DOCUMENT RESUME

ED 281 380 FL 016 657

AUTHOR Cohen, Andrew D.
TITLE Testing Linguistic and Communicative Proficiency: The Case of Reading Comprehension.
PUB DATE Apr 87
NOTE 43p.; Paper prepared for R. L. Politzer Festschrift, "Perspectives on Second Language Teaching," H. B. Altman, Ed.
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Communicative Competence (Languages); Computer Assisted Testing; *Language Proficiency; *Language Tests; Linguistic Competence; *Reading Comprehension; Reading Strategies; *Test Format; Testing Problems; Test Use; Test Validity; Test Wiseness

ABSTRACT

Current issues in the literature on the testing of linguistic and communicative proficiency are reviewed and discussed in relation to reading comprehension testing. Several theoretical issues in language testing are discussed, including testing purposes and test validity. Areas of concern regarding methods of testing reading comprehension are then considered. These include the types of reading to be tested, tapping comprehension at different levels of meaning, comprehension skills, and testing methods (language of response, cloze and C-tests, communicative tests, computer-adaptive testing). The discussion concludes with a look at strategies used by test-takers in dealing with reading comprehension tests, and their implications for test development. A seven-page reference list is appended. (MSE)

Testing Linguistic and Communicative Proficiency: The Case of Reading Comprehension(1)

Andrew D. Cohen

School of Education

Hebrew University of Jerusalem

April 1987

(1) Paper prepared for R.L. Politzer Festschrift edited by H.B. Altman et al., Perspectives on Second Language Teaching. I would like to express my gratitude to Graham Low for his careful reading of this paper, and to Michael Scott for some well-placed comments.

This chapter will provide a brief survey of some of the current issues in the literature on testing linguistic and communicative proficiency, and will narrow the field somewhat by focusing on the testing of reading comprehension as a case in point.(2) We will start by discussing several theoretical issues in language testing, will then consider several areas of concern regarding methods of testing reading comprehension, and will conclude with a look at strategies of test takers in dealing with reading comprehension tests.

Theoretical Issues in Testing

1. Purposes for Testing

It has been demonstrated that tests can be used for administrative, instructional, or research purposes (Jacobs et al. 1981). In fact, the same test of reading comprehension could conceivably be used for twelve different purposes, five administrative purposes -- assessment, placement, exemption, certification, promotion; four instructional purposes -- diagnosis, evidence of progress, feedback to the respondent, evaluation of teaching or curriculum; and three research purposes -- evaluation, experimentation, knowledge about language learning and language use.

---

(2) I owe my expertise in language testing in no small part to Robert Politzer, for it was he who encouraged me to become an evaluator of a bilingual education program, which in turn gave me field experience in psychometrics, which afforded me the credibility which led to offers for work which enabled me to get even more experience.

Given the traditional ways of designing tests of reading comprehension, the average test is not intended to be used for more than several purposes, and the major split is often between proficiency tests intended for administrative purposes and achievement tests for assessment of instructional results.

Current innovations in testing, however, would suggest that the same test could possibly merge these two different sets of purposes under cetain circumstances, i.e., if assumptions of design and use are met. In other words, it is being suggested that tests used to differentiate people according to general level of ability and tests used for certifying the attainment of content be combined in one test (Henning 1985). The suggested means for achieving this merger is through item response theory (specifically, the Rasch model), wherein a latent "acquisition" continuum is inferred both for testing tasks and for the ability level of the respondents. In that both respondents' ability and item or task difficulty are positioned along the same latent continuum, it is thus considered possible to make inferences from examinee performance that are referenced to the performances of other individuals or to the standards imposed by other tasks. It is argued that by merging proficiency and achievement tests in this way, placement can be more in line with what is taught, passing from one level of instruction to the next can be contingent on actual learning, and the curriculum can be more sensitive to individual differences of students at every level.

It is important to piont out that this suggested merger is only possible if a number of assumptions about test design are met. It assumes that the Rasch model provides a "good fit" for the item. In testing language competence, this is problematic since the Rasch model requires the items to be constructed along a single dimension and yet there is usually a degree of multidimensionality in language tests since language competence is not a unitary skill, but rather involves different types of skills (see Wodd and Baker 1985). Other assumptions that are disputed include the claims that the item bank will retain stable properties over a long period and that it is possible to gradually add items without retesting all those in the bank. Woods and Barker add that Rasch provides a "sample-free" estimate of item difficulty only if the Rasch model provides a perfect fit and reflects true item difficulties rather than just estimates. However, according to Woods and Barker, random variation in the respondents rules out the possiblity of a perfect fit. Thus, whereas we need to be open to the possiblity of new groupings of test purposes in accordance with advances in the field, we must proceed cautiously, weighing the pros and cons of each innovation.

2. Test Validity

The next issue we will consider is that of test validity. It is related to testing purpose in that a test can only be considered as valid or invalid with respect to some intended purpose. Although test validity is often discussed, the actual

measure of validity is illusive. Part of the problem is that, as Morrow (1981) points out, there is no such thing as "absolute validity." Validity exists only in terms of specified criteria. So, if the criteria selected are the wrong ones (i.e., not interesting or not useful), then the validity is spurious. Thus, the situation may arise wherein a test with admirable qualities is invalid in that it is used for inappropriate purposes. For example, a test may be an adequate measure of general placement but be of limited utility in diagnosis of specific reading problems.

Another part of the problem is that certain measures of validity lend themselves more easily to more conventional means of investigation, while others do not (Underhill 1983). Concurrent and predictive validity(3) can be readily assessed empirically through correlating results on the test under study with scores on other tests considered to be valid in terms of specified criteria. Construct, content validity, and face validity(4), on the other hand, are referred to by Underhill as

(3) "Concurrent validity" relates to the extent of correlation between the test results and those on another test believed to measure the same function, both taken at the same time. "Predictive validity" deals with the extent to which results on the test enable prediction of performance on another test in the future.

