. The other angle used in the study was item analysis. Both item difficulty and item discrimination were computed. Any item having an index blow .33 was considered to be difficult. They were deemed to be easy if they happened to be above .67. And finally, they were within an acceptable range if they fell within .33 and .67. In terms of content considerations, three basic item types were considered: textually explicit, textually implicit, and finally scriptually implicit. The authors contend that the most important findings concerned the combination of all three sources of information-verbal reports, item characteristics, and question classifications or item types. Strategy 37 "making reference to time", when looked at from these three perspectives, indicates that when students are coping with direct statement-type questions that actually discriminate well between good and poor readers, they do worry about the time allocation. "Stated failure to understand" (strategy 3) was most frequent on items that discriminated among good performers on the test, less frequent on easy items, and on items that simply discriminated between good and poor readers. "Paraphrasing" (strategy 19) was employed particularly with items of medium level of difficulty, and specially with understanding direct statements rather than inferences or the main idea. "Guessing" (strategy 30) occurred especially with inference items and medium difficulty items. And "matching stem with the previous portion of the text" (strategy 34) was used particularly with inference items, but less on the main ideas and direct statements, more often with items of medium difficulty.

McDonough (1995) raises a few questions regarding the previous study. One question is that whether the triangulation study can be carried out with other skills, like listening and writing. Still another question is whether triangulation can be done with other test formats. For example, can triangulation be done with cloze testing or true - false items? Furthermore, he wonders if the study can lend itself to universality. In other words, whether the same strategies, considering the low occurrence of them, are universal strategies or just specific to one testing situation? A third question and a quite pertinent one is what Alderson (1995) is concerned with. This has to do with whether students perform similarly under test and non-test conditions.

Another study is that of Storey (1997). He attempted to see whether protocol analysis could reveal insights into the validity of a cloze test. The cloze test he used was a mini text based on a long passage. There were thirteen blanks in the passage. The validation procedure was an introspective one. In other words, the subjects were asked to think aloud and talk about the strategies they used while answering the questions. The study was in line with Grotjahn (1986) who argues that correlational approaches to validation are necessary, but not sufficient. The blanks were divided into four categories. The first category was related to discourse markers. The second section had to do with anaphoric devices. The third section consisted of lexical substitutes. And finally, the last section had to do with lexical words. The author carried out the study with thirteen subjects. He divided the contents of the passage into distinct parts: discourse markers and cohesive ties. The following conclusions concerning the discourse markers were arrived at:

1- Some testees were involved in "macro" strategies. But some others were involved in "micro" strategies. The distinction between the two has to with the fact that macro strategies operate at inter-sentential level, whereas micro strategies operated at sentence level. A cloze test is supposed to tap beyond sentence level knowledge, but testees' verbalizations shed more light on the fact that the students were treating the passage not as a piece of discourse as they should have done but as isolated sentences. This jeopardizes the validity of the test.

2- It happened that the testees were involved in translating what they thought were ambiguous into terms that could clarify the points.

3- It was observed that some testees did not understand the meanings of certain words.

Concerning the cohesive ties, the following conclusions were arrived at:

1- Most items were treated at a surface level.

2- Students worked with the items without necessarily having to read the passage.

The above points detract from the validity of the test. Still another study dealing with protocol analysis is that of Yi'an (1998). The researcher did the study on listening comprehension. The setting was a Chinese context. An immediate retrospection study used only four subjects to see the effect of test method on test performance. The test method was an interview with six multiple-choice questions. Although the researcher was satisfied with the results as yielded by the retrospection methodology, the quality of the verbal reports hinged on the probing procedures employed. The selection of the subjects was of great importance. The fact that they were picked up from the similar language backgrounds facilitated the job of verbalization and they were free to express whatever their hearts desired. One prime concern of the researcher was whether test method played any role in test performance. If it did, it could lead to the conclusion that test method was a factor in the validity of the measurement process. In fact, it turned out that the test method exerted an influence in two important and fundamental ways. One was that it favored the advanced students and put the less able students at a disadvantage. Another way in which the construct validity was undermined was that the testees chose the correct answers for the wrong reasons. All in all, the researcher was happy with immediate retrospection methodology and called it as "promising" for assessing the essential EFL listening comprehension test-taking processes.

