On taking language tests:  
what the students report

Andrew D. Cohen School of Education, The Hebrew University of Jerusalem

This paper discusses methods for obtaining verbal report data on second-language test-taking strategies; reports on types of findings obtained in a number of studies conducted by university students on how learners take reading tests, in particular cloze and multiple-choice tests; and finally considers the implications of the findings for prospective test takers and test constructors. The main conclusion is that a closer fit should be obtained between how test constructors intend for their tests to be taken and how respondents actually take them.

There is a small but growing literature on how students go through the process of taking language tests — i.e. the steps that they take to arrive at answers to questions. Such research has generally focused on the testing of native language skills, but it has come to encompass second-language testing as well. The purpose of such research has been to explore the closeness-of-fit between the tester’s presumptions about what is being tested and the actual processes that the test taker goes through. The findings have been revealing, both for what has been learned about weaknesses in tests and for what has been learned about successful and unsuccessful test-taking strategies.

Regarding weaknesses in tests, for example, it has been found — through looking closely at the performance of native third-grade test takers — that the test passages such students are requested to read may be seriously flawed (Fillmore, 1981). Such passages have been referred to as a new genre for English written language — with unnatural requirements of lexical choice, grammatical structuring and syntactic alterations to test particular vocabulary items. Regarding the strategies of test takers, it has been found that poor native and non-native readers contort test material in both a semantically unnatural and a grammatically awkward way to fit their schemata or cultural background knowledge (Fillmore and Kay, 1980).

Research has also revealed ‘mainline’ vs ‘fragmented’ reading of test questions. Whereas in mainline reading the respondent quickly gets the main idea and uses this as a point of departure, in fragmented
reading the reader refers primarily to words in the immediate vicinity of the question or to some very strong concept in the text (Dollerup *et al.*, 1981). Likewise, poor performers on cloze tests have been found to be those who among other things do not use ‘forward reading’ (i.e. utilizing the context following the blank) to find clues for supplying the missing word (Homburg and Spaan, 1981).

Furthermore, respondents have been found to do surface matching between information in the text, in the item stems, and in the multiple-choice alternatives — without processing any of these stimuli for meaning. Respondents have also been found to perform tasks by analogy to previous tasks without noticing what may be slight changes in the response procedures (Fillmore and Kay, 1980; Hosenfeld, 1976; Cohen and Aphek, 1979).

The outcome of these various strategies may be that respondents get items wrong for the right reasons or right for the wrong reasons. In other words, a good student may produce a wrong answer as a result of an adventurous inference, while a weak student may choose the appropriate alternative despite faulty reasoning (Mehan, 1974; MacKay, 1974; Dollerup *et al.*, 1981). Thus, due to flaws in the test or due to certain test-taking strategies, students may not be displaying a representative performance of their language competence.

The principal purpose of this paper is to discuss methods for obtaining verbal report data on second-language test-taking strategies, to report on some of the types of findings obtained, and to look at the implications of the findings for prospective test takers and test constructors. The findings will be drawn from a series of unpublished studies conducted by university students — studies dealing with how language learners take reading tests, especially of the cloze and multiple-choice varieties.

### I Methods for obtaining verbal data

We will first describe some of the types of verbal data that can be collected, and then will describe some ways that these data can be collected.

#### 1 Types of verbal data

There are at least two types of verbal data available — those regarding the process of responding and those regarding reactions to items and subtests.

**a The process of responding:** First, there are data concerned with the process of responding to test items and procedures — i.e. how the respondents went about it. For example, did they read the instructions
and pay attention to special procedures associated with the given testing format? For example, if they were requested to first read a passage (with or without deletions) from beginning to end, did they do it or did they read only parts of it and/or jump around? Similarly, if respondents were requested to read all multiple-choice alternatives before answering a given item, did they, in fact, do this?

In their effort to understand passages and items, did the respondents translate as they read (either in their heads or in writing)? How did they go about answering questions? Did they do so systematically — eliminating unreasonable choices? For example, did they guess in a calculated way based on inferences or did they guess impulsively? In filling in cloze items, did they make use of the context preceding and following each blank in choosing an appropriate word? If the students were allowed to use a dictionary while completing the test, did they actually use it and if so, how? Finally, did they copy from another respondent’s test paper?

Reactions to items and subtests: Respondents can report on the ease or difficulty with which they read a given passage, answered particular questions, or performed certain tasks like taking a dictation, giving a short speech, and the like. They can also report on their attitudes toward a certain testing format, such as the cloze test, or toward a certain type of item — e.g. one that calls for inference as opposed to one which is based on information readily available in the text.