(4) "Construct validity" concerns the extent to which the items/tasks match the theory behind them. "Content validity" considers whether the items/tasks in the test match what the test as a whole purports to assess. "Face validity" deals with the issue of whether the test looks like a reasonable test.

forms of "theoretical validation" in that test evaluators must rely on "intuition and introspection" for their assessment.

In recent years the conventional means of assessing validation have been questioned and the more unconventional have been given more credence. It has been pointed out, for example, that assessing both concurrent and predictive validation is not so simple. The argument is made that a high correlation between two tests does not indicate which is preferable, or if either is any good for the given purpose, or whether one can be substituted for the other. Rather, it is suggested that trait-method interaction may be taking place -- i.e., that in a given language use situation, individual respondents will react differently (Low 1985). In addition, it has been suggested that the term "face validity" is unfortunate because of its derogatory overtones. Low (1985) would offer the term "perceived validity" instead. As relates to respondents, then, this form of validation would refer to their perceptions as to: 1) any bias in test content (i.e., whether the content seems to favor a respondent with certain background knowledge or expertise), 2) the nature of the task that they are being requested to perform, and 3) the nature of their actual performance on the test as a whole and on any particular subtests (test-taking strategies employed).

This concern for giving careful consideration to perceived validity comes at a time when mentalistic measures are being called upon to gather verbal reports from respondents regarding the strategies that they are using during the process of taking

tests (Cohen 1984, Cohen 1980, Dollerup et al. 1982). It is being demonstrated that the use of mentalistic measures can yield empirical data that provide considerable information concerning how respondents perceive tests and how they actually deal with them in testing situations. Having looked at the issues of clarifying purposes for testing and of considering respondents' perceptions of these tests, let us now look at key concerns in determining or evaluating methods of testing reading comprehension.

Methods of Testing Reading Comprehension

Reading comprehension items or procedures require of learners that they use a certain type or types of reading, comprehend at a certain level or combination of levels of meaning, enlist a certain comprehension skill or skills, and do all of this within the framework of a certain testing method or methods. In this section, we will look at some of the choices available to the test constructor and considerations of concern to the test user.

1. Type of Reading

Items and procedures can be written so that they implicitly or explicitly call for a given type of reading. For example, a respondent can be given a lengthy passage to read in a limited time frame such that the only way to handle it successfully is to

skim(5) or to scan(6), depending on the task. A distinction is also made between scanning and "search reading," where in the latter case the respondent is scanning without being sure about the form that the information will take (i.e., whether it is a word, phrase, sentence, passage, or whatever) (Pugh 1978). A respondent could also be given a passage to read receptively(7). Yet another approach is to have respondents read responsively, such that the written material acts as a prompt to them to reflect on some point or other and then possibly to respond in writing. Testing formats in which questions are interspersed within running text may especially cater to such an approach, if the questions stimulate an active dialog between the text and the reader.

The type of reading task is raised here because it would appear to be neglected at times in the process of test construction. In other words, reading items and tasks are sometimes constructed without careful consideration as to how the respondent is to read them. It may even be of benefit for the test constructor to indicate explicitly to the respondent the type of reading expected. For example, a certain item could be introduced by the following:

---

(5) Overall rapid inspection with periods of close inspection.

(6) Locating a specific symbol or group of symbols -- e.g., a date, a name of a person or place, a sum of money, etc.

(7) Discovering accurately what the author seeks to convey.

> Read the following text through rapidly (i.e., skim it) in order to get the main points. There will not be time to read the text intensively. When you have completed this reading, answer the questions provided -- without looking back at the text. You will have ten minutes for the exercise.

Another type of reading constituting a test of its own is oral reading. Various oral reading functions could be tested -- such as the giving of a talk from a scripted text, the announcing of public information (as if at a train station, airport, etc.), the reading aloud of the contents of a pamphlet (giving, for example, the operating instructions for some appliance), or the reading of a children's story. Given that the reading of text as oral recitation is not intended to be the same sort of behavior as silent reading (involving the skipping of words and phrases, regressions, and pauses), oral reading needs to be assessed by its own set of criteria, not by those used for assessing silent reading. For example, a scripted talk could be assessed in terms of smoothness of delivery, appropriateness of intonation, and so forth. The successful reading of a pamphlet could be based on whether stress is placed on those items of crucial importance in having the appliance operate successfully.

A possible misuse of oral reading has been as a means for tapping silent reading through assessing miscues -- i.e., the addition, subtraction, substitution, or transposition of material

while reading aloud (Leu 1982). Effective reading comprehension almost invariabley means silent reading. The reader of a scientific paper, for example, may well stop at numerous points and go back to check the precise working of earlier parts of the article, or periodically jump forward to read the footnotes, the references, or pre-read the conclusion (Carre 1981). In short, oral reading as recitation is not the same process as silent reading.

2. Level of Meaning

A test item or procedure can tap comprehension at one of four levels of meaning or at several levels simultaneously: grammatical meaning, propositional meaning, discoursal meaning, and pragmatic meaning (adapted from Nuttall 1982). Note, however, that these categories are presented as a rough rule of thumb, rather than as a hierarchy of discrete levels. Grammatical meaning deals with the meanings that words and morphemes have on their own. Propositional meaning refers to the meaning that a clause or sentence can have on its own, i.e., the information that the clause or sentence transmits. This meaning is also referred to as its "informational value." Discoursal meaning relates to the meaning a sentence can have only when in context. This meaning is also referred to as its "functional value." Pragmatic meaning concerns the meaning that a sentence has only as part of the interaction between writer and reader. This is the meaning that reflects the writer's feelings, attitudes, and the intended effect of the utterance upon the reader.

The level of meaning that has perhaps gotten the most attention in the literature in recent years is the discoursal one, especially the perception of rhetorical functions conveyed by text. For example, an item may overtly or covertly require a respondent to identify where and how something is being defined, classified, exemplified, or contrasted with something else. Often such "discourse functions" are signaled by connectors or "discourse markers." Nontheless, uninformed or unalert readers may miss these signals -- words or phrases such as "unless," "however," "thus," "whereas," and the like. Research has shown that such markers need not be subtle to cause reading problems. Simple markers of sequential points ("first," "also," and "finally") may be missed by a reader as well as more subtle markers (see Cohen et al. 1979).