III. Methodology

A. The Participants

The participants in the qualitative part of the study consisted of 40 PhD candidates. They answered 35 reading items working with a checklist of strategies. The participants were selected on the basis of their availability. They were rewarded for their time and effort with some TOEFL books and CDs. Some students did not agree to participate. Some of them did a hasty job. As the participants were PhD students and were about to do a research themselves, they showed good cooperation.

B. The Instrumentation

1. The UTEPT

The test consists of 100 items. The three sections of the test are grammar, vocabulary, and reading comprehension. The grammar section has 35 items. The first 20 items are multiple choice completion items. The second 15 items are error identification; 10 items (items 36 to 45) deal with grammar and vocabulary tested in context. The next section deals with vocabulary. This section is divided into two parts: part one has 10 items (items 46 to 55) and part two has 10 items (items 56 to 65). The last section is concerned with reading comprehension. This section has 35 items consisting of six passages. Table 1 summarizes the three sections and parts of the UTEPT.

TABLE 1.
DIFFERENT SECTIONS OF THE UTEPT, THE METHOD OF TESTING, NUMBER OF ITEMS AND ITEM NUMBERS ON THE TEST

Section	Method of Testing	No. of Items	Item No.
Grammar	Multiple Choice Completion	20	1-20
	Error Identification	15	21-35
	Contextualized	5	36-40
Vocabulary	Multiple Choice Paraphrases	10	46-55
	Multiple Choice Completion	10	56-65
	Contextualized	5	41-45
Reading Comprehension	Six Short Passages	35	66-100

2. A Checklist of Strategies

To facilitate the process of collecting the protocols, it was thought best to use a checklist of strategies as Anderson et al. (1991) suggested. There were a few options (Cohen & Upton, 2006; Nevo, 1989; Phakiti, 2003; Purpura, 1988). It was decided to use Cohen and Upton's (2006) checklist. The checklists provided by Purpura and Phakiti were concerned with metacognitive and cognitive strategies. These two strategy types were not the main concern of the current study. Nevo's checklist was another alternative. That was also ruled out on the grounds that it was simple and not comprehensive. Still a third option available for the researcher was the checklist taken and expanded by Anderson et al. (1991). The strategies were 47 in number and not as comprehensive as the one provided by Cohen and Upton (2006) which was adopted by the researcher. To sum up, the justification for the use of the checklist are as follows: First, it was the most comprehensive one; second, it was the most recent one available; and third and related to the second one it included recent trends in strategy training.

This checklist arrived at by Cohen and Upton (2006) has two sections. The first section covers strategies used in reading. The second part deals with strategies used in test taking.

C. The Procedure

1. Data Collection

Some 10,000 test takers took the UTEPT. After some three months, some 40 candidates were chosen based on their availability. They were asked to take the reading section of the test again, this time working with a checklist of strategies.

2. Data Analysis

The reading section items were exposed to exploratory factor analysis. Eleven factors were extracted as the result of this analysis. I this study these factors are referred to as item types. For more information the interested reader can refer to Salehi (2011).

IV. Results and Discussion

A. Test Taking Strategies across Item Types

Test taking strategies were also compared with different item types to see if a relationship could be found between the use of test taking strategies and item types. Table 2 shows the distribution of test taking strategies and item types. Item types have been placed on the rows. On the columns, test taking strategies have been put:

