4 Strategies and processes in test taking and SLA

Andrew D. Cohen

A process approach to test taking

One of first requests that more attention be paid to the processes of respondents in giving answers on language tests was issued by Bormuth (1970):

There are no studies, known to the author, which have attempted to analyze the strategies students use to derive correct answers to a class of items. The custom has been to accept the test author's claims about what underlying processes were tested by an item. And, since there were no operational methods for defining classes of items, it was not scientifically very useful to present empirical challenges to the test author's claims. (p. 72)

Bormuth's book outlines the objectives and major components of a theory for writing items for achievement tests, drawing on structural linguistics, semantics, and logic. Subsequently, studies began to appear that entailed observation and description of how learners at different age levels actually accomplish first language (L1) testing tasks. For example, with respect to a teacher's oral questioning of young children, it has been suggested that "the interrogator and respondent work together to jointly compose the 'social fact' we call an answer-to-a-question" (Mehan 1974: 44). On the basis of his research efforts, Mehans indicated that it may be misguided to conclude "that a wrong answer is due to a lack of understanding, for the answer may come from an alternative, equally valid interpretation."

Since the late 1970s, interest has slowly begun to grow in approaching second language (L2) testing from the point of view of the strategies used by respondents going through the process of taking the test (e.g., Cohen & Aphek 1979; Homburg & Spaan 1981; Cohen 1984; MacLean & d'Anglejan 1986; Gordon 1987; Anderson 1989; Nevo 1989). By the 1990s, L2 testing textbooks acknowledged this concern as a possible source of insights concerning test reliability and validity (Bachman 1990; Cohen 1994a).

I wish to acknowledge Lyle F. Bachman, Elaine Tarone, and an anonymous reviewer to Cambridge University Press for their helpful comments in the revision of this chapter. Some sections of this chapter are based on Cohen (1994a).

Tests that are relied upon to indicate the comprehension level of readers may produce misleading results because of numerous test-wise techniques that readers have developed for obtaining correct answers on such tests without fully or even partially understanding the text. As Fransson (1984) put it, respondents may not proceed via the text but rather around it. In effect, then, there are presumptions held by text constructors and administrators as to what is being tested, and there are the actual processes that test takers go through to produce answers to questions and tasks. The two may not necessarily be one and the same. Students may get an item wrong for the right reasons or right for the wrong reasons. Discovering that a respondent used poor logic in attempting to answer a reading comprehension item may be of little interest to the text constructors if the problem resides solely with the respondent. However, if the poor logic was precipitated by an overly ambiguous text passage or by an ambiguous question, then the test constructor may wish to edit the text or revise the question.

Even if the problem resides exclusively with the test taker, a concerned test developer and test administrator may wish to have more information about the items that provide such illogical responses or about the test-taking strategies that result in incorrect answers. Respondents may be consistently using certain strategies that are detrimental to their performance on certain types of items or on an entire test. For example, respondents may plod laboriously through a text only to find that once they reach the multiple choice questions, they have forgotten most of what they read or have failed to focus adequately on those elements being tested. In such a case, the strategy of studying the questions carefully before reading the text might have been more beneficial.

The intent of this chapter is to describe test-taking strategy data emerging from studies of respondents taking different kinds of tests, and to discuss the role these data can and do play in second language acquisition (SLA) research. Traditionally, the difference has been relatively clear-cut between language tasks intended for SLA research purposes and language tests constructed for assessing language achievement. An SLA measure (e.g., a communication task such as relaying directions from a map) is not intended for gate-keeping purposes. In fact, such tasks may purposely encourage risk taking by putting the respondents in a situation where they do not have the vocabulary or other language forms needed to complete the task, so that researchers can determine the strategies that they use. The respondents are usually in a low anxiety situation because their performance would not normally have implications for their lives.

In a traditional achievement testing situation, the testers usually check for control of language that the respondents have been taught. The respondents know that there is a premium put on better performance. Whereas in SLA research tasks the respondents can get points for communication
in spite of inaccuracies, and their performance does not usually affect their grade in a language course, in instructional achievement testing the respondents must perform accurately, often under time constraints, and are held accountable for their responses.  

In recent years, the distinction between assessment for the purposes of SLA research versus assessment for instructional purposes has lessened somewhat. More and more tasks and tests are being used interchangeably. As language tests have become a greater part of SLA research, there has been a growing concern for the reliability and validity of such measures. 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 in order to validate such tests. This chapter mentions efforts in this direction, drawing on insights from both L1 and L2 testing.  

Let us start by looking at the purpose for considering the processes involved in test taking. The main purpose is to determine the effects of test input upon the test taker – specifically, the processes that the test taker makes use of in order to produce acceptable answers to questions and tasks. There is a concomitant concern to determine the respondent’s perceptions about tests before, during, and after test taking.  

What is meant by test-taking strategies? 