3. Comprehension Skill

Not only must a test constructor and user be aware of levels of comprehension, but also of individual skills tested by reading comprehension questions at one or more such levels of meaning. There are numerous taxonomies of such skills. Alderson (1986) offers one which reflects a compilation of others, and includes: (1) the ability to recognize words and phrases of similar and opposing meaning, (2) the identifying or locating of information, (3) the discriminating of elements or features within context; the analysis of elements within a structure and of the relationship among them -- e.g., causal, sequential, chronological, hierarchical, (4) the interpreting of complex

ideas, actions, events, relationships, (5) inferencing -- the deriving of conclusions and predicting the continuation, (6) synthesis, and (7) evaluation. We note that this taxonomy omits the reader-writer relationship -- e.g., the author's distance from the text and the level of participation in the text that the author requires of the reader. With this taxonomy, as with others, the boundaries between skills are assumed to be discrete when, in reality, they may not be.

It is noteworthy that taxonomies of comprehension skills do not necessarily imply that the reading of texts requiring the use of so-called "higher-order" skills necessarily constitutes a more difficult task. In other words, interpreting complex relationships may not be any more difficult and perhaps easier than recognizing that two words are antonyms in a given context. Alderson (1986), for example, reported on a study in which both weaker and more proficient Bombay university students had as much difficulty with lower-order questions as they had with higher-order ones. One explanation given was that whereas the lower-order questions measured language skills, the higher-order ones measured cognitive skills which the lower-proficiency students had no problem with. Another explanation was that the lower- and higher-order distinction was faulty. Apparently ten expert judges at Lancaster disagreed on 27 out of the 40 reading items as to what each of them measured.

4. Testing Methods

Besides considering the type of reading to be performed, the desired levels of comprehension, and the comprehension skills to be tapped, the test constructor and user needs to give careful thought to the testing method. The challenge is to maximize the measurement of the trait -- i.e., the respondent's ability, while minimizing the reactive effects of the method. In order to do this, it is useful to be informed as to the options for testing with each method and what these options yield. We will look at three areas of concern regarding testing method -- the language of response, the cloze and the C-test, and the design of genuinely communicative reading comprehension tests.

a. The Language of Response

In foreign language tests, item responses have usually been in the foreign language, except in translation tasks. In the case of open-ended answers, Laufer (1983) offered three reasons why first-language responses might be preferable. She noted that when responses are in the foreign language, it is possible to copy answers from the text, writing may be of poor quality, and the respondent can be terse in order to play it safe, thus providing not quite enough information to judge whether the response is correct.

Researchers have recently been exploring the effects of mixed language formats -- elicitation in foreign language, response in first language. Shohamy (1984), for example, found that multiple-choice and open-ended responses in first language were easier to answer and were probably processed differently

than in the foreign language. Although she felt that having multiple-choice alternatives in the first language may give clues to the meaning of the text, she saw it as eliminating the use of tricky look-alike items and unknown distractors. She found that with her sample of Israeli twelfth-grade students of English as a foreign language, the language used for responses affected lower proficiency students more. She concluded that in criterion-referenced testing situations, where the purpose was to have every respondent performance at maximal level, then responses to foreign-language items should be in the first language.

In another study, Zupnik (1985b) had twenty Hebrew-speaking intermediate EFL students (in their first year at the university) perform two tasks on an English text. In the first task, the students were requested to read the text and were asked five questions in English, two involving definitions, the other three involving a reason, a relationship, and a process respectively. In this task they were to indicate the precise line(s) in the English text that provided an answer to the question. These responses were collected and then the respondents were asked the same questions again, but in the second task they were to provide open-ended answers for the questions in Hebrew -- first in rough draft, then in a revised version. Finding the relevant line of text in English was intended to reflect those types of questions that can be answered by quoting from the text, thus encouraging superficial reading. The first-language responses were expected to demand a deeper comprehension of the text.

The results showed first-language responses to reflect a lower level of comprehension than the foreign-language responses (42% average correct on the Hebrew version vs. 59% on the English version). Also, although the correlation between performance on the two forms was signficant ($p<.05$), it was low ($r=.45$). The researcher concluded that the two tests were in part testing different things. She pointed out that in reading a foreign-language text, it is possible to recognize that A causes B without understanding what B means. She noted that definitions were particularly easy to identify superficially and harder to explain in the first language. The item discrimination results indicated that the better respondents did better both on "locating abilities" (e.g., skimming and scanning), as called for in the English-language responses, and on reading in depth, as called for in the Hebrew-language responses. The better respondents were also more likely to paraphrase the relevant material from the text when responding in their first-language rather than translating word-for-word (85% of responses from the better students vs. 57% of responses from the weaker students).

b. The Cloze and the C-Test

The origins of the cloze test date back farther than many would think -- to 1897, in fact. At that time, Ebbinghaus proposed a series of tests that had one- or two-word deletions, rational deletion, and partial deletion from the beginning or end of words (Ebbinghaus 1897). There is a controversy concering the cloze test as to whether filling in cloze items is not just a

matter of perceiving local redundancy, but rather, involves an awareness of the flow of discourse across sentences and paragraphs, as Oller (1979, ch. 12) maintains. Whereas recent research would suggest that traditional fixed-word deletion is more of a micro-level completion test (a measure of word- and sentence-level reading ability) than a macro-level measure of skill at understanding connected discourse (Alderson 1983, Klein-Braley 1981), Chavez-Oller et al. (1985) have recently come out with yet another claim that cloze is sensitive to constraints beyond 5-11 words on either side of a blank, based on a reanalysis of earlier data.