TABLE 2.
THE DISTRIBUTION OF TEST TAKING STRATEGIES ACROSS ITEM TYPES

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
TT1	0	0	97(2)	68 66 67	0	80	73(2)	0	82 84	77 80	0
TT2	94	100(2) 88(2)	97(2)	67	93	0	0	0	82(2)	75(3) 80(3)	0
TT3	0	100(3)	0	0	0	0	92	70(2)	82(2) 84	80 77(4)	0
TT4	72(2) 90 86(2)	88(3) 89	97(2)	69(2) 66(3) 67(2) 68(2)	93(3) 85(5)	0	92	73(2) 70(3)	84(2) 82	77(2) 75(3) 80(5)	0
TT5	94 90	88(6) 100		67(9) 68(6) 66(5)	93(3)		92(3)	73(2)	84(6)	75(3) 80 77(3)	91
TT6	72(2) 94	88	97	67(4) 69(2) 69(2)		78	92(3)	73(3) 70		75 80	91(2)
TT7	72(3) 86(2) 90	89	97	66(4) 68	93(2) 85		92	73 70	82	77(2) 75(2)	
TT8	72	89(2)	97	69 68(3)	85		82(2)		82	80(2) 75	
TT9									82 84(2)		
TT 10	86(5) 94(3) 72 90			69(4) 68	85(2)		92(4)			80 75	91
TT 11	94	88 89(2)	97	67				70(2) 73	84		
TT 12	94(2)	89 100	97(2)		93	78	92	73	82	80 75	
TT 13	94(2) 90(3) 72(2) 86(2)			69(3) 68	85(3)				82	75	91(3)
TT 14	72(2) 86 90 94	89 100		68(3) 69(2)	93			73(3)	84(2)		91
TT 15											
TT 16		100		68(2)			92	73(2)		75	
TT 17	86	100(4) 89(2)	97		93			73(2)	82		
TT 18	86(2) 94 90	89	97(2)	66(2) 67(2) 68 69(3)	85(2)		92		82(2) 84(2)	77(2) 80 75	91(2)
TT 19		100	97	66(2) 67(3)			92	70	84	77 75(3)	91(2)
TT 20	86(2) 94 72(2) 86		97	69(2)	85		92(3)	70			
TT 21	90 (11) 94(3) 72(7) 86(5)		97	69(6)	85(5)			73		77	91(2)
TT 22	72 94 86	88 100	97	67 69 66	85			73		80 77 75(2)	
TT 23	72 90(2)	100(3)	97	66 68(4)				73		75(2) 80	91(4)
TT 24	86 94(2) 90 72	100 88	97	68(5) 66(3) 67	85 93	78		73(2) 70(3)	84	80	

TT 25	72	89(2)		68 69 67(2)	85				82 84	77 80(3)	
TT 26		100(2)		67	85						
TT 27	90(3) 72 94(2) 86			66(3)	85(3)					77(2)	91
TT 28	90 72 86(2)	89(2) 89(3)	97(3)		85			70(2)		80	

Before going through the interpretation of the table, one point should be made about the difference between reading and test taking strategies. The former has an overall high frequency over the latter. This, it should be mentioned, runs counter to what is found in literature. For example, Cohen and Upton (2006) found that the opposite was true.

B. Interpretation of Table 2

In the case of strategy 1, the highest frequency belongs to items 97 and 73. Both items require a great deal of thought on the part of the test taker and our prediction is that the test taker is going to reread the question. This is the first strategy: *I go back to the question for clarification: reread the question*. The second strategy is as follows: *I go back to the question for clarification: paraphrase (or confirm) the question or task*. Our presumption is that the strategy is going to be used on items that are relatively difficult. One can easily see that the prediction is borne out. The strategy has been used on main idea items.

As for through strategy three which reads as: *I go back to the question for clarification: wrestle with the question intent, the* strategy is probably going to be used on items that are relatively difficult. This is because the testee wants to go back to the passage. Any idea of "going back" is more true of items that place a little of a challenge on testees. It has the highest frequency on item 77 and the next goes to item 100. Item 77, it is to be recalled, was grouped with items 75 and 80, and they all dealt with inferencing and topic identification all of which are liable to present challenges for testees. Now, lets' turn to item 100 to see if the prediction made can become true. This item is an item that is based on topic organization, an endeavor which would necessitate going through the passage again to gain more insight into it. In a nutshell, the more frequent use of this strategy on items that are of inference and topic organization type lends support to the construct validity of this test.