Language use strategies are mental operations or processes that learners consciously select when accomplishing language tasks. These strategies also constitute test-taking strategies when they are applied to tasks in language tests. For the purposes of this discussion, test-taking strategies will be viewed as those test-taking processes that the respondents have selected and of which they are conscious, at least to some degree. In other words, the notion of strategy implies an element of selection. Otherwise, the processes would not be considered strategies.  

Strategies vary according to context. One strategy is to opt out of the language task at hand (e.g., through a surface matching of identical information in the passage and in one of the response choices). Another is to use shortcuts to arrive at answers (e.g., not reading the text as instructed but simply looking immediately for the answers to the given reading comprehension questions). In such cases, the respondents may be using test-wisenedness to circumvent the need to tap their actual language knowledge or lack of it, consistent with Fransson’s (1984) assertion that respondents may not proceed via the text but rather around it. In some cases, quite the contrary holds true. One Hebrew second language respondent in a study of test-taking strategies in Israel determined that he had to produce a written translation of a text before he could respond to questions dealing with that text (Cohen & Aphek 1979).  

At times, the use of a limited number of strategies in a response to an item may indicate genuine control over the item, assuming that these strategies are well-chosen and are used effectively. At other times, true control requires the use of a host of strategies. It is best not to assume that any test-taking strategy is a good or a poor choice for a given task. That evaluation depends on how individual test takers – with their particular cognitive style profile and degree of cognitive flexibility, their language knowledge, and their repertoire of test-taking strategies – employ the strategies at a given moment on a given task. Some respondents may get by with using a limited number of strategies, and using them well for the most part. Others may be aware of an extensive number of strategies but may use few, if any of them, effectively. So, for example, while a particular skimming strategy (such as paying attention to subheadings) may provide adequate preparation for one test taker on a recall task, the same strategy may not work well on a different text that lacks reader-friendly subheadings. In addition, the strategy, while successful for one respondent, may not work well for another respondent.  

The ability of learners to use language strategies has been referred to as their strategic competence – a component of communicative language use (Canale & Swain 1980). This model puts the emphasis on “compensatory” strategies – that is, strategies used to compensate for or remediate a lack in some language area. Bachman (1990) provided a broader theoretical model for viewing strategic competence. Bachman and Palmer (1996) refined the Bachman (1990) categories somewhat. Their current framework includes an assessment component, whereby the respondents (in the case of language testing) assess which communicative goals are achievable and what linguistic resources are needed; a goal-setting component, wherein the respondents identify the specific tasks to be performed; and a planning component, whereby the respondents retrieve the relevant items from their language knowledge and formulate a plan for their use in a response.  

Despite the theoretical shift away from a primary focus on compensatory strategies, it may still be the case that a fair number of test-taking strategies are, in fact, compensatory. When put on the spot, respondents often omit material because they do not know it, or produce different material from what they would like with the hope that it will be acceptable in the given context. They may use lexical avoidance, simplification, or approximation when the exact word escapes them under the pressure of the test situation or because they simply do not know the word well or at all.  

Thus, in keeping with the Bachman and Palmer model, when respondents are given a situation in which to perform an oral role play, they
may assess the situation and identify the information that is needed in that context. They may also set their general goals. They may plan their specific response and go about retrieving from their language knowledge the grammatical, discourse, and sociocultural features needed for the role play. Then at some point they execute the role play. After they have finished, they may again perform an assessment to evaluate the extent to which the communicative goal was achieved.

As is the case with any theoretical model, test takers may make differential use of the components of this model when performing specific testing tasks. Hence, there are respondents who might not assess the situation before starting the role play. This approach may be successful, or it may lead to the violation of certain sociocultural conventions. For example, a respondent in a Japanese L2 role play may neglect to take into account the older age and higher status of the interlocutor, and may select language forms that are not adequately respectful. In addition, there are respondents who may set general goals for an utterance or string of utterances in the Japanese L2 role play without making a detailed plan of their utterances before producing them. Again, this may well work, or it may lead to ineffective utterances that lack grammatical fine tuning.

By the same token, role-play respondents may also plan specifics without having general goals in mind. In such cases, the respondents may produce one or more L2 utterances that have been carefully monitored for grammatical accuracy but that do not fit into the overall discourse and, hence, come across as incoherent. There may be still other respondents who begin talking without first determining either general goals or a detailed plan. Indeed, the same respondent may assume one or another of these response patterns at different moments during a given test-taking situation and/or in different test-taking situations.

Recent research involving the use of verbal report directly after the performance of oral role-play interaction is just beginning to yield results regarding the extent of assessment and planning that actually takes place before the execution of speech acts, such as apologies, complaints, and requests (Cohen & Olshtain 1993). Clearly, more such work is needed in order to understand how respondents arrive at utterances in complex speech situations.

The use of verbal report in identifying test-taking strategies

The field of research methods has supplied us with suggested approaches for looking at test taking. One means is through observation of what respondents do during tests. Another means is through designing items that are assumed to require the use of certain strategies (e.g., cloze items calling for anaphoric reference) and adding up the correct responses as an indicator of strategy use. A third approach is through the use of verbal report, while the items are being answered, immediately afterward, or some time later on.