As an alternative to the fixed-word deletion, researchers have turned to the rationale deletion cloze, whereby words are deleted according to predetermined, primarily linguistic criteria -- often stressing the area considered to be underrepresented, namely, macro-level discourse links (Levenston et al. 1984). Research by Bachman (1985) with EFL university students found that the rational deletion approach sampled much more across sentence boundaries and somewhat more across clause boundaries within the same sentence than did the fixed-ratio cloze. He concluded that the rational deletion cloze was a better measure of the reading of connected discourse, although he questioned its construct validity. Bachman found that while the rational deletion procedure affords the test developer a better means for making judgements regarding the content validity of such tests, the question remains as to whether such tests "in fact measure

the components of language proficiency hypothesized by the deletion criteria" (Bachman 1985:550) -- i.e., the flow of discourse across sentences and paragraphs within a text. Markham (1985), for example, would contend that even the rational deletion cloze does not measure comprehension of connected discourse. He gave 84 English-speaking university students of German an original and a scrambled version of a rational deletion cloze and found that neither were testing for global reading ability. Thus, the controversy continues.

A suggested alternative to the cloze test, namely the C-test, has been proposed by Klein-Braley and Raatz (Raatz & Klein-Braley 1982, Klein-Braley & Raatz 1984, Klein-Braley 1985, Raatz 1985). In this procedure, the second half of every other word is deleted, leaving the first and the last sentence of the passage intact. A given C-test consists of a number of short passages (maximum 100 words) on a variety of topics. This alternative eliminates certain problems associated with cloze, such as choice of deletion rate and starting point, representational sampling of different language elements in the passage, and the inadvertent assessment of written production as well as reading. With the C-test, being given a clue (half the word) serves as a stimulus for respondents to find the other half. The following is one passage within a C-test (from Raatz 1985):

> Pollution is one of the big problems in the world today. Towns a_____ cities a_____ growing, indu_____

is gro_____, and t_____ population o_____ the wo_____ is gro_____. Almost every_____ causes poll_____ in so_____ way o_____ another. T_____ air i_____ filled wi_____ fumes fr_____ factories a_____ vehicles, a_____ there i_____ noise fr_____ airplanes a_____ machines. Riv_____, lakes a_____ seas a_____ polluted b_____ factories and by sewage from our homes.

At present it would appear that the C-test may well be a more reliable and valid means of assessing what the cloze test assesses, but as suggested above, it is still not clear to what extent the C-test tests more than micro-level processing. Because half the word is given, students who do not understand the macro-context can still mobilize their vocabulary skills adequately to fill in the appropriate discourse connector witout indulging in higher-level processing. This was the finding from research using Hebrew C-tests (Cohen et al. 1984).(8) Extensive research on what processing of C-test items actually entails is currently underway -- using data from protocols of German speakers'verbal reports while taking French and Spanish C-tests, and more information will be available in the near future (Grotjahn 1986).

---

(8) Lo (Personal Communication) suggests that the C-test is a different test for VO languages as opposed to OV languages because verbal affixes and morphology are in different positions. For example, a Gaelic C-test would only give the first letter of a mutation and frequently, the letter given would be for the <u>affix</u> not for the noun stem.

c. Communicative Tests of Reading Comprehension

For years attention has been paid to so-called "communicative tests" -- usually implying tests dealing with speaking. More recently, efforts have been made to design truly communicative tests of other language skills as well, such as reading comprehension. Canale (1984) points out that a good test is not just one which is valid, reliable, and practical in terms of test administration and scoring, but rather one that is acceptable -- i.e., accepted as fair, important, and interesting by test takers and test users.(9) Also, a good test has feedback potential -- rewarding both test takers and test users with clear, rich, relevant, and generalizable information. Canale suggests that acceptability and feedback potential have often been accorded low priority, thus explaining the curious phenomenon of multiple-choice tests claiming to assess oral interaction skills.

Some recent approaches to communicative testing were in part an outgrowth of a theoretical framework proposed by Canale and Swain (1980), which offered a basis for communicative testing. This framework defined four types of competence that need to be considered in assessing communicative ability: grammatical,

(9) This position is an endorsement of the need to take into account "perceived validity" (Low 1985), as discussed above.

discoursal, sociocultural, and strategic.(10) Both Swain and Canale undertook to construct communicative tests consistent with their framework. The particular variety of communicative test that they dealt with has been referred to as a "storyline" test, a test with a line of development. In such a test, there is a common theme running throughout in order to assess context effects. The basis for such an approach is that the respondents learn as they read on, that they double back and check previous content, and that the ability to use language in conversation or writing depends in large measure on the skill of picking up information from past discussion and using it in formulating new strategies (Low, in press).

Swain (1984), for example, developed a storyline test of French as a foreign language for high-school French immersion students. The test consisted of six tasks around a common theme, "finding summer employment." There were four writing tasks (a letter, a note, a composition, and a technical exercise) and two speaking tasks (a group discussion and a job interview). The test was designed so that the topic would be motivating to the students and so that there would be enough new information

(10) "Grammatical competence" refers to mastery of the features and rules of the language, "discoursal competence" to cohesion (local links within the text) and coherence (interpretation and use of connected utterances in a meaningful whole), "sociocultural competence" to sociocultural rules of appropriateness (status, purpose, norms of interaction), and "strategic competence" to ways of compensating for imperfect knowledge of rules (such as through paraphrase, shifts in register, etc.).

provided in order to give the tasks credibility. Swain provided the respondents with sufficient time, suggestions as to how to do the test, and clear knowledge about what was being tested. There was access to dictionaries and other reference material, and opportunity to review and revise their work. Swain's main concern was to "bias for best" in the construction of the test -- to make every effort to support the respondents in doing their best of the test.(11)

Canale also provided a design for a communicative storyline test -- for administration to University-level learners of English as a second language in Ontario (Canale 1984). The example provided had a suggested theme, "a day in the life of a student." It consisted of four phases, a warm-up, a level check, a probe, and a wind-up. The warm-up was intended to put test takers at ease and to familiarize them with the language and the interviewer. The given example was that of "choosing one's courses," intended to afford the respondents an opportunity to decide which form of the test they wanted to try, an easier or a more difficult one. The level check identified the proficiency level at which the test taker performs best. The example provided dealt with "applying for a job or for aid," and consisted of short-answer responses.