The next strategy to be analyzed is strategy 4: *I read the question and consider the options before going back to the passage /portion*. One prediction is in order: this strategy is going to be used on items that require knowledge that might be independent of the texts and passages. This is because the testee considers the options before going through the text. Probably the testee has a chance of getting the item right even without having to read the passage. The highest frequencies go to items 80 and 85. Both items have equal frequency of five. A high frequency on item 85 is quite justifiable on the grounds that the item is a vocabulary one and virtually answerable without any resort to passages. But what is surprising and not easily interpretable is the test taking behavior of the participants on item 80. This item requires the identification of the primary purpose of the passage which cannot be tackled without reading or even rereading the passages.

The next strategy to be discussed is strategy 5: *I read the question and then read the passage/portion to look for clues to the answer, either before or while considering options.* From what the strategy says, one can easily link it to directly answered questions. Let's see if that prediction is borne out. One can see that the prediction was borne out. The highest frequency belonged to item 67. This item is one of those items which can be answered directly. This strategy has also a high frequency of use on other items as well. These are items 84 and 88. The use of the strategy with these items is surprising because these items are not directly answerable.

Now, it is time to analyze strategy 6: *I predict or produce my own answer after reading the portion of the text referred to by the question.* The strategy has been more frequently used on item 67.

The next strategy to be elaborated on is strategy 7: *I predict or produce my own answer after reading the question and then look at the options (before returning to text).*

Probably, reflecting momentarily before going through the options puts one on high alert. Again, the highest frequency belongs to one of the items which is directly answerable. Strategy 8 does not lend itself to discussion because of low frequency of use and on the grounds that the strategy is more appropriate for iBT TOEFL.

The next strategy is strategy 9: *I consider the options and identify an option with an unknown vocabulary*. This strategy was employed for item type nine which include items 82 and 84.

What follows next is strategy 10: *I consider the options and check the vocabulary option in context*. The highest frequency goes to item 86 which is obviously a vocabulary item. Actually, the strategy was employed ten times for item type 1 which is a vocabulary item. The testees might have been led to the use of this strategy more frequently with item type 1 because of the simple association of "vocabulary" in the strategy and most vocabulary items in the passages.

The next strategy is strategy 11: *I consider the options and focus on a familiar option.* The strategy is relatively infrequent, with only 10 items occurring on item types. It will not be discussed for the same reason. Strategy 12 is not going to be discussed on the grounds of low frequency. But strategy 13 will be discussed below:

I consider the options and define the vocabulary option.

Building on previous trends, the expectation was that the testees would use the strategy with vocabulary items. As a matter of fact, the expectation was met. It was used 9 times on item type 1.

Strategy 14 will be discussed next: *I consider the options and paraphrase the meaning.* The presumption is that the strategy is going to be linked to items which were challenging for the test takers. With that thing in mind, let's see what actually happened. The highest frequencies belong to items 68 and 73. Item 68 is not that difficult. But item 73 is more difficult than item 68. The facility indices for the two items are .61 and .44, respectively. Apart from the difficulty level of items, the two items are also different in terms of item types. In fine, the grouping of the two items in terms of strategy use is surprising.

The next strategy is strategy 15: *I consider the options and drag and consider the new sentence in context.* This strategy is specific to iBT TOEFL. The researcher deliberately included this one to test testees' integrity in the responding process. As was predicted, no one chose this strategy because items corresponding to this strategy were simply nonexistent.

Strategy 16 reads as: *I consider the options and postpone consideration of the option.* This strategy was the one most testees had a problem with. They did not know what the strategy meant in the first place. One can rightfully wonder whether that could have any effect on the task they were doing. As a matter of fact, one can realize that the problem with the literal understating of the strategy has taken its toll: the strategy demonstrated a low frequency of use which precludes any discussion of it as well.