2 How the verbal data are obtained

How the data are obtained is a major issue regarding all research in this area of verbal report. First, do we want to obtain ‘think-aloud’ data (simply letting the thoughts flow verbally without trying to control, direct or observe them) or self-observational data (where there is limited or more extensive analysis of thoughts and abstraction about those thoughts)? Self-observational data constitute either introspection (if they relate to thoughts within 10-20 seconds), immediate retrospection (within minutes of the thoughts) or delayed retrospection (after an hour, a day, a week) (Cohen and Hosenfeld, 1981). (Ericsson and Simon (1980) have a somewhat similar breakdown between ‘think-aloud’ data and data reflecting what they call ‘concurrent’ and ‘retrospective verbalizations’.) Sometimes we want a spontaneous, unanalysed report, while at other times we want the learners’ reflective observations on what they did.

Secondly, do respondents write their replies or are they given orally — possibly to be audio or video-taped? If respondents write down their thoughts, then some material is invariably lost in the writing process. They may not express in writing what they would have said.
However, by having respondents write down their thoughts, it is possible to obtain data from large groups of respondents at the same time. Oral responses generally call for small-group or individual sessions, unless the respondents are seated, say, in booths in a language lab and asked to record their replies.

Thirdly, do we obtain the data while the test is actually going on or at some time after the testing session has ended? For example, it is possible to intersperse, within a test, questions that test language performance with questions that call for verbal report about the test itself. Such questions could appear at the end of each subtest. A problem in collecting data this way is the interventionist effect of the verbal report questions. Yet a problem in waiting until after the test is over to ask questions is that some of the most interesting data may no longer be recoverable in that the respondents may have forgotten the processes that they used and the reasoning behind what they did. It is also possible to have the test itself constitute a measure of test-taking strategies. For example, the successful completion of a test could indicate that a respondent was utilizing a given set of strategies. A case in point would be giving respondents multiple-choice questions without giving them the passage upon which the questions are based (an approach to be reported on below). Then calculated guessing could be assessed.

II Some of the data obtained

The following are illustrative of possible data that can be obtained using procedures such as those outlined above. The findings are all drawn from student course papers. Two of these studies looked specifically at strategies in taking cloze tests — one involving 25 Israeli ninth-grade EFL students taking a regular and a discourse cloze (Emanuel, 1982) and another one involving 22 Israeli twelfth-grade EFL students and four native English speakers taking a regular cloze (Hashkes and Koffman, 1982). In these two studies, students were asked questions about how they took the cloze tests — immediately following the tests themselves. In another study, 45 Hebrew University EFL students were given a test comprising a reading passage with multiple-choice and open-ended questions, and a cloze test constructed from a summary of the passage (Roizen, 1982). In this study, the students were asked to react in their native language to each subtest of the test as they completed it.

The final two studies investigated the taking of multiple-choice reading comprehension tests. In one of the studies, 40 ESL students at the University of California, Los Angeles, were asked to describe how they arrived at answers to a 10-item multiple-choice test based on a 400-word reading passage (Larson, 1981). Seventeen signed up
in groups of two or three to meet with the author of the test within twenty four hours after the test. Twenty three students met in groups of five or six led by an interviewer four days after taking the test. In the other study, 32 intermediate and 25 advanced Israeli EFL students were given a title and just the first paragraph of a passage appearing on the previous year's exemption examination, and then were 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).

1 The process of responding

a The cloze: Although the instructions on the cloze test request that the respondent read the entire passage through before answering, only up to one quarter of the respondents in the cloze studies indicated doing this (Emanuel, 1982; Hashkes and Koffman, 1982). The purpose of this request was to have the students benefit as much as possible from the surrounding context. Moreover, as many as 16 per cent of the students indicated not using the context of the preceding or following sentences to find clues for filling in the blank at hand. The bulk of the students did this 'part of the time', and about a third to a half as many did this 'all the time'. When taking the discourse cloze, students reported using context more than with the regular cloze — as would be expected, since the discourse cloze involves pronominal reference, lexical substitution and the like. The majority (64 per cent) of the respondents said that they were most likely to look for a clue to the answer in the same sentence containing the deletion (Hashkes and Koffman, 1982).