---

(11) The point here is that such cases of bias can be viewed as a good thing -- as intentional bias. The aim would be to set up tasks that test takers will be motivated to participate in, such as those that approximate real-life situations (Spolsky 1985).

The probe was intended to challenge test takers with tasks just beyond their identified level in order to verify maximum proficiency and to show the test takers tasks which were still beyond their ability. In this subtest, respondents were asked to select a topic for a course report or take-home exam within their own discipline area. The wind-up was aimed at the test takers' best performance level in order to have them end with a sense of accomplishment. Test takers who took the same discipline-specific subtests were asked to engage in a semi-directed conversation on two themes: what each respondent proposed in the just-completed writing task and what they thought of the testing experience.

Canale (1985) views communicative tests such as that described above as "proficiency-oriented achievement tests," which is consistent with Henning's (1985) suggested "marriage" between proficiency and achievement testing mentioned above. Canale offers five reasons for taking this view:

(1) Such tests put to use what is learned. There is a transfer from controlled training to real performance.

(2) There is a focus on the message, the function, and the form, not just on the form.

(3) There is group collaboration as well as individual work, not just the latter.

(4) The respondents are called upon to use their resourcefulness in resolving authentic problems in language use as opposed to accuracy in resolving contrived problems at the linguistic level.

(5) The testing itself is more like learning, and the learners are more involved in the assessment.

Communicative storyline tests have also received criticism for various reasons (Jones 1984; Liskin-Gasparro 1984; Low, in press). The following are some of the reservations made about such types of tests:

(1) In order to approximate real life more, it is necessary to move away from mass administration and scoring, which is less practical and less objective. Tests that are acceptable (fair, important, and interesting) and give feedback are usually small-scale, classroom tests.

(2) With a thematic organization, there is less efficiency because learners need to produce more text or respond to fewer items.

(3) Such a test limits the variety of language material and thus leads to content bias expressly because the focus of the test is narrow.

(4) There is the possibility of contamination -- that a question relating to the first part of the test will be unintentionally answered in a later section. The fact that learners can use information from earlier parts of the test in answering subsequent questions lowers the test's reliability.

(5) It is difficult to design such tests because of the need to have genuine links between sections without having them too interdependent.

(6) There is a potential shock effect if respondents have not been tested by this approach before.

It would appear that such criticisms need to be taken into account when considering the use of communicative tests. There appear to be clear advantages to pursuing such testing approaches, accepting their limitations. A Hebrew University seminar paper (Brill 1986), for example, had thirty-two ninth-grade Hebrew speakers complete a communicative storyline test, including five tasks dealing with membership in a youth group.(12) The students were then asked to compare their experience on this test and on the traditional multiple-choice one they had taken previously. They almost unanimously endorsed the communicative test as preferable because it was more creative, allowed them to express their opinions, was more interesting, taught them how to make contact with others, and investigated communication skills besides reading comprehension. For these reasons, they felt that it provided a truer measure of their competence than did the traditional test.

d. Computerized Adaptive Testing (CAT)

---

(12) The tasks included: writing a letter as a response to a friend interested in a youth movement the respondent belongs to; presenting questions to the group leader to get more information on the movement; preparing an announcement about the movement to post on bulletin boards; writing out a telephone request for information on how a local foundation could aid the movement, and writing out a telephone response to an invitation by a political group to join a demonstration of theirs.

Computerized adaptive testing (CAT) of reading comprehension implies an approach to testing whereby the selection and sequence of items depends on the pattern of success and failure experienced by the respondent. Most commonly, if the respondent succeeds on a given item, one of greater difficulty is presented, and if the respondent experiences failure, then an easier item is presented. The testing continues until sufficient information has been gathered to assess the particular respondent's ability. At present, such tests are mostly limited to objective formats, such as multiple-choice. Based on item response theory(13), CAT is known to be more efficient and more accurate than conventional fixed-length tests employing multiple-choice items (Tung 1986).

Among the advantages of CAT are the following: individual testing time may be reduced, frustration and fatigue are minimized, boredom is reduced, test scores and diagnostic feedback may be provided immediately, test security may be enhanced (since it is unlikely that two respondents would receive the same items in the same sequence), record-keeping functions are improved, and information is readily available for research purposes (Henning, in press). The main disadvantage is that given its present item-response-theory basis, CAT requires that

---

(13) Item response theory (also referred to as "latent trait measurement") refers primarily to analytical procedures for quantifying the probability of individual item and person response patterns given the overall pattern of responses in a set of test data (Henning 1984). Reference was also made to item response theory above under "Theoretical Issues in Testing."

the construct to be measured be unidimensional -- i.e., be assumed to involve only one major factor or underlying trait. It is suggested that such an assumption threatens to trivialize and compromise the existing theories of reading comprehension, which include multiple dimensions, such as world knowledge, language and cultural background, type of text, reading styles, and so forth, and fails to take into consideration various subcomponents of reading, along with the influences of instruction (Canale 1986).

The line of development that Canale (1986) would propose for CAT is that it move from simply mechanizing existing product-oriented reading comprehension item types to the inclusion of more process-oriented, interactive tasks that can be integrated into broad and thematically coherent language use/learning activities, such as "intelligent tutoring systems."(14)

Test-Taking Strategies

The strategies that respondents use in taking tests have implications both for the issue of test validity and "bias for best." Tests that are relied upon to indicate the comprehension level of readers may produce misleading results because of numerous techniques that readers have developed for obtaining correct answers on such tests without fully or even partially

(14) In intelligent tutoring systems, the computer diagnoses the students' strategies and their relationship to expert strategies, and then generates instruction based on this comparison.

understanding the text. As Fransson (1984) puts it, respondents may not proceed via the text but rather around it. In effect, then, there are presumptions held by test 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. It may also be that the strategies the respondents are using are detrimental to their overall performance, or at least not as helpful as others they could be using.