Verbal report techniques have constituted a major tool in the gathering of data on test-taking strategies. Such techniques were initially developed in first language and then second language acquisition research in order to study the processes of reading and writing. Verbal reports have helped determine how respondents actually take tests of various kinds (Nicker-son 1989; Norris 1989). Moreover, innovative research on test taking has helped to refine the research methodology for tapping such test-taking strategies. For example, studies have found that it is possible to collect introspective and retrospective data from students just after they have answered each item on a multiple choice reading test. Several such studies are described in this section. Earlier work reported on approaches that involved at most a request of respondents to reflect back on the strategies they used in arriving at answers to a subtest or group of items, producing data of more questionable reliability and validity (Cohen 1984).

Just as there is a keen interest in using verbal report methods to improve the reliability and validity of assessment instruments, there needs to be an ongoing concern for assuring the reliability and validity of the very verbal report methods that are being used to collect test-taking strategy data. They cannot be immune from such scrutiny. A recent article by the author deals extensively with the issue of improving the reliability and validity of verbal reports (Cohen, in press), especially concerns about the appropriate use of such measures and the nature of reports that include the findings from the use of such measures. The issues discussed include the effects of the immediacy of the verbal report, the potential value of having the respondents themselves interpret the data, the benefits of prompting for specifics in verbal report, the advantages of providing guidance in how to produce verbal report, and the effects of verbal report on task performance.

The article concludes by focusing on what needs to be included in the write-ups so that others will understand fully what was done, enabling comparisons with other studies and replication of the study. The following areas are addressed: (1) the subjects’ characteristics, (2) the characteristics of the materials, (3) the nature of the criterion task, (4) the extent of guidance in verbal reporting, (5) the methods of analysis used, (6) the

1 Discussions of reliability and validity of verbal reports can be found in Ericsson and Simon (1984) and, with respect to SLA research, in Grootjahn (1987) and Haastrop (1987). Extensive discussions of reliability and validity in qualitative research can be found in Kirk and Miller (1946), Miles and Huberman (1994), and Denzin and Lincoln (1994).
categories used to score verbal report protocols, (7) the results of inter-rater reliability checks, (8) the criteria for selecting verbal report excerpts for inclusion in research reports, and (9) any theories used in framing the verbal report study. Care in the write-up can help to dispel arguments that such methodological approaches are not adequately rigorous. Since the use of verbal report techniques is becoming more prevalent in investigating test-taking strategies, there is a need to provide greater systematicity both in the collection of such data and in the reporting of such studies in the research literature.

Because verbal report techniques can be said to have emanated in part from literacy studies within the field of SLA — that is, studies eliciting strategies from writers as to how they plan, generate, and revise their text — it can be said that SLA research methods have provided the testing field with a means for validating not only non-testing tasks but testing tasks as well. By the same token, experts in the field of testing in turn contribute to SLA research by reminding researchers of the need to validate their research instruments.

Strategies for taking tests of reading and writing

In considering strategies on tests of reading and writing skills, I focus first on two more indirect testing formats, multiple choice and cloze, and on strategies for three more direct formats, namely, summarization tasks, open-ended questions, and compositions. The chapter closes with several suggestions that may lead to more effective test taking. Unless it is specified that the activity constituted a research task, the tasks and tests for which the respondents reported strategy use contributed to their course grades in language classes.²

Indirect testing formats

Indirect formats for testing — in other words, those formats that do not reflect real-world tasks — may prompt the use of strategies solely for the purpose of coping with the test format. Let us look at two such formats, multiple choice and cloze, and at some of the research findings regarding strategies used in taking such tests.

MULTIPLE CHOICE TEST FORMATS

Investigating a standardized test of English L1 reading (the Cooperative Primary Test, Form 12 A) by sitting down with individual first grade learners and going over each item separately after the testing session, MacKay (1974) found that learners did not necessarily link the stem and the answer in the same way that the test constructor assumed was correct. MacKay determined that the test had a somewhat arbitrary frame of reference and did not provide concrete information about how children reasoned. For example, pictures were sometimes ambiguous. In an item requiring the student to link the expression “The bird built his own house” to a picture, a student chose a nest of twigs with eggs in the middle over a wooden birdhouse because he claimed that some big birds could not fit in the hole of the birdhouse. MacKay noted that the student chose the right picture for the wrong reason. The student missed the element that people, not birds, are responsible for building wooden birdhouses with perches.

A test constructor receiving such feedback might wish to alter the item a bit, requiring, for example, that the respondents give their rationale for choosing one alternative over another. Naturally, there would be constraints on such an approach — the natural constraints of group testing, the age and ability of the respondents to provide this extra information, and so on. It would also be interesting to know whether this is just one isolated anecdote or whether this faulty logic is shared by other respondents as well.