When students were asked what they did when they did not know how to fill in a blank, poor students indicated that they just left the space blank — that they were reluctant to guess. The better students reported that they were more likely to guess. Most of these guesses (82 per cent) were based on the immediate or microcontext (Hashkes and Koffman, 1982). Another strategy that respondents used was that of translation. Twenty seven per cent reported translating parts of the text as they went along, while 18 per cent said they regularly translated the immediate sentence of the blank. The researcher found that those who reported translating while doing the cloze also got poorer scores (Hashkes and Koffman, 1982).

The investigators were also interested in whether successful performance on the cloze indicated that the student had understood the passage. Accordingly, students were requested to give a summary of the passage in their native language. There was found to be a low correlation ($r = .24$) between these summaries and performance on
the cloze (Hashkes and Koffman, 1982). This finding seems to be in keeping with some of the recent literature (Alderson, 1979; Klein-Braley, 1981) suggesting that the 'classical' cloze test is more of a measure of word and sentence-level reading ability than of discourse-level reading, as was originally claimed.

The Hashkes and Koffman study also investigated whether native students filled in blanks on the cloze test in a way different from non-natives. They found that the natives used the context as much as possible to find clues to missing words, more so than most of the non-natives. This often involved rereading the sentences several times. This strategy was used extensively by the one native who correctly supplied the exact word for all 20 deletions.

b Multiple-choice reading comprehension tests: Whereas the instructions asked students to read the passage before answering the questions, students reported either reading the questions first or reading just part of the article and then looking for the corresponding questions (Larson, 1981; Roizen, 1982). Respondents were also advised to read all the alternatives before choosing one as the correct answer. Larson’s study found that students would stop reading the alternatives as soon as they found one that they decided was correct. Larson also found that students would use a strategy of matching material from the passage with material in the item stem and in the alternatives. They preferred this surface-structure reading of the test items to one that called for more in-depth reading and inferencing. It was found that this superficial matching would sometimes result in the right answer.

One example involves the following item from the Larson study:

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 per cent 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 which reappeared in the item stem:

   . . . The investigating travel agency researchers discovered that the illiteracy rate of the people is 20 per cent, 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 centres.
   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 — e.g. making calculated inferences about vocabulary items — would lead to incorrect answers. The following item provided an example of this:

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 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 forests 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. The point is that his was a reasoned guess, not that of, say, surface matching, as in the previous example.

In another study, students reported being able to answer questions correctly on the basis of their prior knowledge of the topic and their general vocabulary knowledge (Roizen, 1982). In fact, this is what the student in the above example about developed industries was trying to do. Moreover, in the study where the students were purposely not given the passage to work with (Israel, 1982), the rate of success on the multiple-choice items was still surprisingly high — 49 per cent for the advanced group and 41 per cent for the intermediates. These results are far better than the 25 per cent success rate that would be expected on the basis of chance alone.1 When the students were given the test with the complete passage and questions two weeks later, the advanced group now scored 77 per cent and the intermediates 62 per cent. The score necessary for exemption from further EFL study was 60 per cent. 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.

c Other processing issues: Roizen (1982) found 17 per cent of her sample reported cheating. This was on a test that was anonymous, not counting toward their grades. Perhaps this is why cheating went on, or perhaps it simply reflected that cheating was a regularly used strategy for a small subsample of the group. On another issue, since students were allowed to use dictionaries on that same test, Roizen asked students whether they actually used their dictionaries (Hebrew—English) and if so, how often. She found that three quarters needed to refer to the dictionary to get through the passage, about one quarter looked up 10 words or more, and some looked up more than 20 words. What might such extensive use of the dictionary suggest about the respondent's test-taking strategies? It is unlikely that excessive use of the dictionary is a substitute for contextual guessing, and most likely an unproductive strategy, when used extensively.

2 Reactions to item types

Sometimes respondents were relatively uniform in their reactions to item types. For example, they generally felt that it was harder to answer multiple-choice items based on reading passages than open-ended questions (Roizen, 1982). They also did not like items that involved inferencing or in which the original text was disguised through the use of paraphrase (Larson, 1981). In other cases, the students were more or less split in how they reacted to items or sub-tests. For example, 60 per cent of the respondents in the Roizen study did not like the cloze. They said that it made them nervous, that it required too much patience and concentration and that it did not help them comprehend the text. The other 40 per cent liked it because it required thinking and logic. A recent study by Madsen (1982) also found that the cloze was a high anxiety-producing test.