Mentalistic measures using verbal report have helped determine how respondents actually take reading comprehension tests as opposed to what they may be expected to be doing (Cohen 1984). Studies calling on respondents to provide immediate or delayed retrospection as to their test-taking strategies regarding reading passages with multiple-choice items have, for example, yielded the following results:

(1) Whereas the instructions ask students to read the passage before answering the questions, students have reported either reading the questions first or reading just part of the article and then looking for the corresponding questions.

(2) Whereas advised to read all alternatives before choosing one, students stop reading the alternatives as soon as they have found one that they decide is correct.

(3) Students use a strategy of matching material from the passage with material in the item stem and in the alternatives, and prefer this surface-structure reading of the test items to one that calls for more in-depth reading and inferencing.

(4) Students rely on their prior knowledge of the topic and on their general vocabulary.

Recent Hebrew University student seminar papers have provided innovations in two areas of investigation regarding test-taking strategies -- in the use of first-language responses to foreign-language passages and in the use of a response-strategy checklist used after each response. The first study had two Hebrew-speaking respondents, a strong and a weak reader respectively, engage in reading comprehension testing tasks (Zupnik 1985a). The students read an EFL text and answered five questions in English by indicating the line(s) in the text that provided an answer to the question, and then answered the same questions again, this time providing open-ended answers for the questions in Hebrew (as in Zupnik 1985b, mentioned above).

Both respondents were trained to produce think-aloud and self-observational data(15), and were then asked to provide such data regarding both language tasks before answering the questions in writing. The poor reader was found to use four times as many reading strategies on the English response task than did the

---

(15) "Think-aloud" data reflect stream-of-consciousness disclosure of thought processes while the information is being attended to. Such data are basically unedited and unanalyzed. "Self-observation," on the other hand, refers to the inspection of specific reading behavior, either while the information is still in short-term memory, i.e., introspectively, or after the event, i.e., retrospectively (usually after 20 seconds or so). It does entail analysis and editing of the data to a lesser or greater degree.

strong reader (using Sarig's taxonomy of reading strategies(16)). Both readers used a similar number of strategies on the Hebrew response task. As to the type of reading strategies used, it was found that the better reader used monitoring strategies most of all in both languages, while the poorer reader relied mostly on clarification and simplification strategies, with very limited use of monitoring strategies. Furthermore, most of the strategies of the stronger reader were comprehension-promoting, while those of the poorer reader were often comprehension-deterring. As in the companion group study (Zupnik 1985b), this case study confirmed the hypothesis that quoting rhetorically-focused foreign-language segments from text encourages more superficial reading than answering in the first language.

---

(16) On the basis of protocol analysis of high-school students reading Hebrew as a first language and English as a foreign language, Sarig (1987) designed a taxonomy of "reading move types," which includes four broad categories of moves or strategies: technical-aid moves (reading acts undertaken to facilitate higher-level moves -- e.g., skimming for the purpose of determing the macro-frame of the text -- and notes taken while reading), clarification and simplification moves (semantic-decoding moves, involving paraphrase to simplify syntax, vocabulary, ideas, or rhetorical functions), coherence-detecting moves (using textual or extra-textual clues to make the text meaningful -- e.g., through textual and content schemata, rhetorical functions, ideas and views expressed), and monitoring moves (conscious strategies for checking on the reading process -- e.g., awareness of the task being performed, identification of misunderstanding and incompaatibility of formerly interpreted material with newly interpreted material, awareness of other failures in comprehension, and awareness of resources for remedy and likelihood of success).

The second piece of innovative research on test taking dealt with the refining of a research methodology for tapping test-taking strategies. The issue under study was whether it is possible to collect introspective and retrospective data from students just after they have answered each item on a test. The approaches reported on in previous work have involved at most a request of respondents after they have finished a subtest or group of items that they reflect back as to the strategies that they used in arriving at answers to those items (Cohen 1984). In an effort to provide immediate verbal report data, Nevo (1985) 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 as to 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 checklist (see Figure 1). She administered a multiple-choice reading comprehension test in Hebrew first-language and French foreign-language to forty-two 10th graders, and requested that they indicate for each of the ten questions on each test, the strategy that was most instrumental in their arriving at an answer as well as that which was the second most

instrumental. 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 report.

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 first language to 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 test than in the first-language version. 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.

Futhermore, the respondents reported benefiting greatly from the opportunity to become aware of how they took reading tests. They reported being basically unaware of their strategies prior to this study.(17)

In terms of the actual strategies used for answering the multiple-choice tests in Hebrew as a first language and French as a foreign language, Nevo found that "returning to the text to

---

(17) What was not looked at were the carry over effects of this study on those same respondents the next time that they took a reading test. Such research would help to determine whether this awareness is only temporary or whether it has a lasting effect.

look for the correct answer after reading the questions" and "looking for clues to the answer in the section of text that the question referred to" were the two most frequently reported strategies, both in first and in foreign language. In foreign language, however, respondents were somewhat less likely to return to the text in general, probably reflecting the greater processing difficulties this involved. The major difference in first-language vs. foreign-language test-taking strategies was that in first language, "guessing without any particular considerations" was rarely utilized, while in foreign-language responses, it was reportedly used for 20%-30% of the items on the test. Nevo's study pinpointed not only the frequency of guessing, but the specific items for which it was reported.(18)

From these findings and from others, there is emerging a description of what respondents do to answer questions. Unless trained to do otherwise, they may use the most expedient means of responding available to them -- such as relying more on their previous experience with seemingly similar formats than on a close reading of the description of the task at hand. Thus, when given a passage to read and summarize, they may well perform the task the same way they did the last summary task, rather than paying close attention to what is called for in the current one.

---

(18) This study made a dichotomy between guessing without any particular considerations and not guessing. In reality, there is a continuum from guessing without considerations to thoughtful guessing to non-guessing.

Often, this strategy works, but on occasion the particular task may require subtle or major shifts in response behavior in order to perform well.