MacKay also gave an example of an item missed for the wrong reasons. The statement “The cat has been out in the rain again” had to be linked to one of three pictures, which looked roughly like those in Figure 4.1. The student perceived the dotted wallpaper as snow and decided that this picture was of the exterior of the house. Thus, he gave the dripping raincoat as the correct answer. Once the child had perceived the wallpaper as snow and thus had eliminated the third picture, his selection of the first picture, the dripping raincoat, rather than the second, was perfectly reasonable — even though cats do not wear raincoats. If this kind of logic was shared by numerous respondents, it might suggest the need to improve the pictures in order to eliminate any ambiguities in them.

² I make this point because it is possible that respondents to a research task would use different strategies from those used when the results "count" toward a grade.
Haney and Scott (1987) found patterns similar to those reported by MacKay with regard to the sometimes dubious fit between elementary school children getting the L1 language arts item right or wrong and their verbal report as to whether they had applied the skill meant to be tested. For example, they found unusual and perceptive interpretations of questions. The interpretations resulted in the wrong answer in the item “Which needs least water?” (followed by pictures of a cactus, a potted plant, and a cabbage). The respondent answered “cabbage” because it had been picked and therefore needed no water at all, whereas the expected correct answer was “cactus.” The interpretation resulted in the right answer in an item asking why Eva likes to watch TV. The respondent reported personalizing the item and responding according to why he liked to watch TV, in this case producing the correct answer without relating to Eva at all.

With respect to older respondents, the patterns are relatively similar, as in a study of 40 college ESL respondents that used retrospective verbal report to gain insights about test-taking strategies (Larson 1981, in Cohen 1984). As a research task, the students were requested to describe how they arrived at answers to a 10-item multiple choice test based on a 400-word reading passage. This formed part of the midterm for the advanced ESL course. Seventeen students met with the author of the test in groups of two or three within 24 hours after the test, while 23 students met in groups of five or six 4 days after taking the test. The investigator found that the respondents used the following strategies: (1) they stopped reading alternatives when they got to the one that seemed correct to them, (2) they matched material from the passage with material in the item stem and in the alternatives (e.g., when the answer was in the same sentence with the material used to write the stem), and (3) they preferred a surface-structure reading of the test items to one that called for more in-depth reading and inferencing (Larson 1981, in Cohen 1984). This superficial matching would sometimes result in the right answer. One example was as follows:

5) The fact that there is only one university in Filanthropia might be used to show why...
   a) education is compulsory through age 13.
   b) many people work in the fishing industry.
   c) 20 percent of the population is illiterate.
   d) the people are relatively happy and peaceful.

Students were able to identify (c) as the correct answer by noticing that this information appeared earlier in the same sentence with the information that reappeared in the item stem:

The investigating travel agency researchers discovered that the illiteracy rate of the people is 20 percent, which is perhaps reflective of the fact that there

is only one university in Filanthropia, and that education is compulsory, or required, only through age 10.

They assumed that this was the correct answer without understanding the item or the word illiterate. They were right.

In another example, students did not have to look in the text for surface matches. They were able to match directly between the stem and the correct alternative:

2) The increased foreign awareness of Filanthropia has...
   a) resulted in its relative poverty.
   b) led to a tourist bureau investigation.
   c) created the main population centers.
   d) caused its extreme isolation.

Students associated foreign in the stem with tourist in option (b), without understanding the test item.

It was also found that more reasoned analysis of the alternatives – for example, making calculated inferences about vocabulary items – would lead to incorrect answers, as in the following item:

4) The most highly developed industry in Filanthropia is...
   a) oil.
   b) fishing.
   c) timber.
   d) none of the above.

This item referred to the following portion of the text:

... most [dollars] are earned in the fishing industry. ... In spite of the fact that there are resources other than fish, such as timber in the forest of the foothills, agriculture on the upland plateaus, and, of course, oil, these latter are highly underdeveloped.

One student read the stem phrase “most highly developed industry” and reasoned that this meant “technologically developed” and so referred to the “oil industry.” He was relying on expectations based on general knowledge rather than on a careful reading of the text. His was a reasoned guess rather than a surface match, as in the previous example.

It needs to be stressed that the Larson study was a student course project and therefore limited in scope. If the test were to be used widely, then it would be advisable to conduct a series of such investigations to determine both the reliability and the validity of the verbal report techniques used to elicit the information, as well as the generalizability of the findings across a range of different students.

In an effort to investigate the extent to which multiple choice questions are answered on the basis of prior knowledge of the topic and general vocabulary knowledge, 32 intermediate and 25 advanced Israeli EFL students were given a research task the title and the first paragraph of
a passage appearing on the previous year’s exemption examination. They were then asked to answer 12 questions dealing with the portion of text not provided. Two weeks later they were given the text in full along with the questions, and once again were asked to respond (Israel 1982, in Cohen 1984). The rate of success on the multiple choice items was surprisingly high the first time – 49 percent for the advanced group and 41 percent for the intermediates. These results were far better than the 25 percent success rate that would be expected on the basis of chance alone.3 When the students were given the test with the complete passage and questions two weeks later, the advanced group scored 77 percent and the intermediates 62 percent. The score necessary for exemption from further EFL study was 60 percent. The fact that the average performance on the test was low even when the passage was provided makes the results without the passage that much more striking.