What comes next comes is strategy 17: *I consider the options and wrestle with the option meaning*. The researcher was walking in the room trouble shooting the points the testees had problems with. The literal understating of the strategy placed an enormous challenge on the testees. What has happened leads one to draw, at least, two conclusions: one is that testees had not figured out the literal meaning of the strategy and so had not used it with many items. The other is that they had properly understood the meaning of the strategy and had used it as such. Both conclusions are sound.

In the case of strategy 18: *I make an educated guess (e.g., using background knowledge or extra-textual knowledge*), before discussing the strategy one thing is certain: the strategy is not obviously going to be used on inference or topic identification items. If this happens, then the construct validity of the test can be under question. That did not happen. As a matter of fact, the strategy occurred more frequently on two different groups of items. One group is vocabulary items. The other group is the items which are directly answerable. The use of the strategy with both groups of items speaks to the construct validity of the test.

As far as strategy 19 is concerned: *I reconsider or double-check the response,* it is a typical strategy among Iranian students. The best prediction is that the strategy will have a high frequency. The prediction was not borne out. One interesting point is that it was not used on vocabulary items except item 91 which was a separate vocabulary and hard to interpret factor. The best guess was that the strategy might be used on items which are of inference type.

What follows next is strategy 20: *I look at the vocabulary item and locate the item in context.* One obvious point that springs to mind is that the strategy is vocabulary specific and use of it with other item types is simply unthinkable. As a matter of fact, it is easy to spot six instances of the use of the strategy with vocabulary items.

The next strategy to be discussed is strategy 21: *I select options through background knowledge.* The mere superficial similarity of this strategy and strategy 18 might drive certain individuals to use the two strategies on identical or similar items. As a matter of fact, the testees performed as predicted. There are twenty six instances of occurrence of this strategy on item type one.

So far as strategy 22 is concerned: *I select options through vocabulary, sentence, paragraph, or passage overall meaning (depending on item type),* some testees told the present researcher that the strategy is a reiteration of the previous ones, and there was nothing novel about the strategy. One interesting point about the strategy is that it was used only once on various items except item 75.

What comes next is strategy 23: *I select options through elimination of other option(s) as unreasonable based on background knowledge*.

There are two points that should be made about the strategy. First of all, it repeats one thing that it did before, namely the use of background knowledge. The other point is that it mentions the use of elimination strategy. The highest frequency belongs to items 68 and 91 with 9 occurrences on each. Item 68 is a directly answerable item and item 91 is a vocabulary item. In both cases, the use of the strategy is justifiable.

The next strategy to be discussed is strategy 24: *I select options through elimination of other option(s) as unreasonable based on paragraph/overall passage meaning.* The highest frequency has gone to item 68. This is a directly answerable one. As a matter of fact, other items in the item type have likewise a good frequency of this strategy use.

The next strategy to be discussed is strategy 25: *I select options through elimination of other option(s) as similar or overlapping and not as comprehensive.*

Some testees did find the strategy irritating as it contained the word "eliminate". One can see if that is obvious in their performances as well. No distinct pattern emerged. It has the highest frequency on item 80 which is not interpretable as this item would not conveniently lend itself to the elimination procedure.

Turning to strategy 26: *I select options through their discourse structure*, the best prediction is that the strategy will be frequent among items which are related to topic organization, references, etc. Of course, the testees had a little problem with the meaning of the phrase "discourse structure" which was explained to them. Disappointingly, the strategy was reported to have been employed only four times. This can be attributed to the literal understating of the strategy itself despite the fact that it was explained to the testees.

As for strategy 27: *I discard option(s) based on background knowledge*, the testees might have viewed the strategy as being no different from strategy 23. The strategy has a high frequency on item type 1 which is a vocabulary factor and in which options can be easily eliminated.

Finally, the last strategy, i.e. 28: *I discard option(s) based on vocabulary, sentence, paragraph, or passage overall meaning as well as discourse structure*, has the highest frequency on item type 2. More frequency was expected from other items.