There appears to be a further insight to be gained from the test strategy literature, namely, that indirect testing formats -- i.e., those which do not reflect real-world tasks (e.g., multiple-choice, cloze, etc.) -- may prompt the use of strategies solely for the purpose of coping with the test format. More direct formats such as summarizing a test may be free of such asdded testing effects. However, as long as the task is part of a test, students are bound to use strategies they would not use under non-test conditions. It is largely the responsibility of test constructors and of those who administor such tests to be aware of what their tests are actually measuring. Verbal report techniques can assist the test developer and user in obtaining such information.

Insights about the way in which respondents go about performing different testing tasks can be used to make informed decisions as to: (1) the choice of testing format, (2) the choice and wording of instructions, and (3) the value and feasibility of coaching the respondents in how to take language tests. Work by O'Malley (1986) and others has already made use of research findings in designing training modules for the learning of test-taking skills.

Conclusions

This chapter has not attempted to survey the whole field of testing as it relates to linguistic and communicative proficiency. Rather, it has touched on some of the issues regarding the testing of reading comprehension that have been of major concern to test developers, test users, and test takers during recent years. Reconsideration of the purposes for tests and of how to combine purposes has been a key interest in this chapter, as have questions of test validation. Sometimes careful attention is given to innovation in testing method -- whether through cloze, C-testing, or through computerized adaptive testing -- without commensurate attention paid to the type of reading being called for, the level of comprehension desired, and the comprehension skills to be elicited. For this reason, attention was given to these factors here.

During this period of awakened interest in learners' processing of language, it seems fitting that we should pay extra attention to the actual strategies being used in test taking. There is no doubt that test constructors and test users can receive beneficial feedback from inquiries into what the given tests actually prompt respondents to do -- beyond their expectations or assumptions. As for test takers, they are sometimes if not frequently oblivious to how they are answering test items, possibly to their detriment. It is possible that they would become more effective at taking tests if they were informed as to what they are doing at present and as to what they could be doing that would yield better results.

Figure 1

Strategies for Answering Multiple-Choice Reading Comprehension Questions (From Nevo 1985)

1. Background knowledge: general knowledge outside the text.
2. Guessing: guessing without any particular considerations.
3. Returning to the passage: returning to the text to look for the correct answer, after reading the questions and multiple-choice alternatives.
4. Chronological order: looking for the answer in chronological order in the passage.
5. Clues in the text: locating the area in the text that the question referred to and then looking for clues to the answer in that context.
6. Ceasing search at plausible choice: reading the alternative choices until reaching one that was thought to be correct. Not continuing to read the rest of the choices.
7. Process of elimination: selecting an alternative not because it was thought to be correct but because the others did not seem reasonable, seemed similar, or were not understandable.
8. Choosing the exception: suspecting a choice to be the correct answer because it constituted an exception or had something different about it.
9. Length: being drawn to an alternative because it was longer/shorter.
10. Location: being influenced by the location of the alternative within the set of alternatives.
11. Common word: choosing an alternative because it had in it a word that was common -- that was heard all the time.
12. Key word: arriving at an alternative because it had in it a word that appeared to be a key word.
13. Matching the stem with an alternative: selecting an alternative because it had in it a word/words that appeared in the item stem as well.
14. Association: selecting the alternative because it had a word in it that evoked an association with a word in the first language or in another language.
15. Matching the question with the text: selecting an alternative because it had a word/words that also appeared in the text, because it had words similar in sound, meaning, or belonged to the same word family, or because it just seemed to be related.
16. Other strategy

REFERENCES

Alderson, J.C. The cloze procedure and proficiency in English as a foreign language. In J.W. Oller, Jr. (Ed.), Issues in language testing research. Rowley, MA: Newbury House, 1983, 205-228.

Alderson, J.C. Cognition and Reading. Lancaster: Institute for English Language Education, University of Lancaster, 1986. (Paper presented at the Colloquium on Reading in a Second Language, Annual TESOL Convention, Anaheim, CA, April 4, 1986.)

Bachman, L.F. Performance on cloze tests with fixed-ratio and rational deletions. TESOL Quarterly, 1985, 19 (3), 535-556.

Brill, H. Developing a communicative test of reading comprehension and determining its effectiveness. Seminar paper, School of Education, Hebrew University, Jerusalem, 1986. (In Hebrew)

Canale, M. & Swain, M. Theoretical bases of communicative approaches to second-language teaching and testing. Applied Linguistics, 1980, 1 (1), 1-47.

Canale, M. Considerations in the testing of reading and listening proficiency. Foreign Language Annals, 1984, 17 (4), 349-357.

Canale, M. Proficiency-oriented achievement testing. Toronto: Franco-Ontarian Centre and Curriculum Department, Ontario Institute for Studies in Education, 1985.

Canale, M. The promise and threat of computerized adaptive assessment of reading comprehension. In C.W. Stansfield (Ed.), Technology and language testing. Washington, D.C.: TESOL, 1986.

Carre, C. Language teaching and learning. 4. Science. London: Ward Lock Educational, 1981.

Chávez-Oller, M.A., Chihara, T., Weaver, K.A., & Oller, J.W., Jr. When are cloze items sensitive to constraints across sentences? Language Learning, 1985, 35 (2), 181-206.

Cohen, A.D. Testing language ability in the classroom. Rowley, MA: Newbury House, 1980, 46-53.

Cohen, A.D. On taking language tests: What the students report. Language Testing, 1984, 1 (1), 70-81.

Cohen, A.D. Using verbal reports in research on language learning. In C. Faerch & G. Kasper (Eds.), Introspection in Second Language Research. Clevedon, England: Multilingual Matters, in press.

Cohen, A.D., Glasman, H., Rosenbaum-Cohen, P.R., Ferrara, J., & Fine, J. Reading English for specialized purposes: Discourse analysis and the use of student informants. TESOL Quarterly, 1979, 13 (4), 551-564.

Cohen, A.D., Segal, M., & Weiss Bar-Simon-Tov, R. The C-test in Hebrew. Language Testing, 1984, 1 (2), 221-225.

Dollerup, C., Glahn, E., & Rosenberg-Hanson, C. Reading strategies and test-solving techniques in an ESL-reading comprehension test -- a preliminary report. Journal of Applied Language Study, 1982, 1 (1), 93-99.