In a research study with 30 tenth grade EFL students – 15 high proficiency readers and 15 low proficiency readers – respondents were asked to verbalize thoughts while finding answers to open-ended and multiple choice questions (Gordon 1987). The researcher found that answers to test questions did not necessarily reflect comprehension of the text. Both types of reading comprehension questions were regarded by the respondents as “mini” reading comprehension tests. With respect to test-taking strategies, the low proficiency students tended to process information at the local (sentence/word) level, not relating isolated bits of information to the whole text. They used individual word-centered strategies like matching words in alternatives to text, copying words from the text, translating word for word, or formulating global impressions of text content on the basis of key words or isolated lexical items in text or test questions. The high proficiency students, on the other hand, were seen to comprehend the text at a global level – predicting information accurately in context and using lexical and structural knowledge to cope with linguistic difficulties.

In an effort to provide immediate verbal report data, Nevo (1989) designed a testing format that would allow for immediate feedback after each item. She developed a response strategy checklist, based on the test-taking strategies that have been described in the literature and on her intuitions about strategies respondents were likely to select. A pilot study had shown that it was difficult to obtain useful feedback on an item-by-item basis without a checklist to jog the respondents’ memory as to possible strategies.

Nevo’s checklist included fifteen strategies, each appearing with a brief description and a label meant to promote rapid processing of the check-

list. As a research task, she administered a multiple choice reading comprehension test in Hebrew L1 and French as a foreign language to 42 tenth graders, and requested that they indicate for each of the ten questions on each test the strategy that was most instrumental and second most instrumental in arriving at their answer. The responses were kept anonymous so as to encourage the students to report exactly what they did, rather than what they thought they were supposed to do.

It was found that students were able to record the two strategies that were most instrumental in obtaining each answer. The study indicated that respondents transferred test-taking strategies from their first language to the foreign language. The researcher also identified whether the selected strategies aided in choosing the correct answer. The selection of strategies that did not promote choice of the correct answer was more prevalent in the foreign language than in the first language test. The main finding in this study was that it was possible to obtain feedback from respondents on their strategy use after each item on a test if a checklist was provided for quick labeling of the processing strategies utilized. Furthermore, the respondents reported benefiting greatly from the opportunity to become aware of how they took reading tests, having been largely unaware of their strategies prior to this study.

Another study of test-taking strategies among nonnatives (Anderson et al. 1991) revealed that respondents used certain strategies differently, depending on the type of question that was being asked. For example, the strategies of “trying to match the stem with the text” and “guessing” were reported more frequently for inference questions than for direct statement and main idea question types. The strategy of “paraphrasing” was reported to occur more in responding to direct statement items than with inference and main idea question types.

Anderson et al.’s study originated as a doctoral dissertation in which 28 native speakers of Spanish studying at an intensive ESL language school in Austin, Texas, took as research tasks three measures of reading comprehension: a reading comprehension subtest from a test of language skills, a measure of ability to read college-level textbook prose (Textbook Reading Profile, Segal 1986), and a second form of the standardized reading comprehension test (Anderson 1989). After the first two tasks, the participants provided retrospective think-aloud protocols describing the strategies they used while reading the textbook material and answering the comprehension questions. The respondents also provided think-aloud protocols along with the final test. The data were categorized into a list of 47 processing strategies.

In the follow-up phase of the research, data from the participants’ retrospective think-aloud protocols of their reading and test-taking strategies were combined with data from a content analysis and an item analysis to obtain a truly convergent measure of test validation (Anderson et

3 These results are also consistent with those for native English readers, where the results were far better than chance (Tuinman 1973–4; Fowler & Kroll 1978).
The content analysis of the reading comprehension passages and questions was comprised of the test designer's analysis and one based on an outside taxonomy, and the item performance data included item difficulty and discrimination. This study marked perhaps the first time that both think-aloud protocols and more commonly used types of information on test content and test performance were combined in the same study in order to examine the validation of the test in a convergent manner.

Another recent study corroborated earlier test-taking strategy research in finding that examinees focused on the search for answers to test questions. The English L1 respondents in the study paid little attention to strategies that provided an overall understanding of the native language passage (Farr, Pritchard, & Smitten 1990: 223). The investigators concluded that a reading comprehension test is a special kind of reading task, in which skilled examinees contemplate answer choices, use background knowledge, weigh choices, skim and reread portions of the reading selection, and refrain from making choices until they feel confident about an answer. They suggested that the types of questions following the passage will determine whether the reading focuses only on the surface meaning of the text.