V. Conclusions

It was thought best to use a checklist of strategies as opposed to introspection methods. One good thing about the checklist was it that it ensured objectivity as it was structured. On the downside, it limited the strategies. There might have a myriad of strategies employed by test takers which could be possibly captured by the checklist. This could have been alleviated by asking the test takers to come up with additional strategies if they could. So, this is a delimitation of the study.

All in all, different strategies employed by the test takers as revealed via a check list of strategies spoke to the validity of the test in question. What this boils down to is that the right type of strategies were used for the right type of item types which were determined by an exploratory factor analysis on the items. There were strategies which were used infrequently. Henceforth, they were discussed for the same reason. Frequency was used as yardstick to draw conclusions about the validity issues. In other words, the more frequent the proper strategies were used on the right type of item types, the more valid our inferences were. Another point worthy of note was that the difficulty level of items was used as another indicator of strategy use. In other words, strategies of guessing or any kind of non-contributory strategies (Nevo, 1989) cannot be validly used for items that are difficult in nature. Hopefully, it was seen that it was the case that contributory strategies were used for items of high caliber in terms of item difficulty.

References

[1] Anderson, N., 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, 41-66.*

[2] Bachman, L. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press.

[3] Camps, J. (2003). Concurrent and retrospective verbal reports as tools to better understand the role of attention in second language tasks. *IJAL, 13(2),* 201-221.

[4] Cohen, A. (1984). On taking tests: what the students report. *Language Testing*, I, 70-81.

[5] Cohen, A. (1994). Verbal reports on learning strategies. *TESOL Quarterly, 28*, 678-682.

[6] Cohen, A. (1998). Strategies and processes in test taking and SLA. In L. Bachman & A. Cohen. (Eds.), *Interfaces between SLA and language testing research* (pp. 90-111). Cambridge: CUP.

[7] Cohen, A. (In press). The coming of age of research on test-taking strategies. In J. Fox, D. Bayliss, L. Cheng, C. Tuner, & M. Wesche (Eds.), *Language testing reconsidered: Selected papers from LTRC 2005*. Ottawa: Ottawa University Press.

[8] Cohen, A., & Upton, T. (2006). Strategies in responding to the new TOEFL reading tasks. Monograph Series: ETS.

[9] Cohen, A., & Upton, T. (2007). "I want to go back to the text": Response strategies on the reading subset of the new TOEFL. *Language Testing,* 24, 209-250.

[10] Grotjahn, R. (1986). Test validation and cognitive psychology: Some methodological considerations. *Language Testing, 3*, 159-185.

[11] Kasper, G. (1998). Analyzing verbal protocols. *TESOL Quarterly, 32*, 358-362.

[12] Kormos, J. (1998). The use of verbal reports in L2 research. *TESOL Quarterly, 32*, 353-358.

[13] McDonough, H. (1995). Strategy and skill in learning a foreign language. London: Edward Arnold.

[14] Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer. & H. Braun (Eds.), *Test Validity* (pp. 33-45). Hillsdale, NJ: Erbaum.

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

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

[17] 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, 333-379.

[18] Salehi, M. (2011). On the factor structure of a reading comprehension test. *English Language Teaching*. Vol 4, No 2.

[19] Sasaki, T. (2003). Recipient orientation in verbal report protocols: Methodological issues in concurrent think-aloud. *Second Language Studies, 22* (1), 1-54.

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

[21] 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*, 21-44.

Mohammad Salehi is currently a faculty member of the Languages & Linguistics center at Sharif University of Technology and holds a PhD degree in TEFL from the University of Tehran. He got his MA degree from the University of Allameh Tabatabai in TEFL. He earned a BA degree in English literature from Shiraz University. He has taught in University of Kashan, University of Teacher Training, University of Amirkabir, Azad University of Karaj, University of Tehran, and University of applied sciences. He has presented articles in Turkey, Cyprus, Dubai, Armenia, Australia and Iran. His research interests include language testing and second language acquisition research. He has also written books on language testing and vocabulary.