Ebbinghaus, H. Über eine neue Methode zur Prufung geistiger Fahigkeiten und ihre Anwendung bei Schulkindern. In S. Exner et al. Zeitschrift fur Psychologie und Physiologie der Sinnesorgane. Leipzig: Barth, 1897.

Fransson, A. Cramming or understanding? Effects of intrinsic and extrinsic motivation on approach to learning and test performance. In J.C. Alderson & A.H. Urquhart (Eds.), Reading in a Foreign Language. London: Longman, 1984, 86-121.

Grotjahn, R. Test validation and cognitive psychology: Some methodological considerations. Language Testing, 1986, 3 (2).

Henning, G. Advantages of latent trait measurement in language testing. Language Testing, 1984, 1 (2), 123-133.

Henning, G. Proficiency testing and achievement testing: A proposal of marriage. Los Angeles: ESL Section, Dept. of English, University of California, 1985.

Henning, G. Computer adaptive testing: Assertions, assumptions, illustrations, and estimations. Language Testing, in press.

Jacobs, H.L. Zinkgraf, S.A., Wormuth, D.R., Hartfiel, V.F., & Hughey, J.B. Testing ESL Composition. Rowley, MA: Newbury House, 1981.

Jones, R.L. Testing the receptive skills: Some basic considerations. Foreign Language Annals, 1984, 17 (4), 365-367.

Klein-Braley, C. Empirical investigations of cloze tests: An examination of the validity of cloze tets as tests of general language proficiency in English for German university students. Unpublished doctoral dissertation, University of Duisburg, Duisburg, West Germany, 1981.

Klein-Braley, C. A cloze-up on the C-test: A study in the construct validation of authentic tests. Language Testing, 1985, 5 (1), 76-104.

Klein-Braley, C. & Raatz, U. A survey of research on the C-test. Language Testing, 1984, 1 (2), 134-146.

Laufer, B. Written answers in Hebrew to English comprehension questions -- some advantages. English Teachers' Journal (Israel), 1983, 29, 59-64.

Leu, D.J., Jr. Oral reading error analysis: A critical review of research and application. Reading Research Quarterly, 1982, 17 (3), 420-437.

Levenston, E.A., Nir, R., & Blum-Kulka, S. Discourse analysis and the testing of reading comprehension by cloze techniques. In A.K. Pugh & J.M. Ulijn (Eds.), Reading for professional purposes. London: Heinemann, 1984, 202-212.

Liskin-Gasparro, J.E. Practical considerations in receptive skills testing. Foreign Language Annals, 1984, 17 (4), 369-373.

Low, G. Validity and the problem of direct language proficiency tests. In J.C. Alderson (Ed.), Lancaster practical papers in English language education 6. Oxford: Pergamon, 1985, 151-168.

Low, G.D. Storylines and other developing contexts in use-of-language test design. Indian Journal of Applied Linguistics, in press.

Markham, P.L. The rational deletion cloze and global comprehension in German. Language Learning, 1985, 35 (3), 423-430.

Morrow, K. Communicative language testing: Revolution or evolution? In J.C. Alderson & A. Hughes (Eds.), Issues in language testing. London: British Council, 1981, 9-25.

Nevo, N. Strategies in taking a multiple-choice reading comprehension test. Seminar paper, School of Education, Hebrew University of Jerusalem, 1985. (In Hebrew)

Nuttall, C. Teaching reading skills in a foreign language. London: Heinemann, 1982, 80-81.

Oller, J.W., Jr. Language tests at school. Ch. 12. London: Longman, 1979.

O'Malley, J.M. Test-taking strategies for ESL students. Rosslyn, VA 22209: InterAmerica Research Associates (1555 Wilson Blvd., Suite 600), 1986.

Pugh, A.K. Silent reading. London: Heinemann, 1978, 53.

Raatz, U. Tests of reduced redundancy -- The C-test, a practical example. Fremdsprachen un Hochschule, Bochum, West Germany: AKS-Roundbriefe 13-14, 1985, 14-19.

Raatz, U. & Klein-Braley, C. The C-test -- a modification of the cloze procedure. In T. Culhane et al. (Eds.), Practice and problems in language testing 4. Colchester: U. of Essex, 1982.

Raatz, U. & Klein-Braley, C. How to develop a C-test. Fremdsprachen un Hochschule, Bochum, West Germany: AKS-Roundbriefe 13-14, 1985, 20-22.

Sarig, G. High-level reading tasks in the first and in the foreign language: Some comparative process data. In J. Devine, P.L. Carrell, & D. Eskey (Eds.), Research in reading in a second language. Washington, D.C.: Teachers of English to Speakers of Other Languages, 1987.

Shohamy, E. Does the testing method make a difference? The case of reading comprehension. Language Testing, 1984, 1 (2), 147-170.

Spolsky, B. Intentional and unintentional bias. Ramat Gan: English Department, Bar-Ilan University, 1985.

Swain, M. Large-scale communicative language testing: A case study. In S.J. Savignon & M.S. Berns (Eds.), Initiatives in communicative language teaching. Reading, MA: Addison-Wesley, 1984, 185-201.

Tung, P. Computerized adaptive testing: Implications for language test developers. In C.W. Stansfield (Ed.), Technology and language testing. Washington, D.C.: TESOL, 1986.

Undernill, N. Commonsense in oral testing: Reliability, validity, and affective factors. In M.A. Clarke & J. Handscombe (Eds.), On TESOL '82. Washington, D.C.: TESOL, 1983, 125-139.

Woods, A. & Baker, R. Item response theory. Language Testing, 1985, 2 (2), 117-140.

Zupnik, Y. A comparison of cognitive processes in foreign-language and first-language responses. Seminar paper, School of Education, Hebrew University of Jerusalem, 1985a.

Zupnik, Y. A comparative study: English/Hebrew responses to open-ended reading comprehension test questions. Seminar paper, School of Education, Hebrew University of Jerusalem, 1985b.