Emerging from these various studies on multiple choice tests of reading comprehension is a series of strategies that respondents may utilize at one point or another in order to arrive at answers to the test questions. Whether these strategies are of benefit depends to a large extent upon when they are used and how effectively they are used. Table 4.1 presents a composite list of some of the more salient test-taking strategies appearing in one or more of the studies mentioned here. There is probably not a single strategy on this list that has not been written up in the testing literature somewhere; the innovation is to pinpoint who uses such strategies, when they use them, why they use them, and (ideally) their relative success at using them. This last issue is perhaps the most difficult to ascertain.

In recent years some attention has been focused on explicit training for ESL respondents in test-wiseness. For example, Allan (1992) developed a test of test-wiseness that included stem option cues, in which it was possible to match information from the stem with information in the correct option; grammatical cues, where only one alternative matched the stem grammatically; similar option cues, where several distractors could be eliminated because they essentially said the same thing; and item giveaway, where another item already gave away the information. In preliminary validation work, Allan tested three groups of students (N = 51), having one group write a brief explanation of how they selected their answers. Even though the items were meant to be content-free, it turned out that prior knowledge and guessing were still possible. The reliabilities for

<table>
<thead>
<tr>
<th>Table 4.1. Strategies for taking a multiple choice reading comprehension test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Read the text passage first and make a mental note of where different kinds of information are located.</td>
</tr>
<tr>
<td>2. Read the questions a second time for clarification.</td>
</tr>
<tr>
<td>3. Return to the text passage to look for the answer.</td>
</tr>
<tr>
<td>4. Find the portion of the text that the question refers to and then look for clues to the answer.</td>
</tr>
<tr>
<td>5. Look for answers to questions in chronological order in the text.</td>
</tr>
<tr>
<td>6. Read the questions first so that the reading of the text is directed at finding answers to those questions.</td>
</tr>
<tr>
<td>7. Try to produce your own answer to the question before you look at the options that are provided in the test.</td>
</tr>
<tr>
<td>8. Use the process of elimination – i.e., select a choice not because you are sure that it is the correct answer, but because the other choices do not seem reasonable, because they seem similar or overlapping, or because their meaning is not clear to you.</td>
</tr>
<tr>
<td>9. Choose an option that seems to deviate from the others, is special, is different, or conspicuous.</td>
</tr>
<tr>
<td>10. Select a choice that is longer/shorter than the others.</td>
</tr>
<tr>
<td>11. Take advantage of clues appearing in other items in order to answer the item under consideration.</td>
</tr>
<tr>
<td>12. Take into consideration the position of the option among the choices (first, second, etc.).</td>
</tr>
<tr>
<td>13. Select the option because it appears to have a word or phrase from the passage in it – possibly a key word.</td>
</tr>
<tr>
<td>14. Select the option because it has a word or phrase that also appears in the question.</td>
</tr>
<tr>
<td>15. Postpone dealing with an item or selecting a given option until later.</td>
</tr>
<tr>
<td>16. Make an educated guess – e.g., use background knowledge or extratextual knowledge in making the guess.</td>
</tr>
<tr>
<td>17. Budget your time wisely on this test.</td>
</tr>
<tr>
<td>18. Change your responses as appropriate – e.g., you may discover new clues in another item.</td>
</tr>
</tbody>
</table>

**Note:** The order in which the strategies are listed does not constitute a hierarchy of importance, nor does it reflect the frequency with which the given strategies have been found to occur.

the stem option and similar option items were low, suggesting that these cues were only sometimes recognized by respondents or that only some of the items in the subscales were measuring those test-wise phenomena.

**CLOZE TEST FORMATS**

Research regarding strategies for taking cloze tests is of interest in that it has helped to determine whether such tests actually measure global reading skills, as they are commonly purported to do. As more studies have
been undertaken on the cloze test, it has become clearer that the instrument elicits more local, word-level reading than it does macro- or discourse-level reading (Klein-Braley 1981; Alderson 1983; Lado 1986), contrary to the claims of its early supporters, who have maintained that cloze assesses global reading (see, e.g., Chihara et al. 1977; Chávez-Oller et al. 1985). In the following excerpt from a cloze passage, items 1 and 3 can be answered on a microlevel, whereas item 2 would call for macro- or discourse-level comprehension:

People today are quite astonished by the rapid improvements in medicine. Doctors (1) ________ becoming more specialized, and (2) ________ drugs are appearing on (3) ________ market daily. (based on Cohen 1994a: 234)

That is, the local context provides the answer to (1), “are.” The answer to (2), “new,” depends on an understanding of the opening sentence, so this would be a discourse-level response. Once item (2) is filled in, item (3) is simply a local response based on the immediate context of that phrase. More proficient readers are more skilled at correctly completing cloze items such as (2), which assess discourse-level reading, whether reading in the native or in a foreign language (Bachman 1985).