Another interesting split in attitudes was over items testing for anaphoric reference. Some respondents (55 per cent) felt that these items helped with their overall comprehension of the passage because the items made them return to certain words and structures, causing them to link sentences and ideas. Others (45 per cent) found that looking for specific points distracted them from the larger context. They felt that they could only understand such questions if they understood the entire text, and that if they understood the entire text, then such questions were a waste of time. They felt that questions like 'What does the 'it' in line 7 refer to?' were only testing grammar, not reading comprehension (Roizen 1982).
III Discussion

There are some basic problems with self-report approaches to understanding the test-taking process. One problem is that the most authentic data are obtained for tests which really count. But on such tests respondents may be unwilling to supply honest information out of the fear that this information might adversely affect their grade. And if the tests are not for credit (as in the case of all but the Larson study above), then perhaps we are not obtaining a true picture of how a student takes a test when it does count (David Nevo, personal communication). There is a further problem that some of the information is already at a level of generalization removed from the actual completion of the item or task. In other words, if I really wanted accurate information on how a student fills in blanks on a cloze test, I would need the respondent to report on how each blank was filled in, immediately after that blank was filled in. One way to obtain some information on the answering of multiple-choice items is to have the respondent indicate why alternative responses were rejected. But this still would not necessarily get at all the processes leading to the selection of a multiple-choice alternative.

Despite the reservations associated with this line of investigation, there do appear to be some implications of the work to date. One message to test takers would be to read all instructions very carefully. The implicit message here is that somehow by doing exactly what the instructions call for, the respondents will do the best job on the task at hand. Thus, the message is also directed at the test constructors — namely, to make sure that the instructions reflect the most efficient approach to taking the particular test. For example, perhaps respondents should be encouraged to read the multiple-choice questions before reading a test passage. As a teaching strategy, we often ask questions before having students read a passage in order to motivate the students in their reading. Respondents could also be given several questions after a short portion of text as a way of dealing with this issue. With respect to the issue of students not reading all the alternatives, it would be possible to have the students indicate their rationale for rejecting the alternatives that they do not select (see Cohen, 1980: 24–26).

It would also appear that students need to be taught how to take a cloze test. A research study some years ago found that practice sessions in how to complete cloze tests did not produce significantly greater improvement among those receiving the sessions than among control-group students (Kirn, 1972). Klein-Braley (personal communication) suggests that the lack of significant improvement was due to substantial differences in the behaviour of the individual cloze tests. She found that the reliability and validity of nth-word deletion
cloze tests vary greatly from one test to the next (Klein-Braley, 1981). Assuming we obtained reliable and valid cloze tests, perhaps we could then train people to take such tests more successfully. For example, students could be offered techniques on how to read a passage that has a series of blanks in it. Special attention could be given to the use of preceding and following context — and not just within the sentence containing the deletion. Respondents could be shown how to guess contextually, using all the available clues. Perhaps such sessions could now be made effective enough to produce significant results.

In summary then, this paper has demonstrated ways that we can obtain verbal report data on second-language test-taking strategies, and has illustrated some of the types of data obtainable. The main conclusion seems to be that there is value in striving for a closer fit between how test constructors intend for their tests to be taken and how respondents actually take them. This may involve changing the format of the test or training the respondents to deal with that format more effectively. An avenue for further research would be to explore the effects of training in taking different types of language tests.

**Acknowledgements**
This is a revised version of a paper presented at the Second Meeting of the Israeli Academic Committee for Research on Language Testing (ACROLT), Kiryat Anavim, 4–5 January 1983. My thanks to David Nevo, Bernard Spolsky, Chris Klein-Braley and Joan Abarbanel for their comments on the original version.

**IV  Notes**

1. A study conducted with 1200 native elementary-school readers of English found that the probability of correct response on four-choice items from five standardized tests without the passage ranged from .32 to .50 (Tuinman, 1973—74). Another study with 101 second-quarter college freshmen and 17 English department faculty members found that freshmen got 52 per cent correct and faculty 75 per cent correct on five-choice items from a scholastic aptitude test without the passage (Fowler and Kroll, 1978). Thus, the results for the non-natives are consistent with results for natives.

2. Klein-Braley (1980) indicated success at training students to take cloze tests.

**V  References**


Cohen, A.D. and Aphek, E. 1979: Easifying second language learning. A research report under the auspices of Brandeis University and submitted to the Jacob Hiatt Institute, Jerusalem, ERIC ED 163 753.


Klein-Braley, C. 1980: The assessment of reading comprehension skills. English Department, University of Duisburg.