Studies on strategies for taking cloze tests have shown that perhaps only a quarter of nonnative respondents read the entire EFL cloze passage before responding (Emanuel 1982 and Hashkes & Koffman 1982, in Cohen 1984). A case study shed some light on the issue of reading the text before completing an L1 cloze test (Kleiman et al. 1986). Verbal protocol data provided by a seventh grade Brazilian girl filling in two cloze passages—one as a warm-up and the other as an exercise in Portuguese L1—indicated that the respondent was preoccupied with local clues from isolated elements of text. What emerged was that she did not use global clues until she had completed a substantial number of blanks on the cloze. In other words, it is easier to read the cloze passage once it has been partially completed and the respondent has some idea of what it is about, much as a child may have an easier time of connecting numbered dots once the picture that the dots are forming becomes clearer.

One of the early studies of strategy use in completing a cloze passage involved indirect assessment of strategies used. The researchers administered a rational deletion cloze with 23 blanks to 39 EFL subjects from three levels (Homburg & Spaan 1981). One of four strategies was hypothesized to be necessary in finding a correct word for each of the blanks: recognition of parallelism, sentence-bound reading, forward reading, or backward reading. Success at items calling for “forward reading” (cataphora) was significantly associated with success at understanding the main idea. In verbal protocol studies, it was found that nearly 20 percent of the respondents did not use the preceding or following sentence for clues to blanks but rather guessed on the basis of the immedi-ate context (Emanuel 1982 and Hashkes & Koffman 1982, in Cohen 1984).

The picture regarding the taking of cloze tests in a native language does not appear to be much different. One study, for example, involved 18 fifth graders at three levels of reading (high, intermediate, low) who were given a rational deletion cloze test in Hebrew L1 and were asked to think aloud as they completed it (Kesar 1990). An analysis of the verbal report protocols yielded at least 26 different strategies, which were grouped into seven categories: word level/part of sentence; sentence level; and five categories at the level of discourse-intersentential, whole-text level, extratextual level, metacognitive level, and “other.” The results demonstrated that although the better readers were more likely to use macro-level schemata and strategies in completing the cloze and also did better on the completion task as a whole, all respondents favored micro-level reading at the sentence level.

Thus, the research on strategies in taking cloze tests would suggest that such tests assess local-level reading more than they measure global reading ability. Furthermore, such tests are more likely to test for local-level reading when they are in a foreign language (see, e.g., MacLean & d’Anglejan 1986).

More direct formats

SUMMARIZATION TASKS

Whereas more direct formats for testing, such as text summarization, are less likely than indirect formats to elicit test-taking strategies that take the place of genuine language use strategies, responses to such measures are still influenced by test-wisieness. As long as the task is part of a test, students are likely to use strategies that they would not use under non-test conditions. In the case of a summary task, the respondent is invariably summarizing a text for a reader who already has a notion of what the summary should look like; therefore, the respondent is reacting to a set of perceived and real expectations on the part of the reader. In the real world, we usually summarize a text for our own future use or for the benefit of someone who has not read it, in which case the set of expectations of the summarizer may be quite different.

Case study research concerning the strategies used in producing summaries in Portuguese L1 on a test of EFL texts has suggested that respondents might use various shortcut measures (Cohen 1994b). One strategy was to summarize by lifting material directly from the passage rather than restating it at a higher level of abstraction or generality, in the hope that the raters’ ambiguity over whether the respondent understood the material would work in the respondent’s favor. Furthermore,
when respondents were in doubt about whether to include or exclude material, they might be prone to include it (as in the case of one less proficient student), assuming that a longer summary would probably be preferred by the raters to one that is too terse. The case study found that the respondents spent more time on their strategies for reading the texts to be summarized than they did on the production of their summaries, so — not surprisingly — the summaries were not particularly coherent or polished.

**Open-ended Questions and Compositions**

Like summarization tasks, open-ended tasks allow respondents to copy material directly from a text in their response, so that raters cannot tell whether the respondent in fact understands the material. Such copying may produce linguistically awkward responses. For example, in a study of 19 college-level learners of Hebrew L2 that involved a retrospective verbal report one week after the learners took their final exam, it was found that students lifted material intact from an item stimulus or from a text passage for use in their written responses (Cohen & Aphek 1979). Results included verb forms incorrectly inflected for person, number, gender, or tense; verbs reflecting the correct root but an incorrect conjugation; and so forth. Various strategies were observed for producing a verb form when the rules for production had not been learned. If students did not know the correct verb form, they would use the infinitive, take a form from a tense that they knew, take one inflectional ending and generalize it across person and gender, take an inappropriate tense from the stimulus and simply add the prefix for person, and so on.

Another strategy that the learners in the Cohen and Aphek (1979) study used in their writing was to introduce prepackaged, unanalyzed material and combine it with analyzed forms. For example, given that Hebrew prepositions like mi (“from”) can be prefixed to the object of the preposition through elision (mi + tsad “side” = mitsad), one student learned this form as one word and then affixed another preposition to it on an exam: *bemitsad “on from a side” intending “on a side.”

The interpretation of essay topics is a problem that is related to inadequate attention to instructions. Usually, an essay topic is presented in the form of a minitext that the respondent needs to understand and operationalize. Ruth and Murphy (1984) note cases where English L1 students misinterpreted words in the prompt, such as confusing the words *profit* and *prophet*, thus shifting the meaning of the topic entirely. Perhaps of greater consequence are the strategies the respondents have to evaluate the nature of the task. Ruth and Murphy give the example of a supposedly friendly letter topic wherein what the raters of the “letter” will actually value is a response at a higher level of formality than might be reflected in an authentic friendly letter. The guideline, then, for respondents in writing tasks is to be especially careful in interpreting the genuine intention of instructions for completing testing tasks.

**Conclusion**

This chapter has examined process approaches to language testing, which have usually entailed the use of verbal report techniques to better understand these processes and the test-taking strategies that respondents use. There are not yet abundant data linking the specific use of test-taking strategies with success or failure on language tests. On the other hand, the use of qualitative methodologies such as verbal report provides a valuable source of information — perhaps the most focused possible — on the strategies respondents used in their responses and why they did so. Verbal report can help us see what items are actually testing, aiding us in making decisions about which items to keep and which to throw out. One could go so far as to say that it is now close to essential to have verbal report as part of pretesting/piloting. If a respondent has legitimate reasons for marking an item wrong, then the item needs to be rewritten.

The emphasis in this chapter has been on the relationship between the characteristics of the test: task and the strategies used, and on the connection between testing methods and SLA research — especially on the types of investigations needed to provide information as to which testing methods would be potentially more or less reliable and valid for SLA research. For example, we already know that the multiple choice format poses problems. It has come under strong criticism in recent years, as pointed out in Hughes (1989: 60–62), who includes the following in his list of shortcomings:

1. Items meant to assess just grammar may also test for lexical knowledge as well.

---

4 Studies are beginning to appear that link the selection of a given strategy with rated success at performance in a foreign language. A study by Cohen, Weaver, and Li (1996), on intermediate college learners of French and Norwegian, for example, identified a link between the frequency of use of a given strategy and performance on the speaking task for which that strategy was chosen. It was found that an increase in certain preparatory strategies (e.g., translating specific words, writing out sentences, and practicing the pronunciation of words) and monitoring strategies (e.g., monitoring for grammar, paying attention to the pronunciation of words, and analyzing a story for its key elements) related to an increase in ratings by outside raters on one or more of the rating scales — self-confidence, grammar, vocabulary, and identifying and ordering elements in a story. A recent study by Purpura (1996), discussed by Bachman and Cohen in this volume (Chapter 1), provides an example of a quantitative approach to investigating the relationships among cognitive and metacognitive strategies reported by test takers and their scores on an English L2 test.
2. Distractors may be eliminated as being absurd.
3. Correct responses may be common knowledge.

It would appear that the test-taking strategy research can be utilized to substantiate or refute claims about multiple choice items, at least with respect to a given test in a given testing situation with given respondents. In other words, criticisms of the multiple choice format are not new and did not originate with qualitative investigations. However, qualitative investigations can help us to move beyond superficial pronouncements to determine how decisions are actually made. So, for example, a study could be designed whereby students need to indicate through retrospective verbal reports the process whereby they arrived at answers to multiple choice grammar items. The intention would be to determine if it was actually grammatical knowledge that was being tapped in each item or whether the deciding element was, for example, control of one or another vocabulary word in the stem or in the distractors.

The results of test-taking strategy studies on cloze tests would also appear to provide crucial information regarding what those tests actually measure. The various types of cloze tests have been subjected to careful scrutiny in recent years, and, of the studies carried out, those that deal with response strategies are perhaps among some of the most insightful. Thus, while the reliability of a given cloze test may be high because the individual items are interrelated, the validity as a measure of global reading ability could be questioned if the respondents indicate that they answered most of the items by means of local micro-level strategies.

It would appear that the nature of test-taking strategies with respect to the more open-ended formats, such as summarization, open-ended responses to questions, and essays, has yet to be fully investigated. Since the assessment of summaries and essays depends on judgments made by raters, there is a concomitant need for research on strategies used in doing the ratings, such as the work conducted by Hamp-Lyons (1989), Vaughan (1991), Connor and Carrell (1993), and Cushing Weigle (1994).

Given the results from test-taking strategy research, however embryonic they may be at present, SLA researchers would probably want to consider validating the testing measures that they use through triangulation, which would include the collection of test-taking strategy data on subsamples of respondents, as in the Anderson et al. (1991) study. Even though the field of test-taking strategy research is a fledgling one, and even though these techniques are still in need of refinement, researchers can find useful descriptions in the literature of techniques for identifying the strategies used by respondents. Consideration of the findings from this growing research area will undoubtedly prove beneficial at all points in constructing, administering, and interpreting language tests.

References

tional deletion cloze. Master's thesis (in Hebrew), School of Education, Hebrew University of Jerusalem, Jerusalem, Israel.


