Language Testing Reconsidered

Fox, Janna, Wesche, Mari, Bayliss, Doreen, Cheng, Liying

Published by University of Ottawa Press

Fox, Janna, et al.
Language Testing Reconsidered.

For additional information about this book
https://muse.jhu.edu/book/4459
THE COMING OF AGE FOR RESEARCH ON TEST-TAKING STRATEGIES

Andrew D. Cohen
University of Minnesota

Abstract

In this selective look at research on test-taking strategies over the last twenty-five years, brief mention is made of the beginnings of test-taker strategy research and then important developments in its evolution to the present are discussed, focusing on conceptual frameworks for classifying strategies, L1- and L2-related strategies, proficiency level and test-taking strategies, strategies as a function of testing method, and the appropriateness of the research methods. The review notes the valuable role that verbal report methods have played in the process of understanding what tests actually measure. The conclusion is that while test-taking strategy research has come of age over the last twenty-five years, there still remain numerous challenges ahead, such as in arriving at a more unified theory for test-taking strategies. Another challenge is to continue finding ways to make the research effort as unobtrusive as possible, while at the same time tapping the test-taking processes.

Three decades ago, L2 assessment validation research was focused for the most part on the outcomes of testing — namely, on how tests fared in terms of item performance (item difficulty and item discrimination), test reliability, the intercorrelation of subtests and the relationship between the test and other tests or criterion variables (e.g., GPA), and the effects of different test methods. What was missing was the aspect of test validation that related to respondents' behaviours in taking the tests: little was known about what they were actually doing to produce answers to questions and how it corresponded to the abilities one sought to test.

At that time there was only a small group of assessment specialists who were concerned that claims of test validity required attention as to how the respondents arrived at their answers. As formulated in early studies (see Cohen, 2000, for details), this meant paying attention to the kinds of strategies that respondents were drawing upon as they completed language tests — that is, the consciously selected processes that the respondents used for dealing with both the language issues and the item-response demands in the test-taking tasks at hand. More precisely, the focus was on both language learner strategies (i.e., the ways that respondents operationalized their basic skills of listening, speaking, reading, and writing, as well as the related skills of vocabulary
learning, grammar, and translation), the separate set of test-management strategies (i.e., strategies for responding meaningfully to the test items and tasks), and a likewise separate set of test-wiseness strategies (i.e., strategies for using knowledge of test formats and other peripheral information to answer test items without going through the expected linguistic and cognitive processes).

It proved a formidable task to obtain information about what respondents were doing without being obtrusive, while efforts to be unobtrusive often left us in the realm of speculation. Verbal report became a primary research tool for this endeavor, as reported in the author's article on test-taker strategies in the first issue of *Language Testing* (Cohen, 1984). Verbal reports include data that reflect one or more of the following types of data:

1. **Self-report**: learners' descriptions of what they do, characterized by generalized statements, in this case, about their test-taking strategies — for example, “On multiple-choice items, I tend to scan the reading passage for possible surface matches between information in the text and that same information appearing in one of the alternative choices.”

2. **Self-observation**: the inspection of specific, contextualized language behaviour, either introspectively, that is, within 20 seconds of the mental event, or retrospectively — for instance, “What I just did was to skim through the reading passage for possible surface matches between information in the text and that same information appearing in one of the alternative choices.”

3. **Self-revelation**: think aloud, stream-of-consciousness disclosure of thought processes while the information is being attended to — for example, “Hmm ... I wonder if the information in one of these alternative choices also appears in the text.”

In the intervening years, the use of verbal report to gain a better understanding of the testing process has evolved from simply describing and codifying strategies that respondents use to respond to different item types and testing procedures to more theoretically based, rigorous, and statistically sophisticated research efforts. Examples include identifying learner's use of test-taking strategies to validate testing formats and specific tests, investigating how proficiency level and other learner characteristics relate to strategy use and test performance, and studying the impact of strategy instruction on learners' performance on standardized tests.

In this selective look at research on test-taking strategies over the last twenty-five years, an attempt will first be made to characterize the beginnings of test-taker strategy research and then important developments in its evolution to the present will be discussed, focusing on the issues studied, the research methodology used, and the significance of the findings for the field of language testing.
Early Work on Test-Taking Strategies

The student research studies that were reviewed in the author's 1984 *Language Testing* article constituted some of the early efforts in the field of L2 test-taking strategies were inspired by several key studies. Danish researchers had used introspection and retrospection to study the responses of high school and college EFL students to three multiple-choice test item types embedded in connected text. Students explained which alternatives they would choose and why they thought the selected alternative was the correct one (Dollerup, Glahn, and Rosenberg Hansen, 1982). Their findings that each item produced an array of strategies and that even erroneous decoding could produce the correct answer demonstrated the potential relevance of information about learner response processes to test validity. Another influential early study dealt with strategy use in responding to random and rational deletion cloze tests (Homburg and Spaan, 1981). Based on respondents' reports of strategy use and success with different random-deletion item types, a rational-deletion cloze was constructed with four item types presumed to require different strategies: recognition of parallelism across phrases, processing within a given sentence, or the use of cues beyond the sentence in either forward or backward reading, depending on the location of these cues in the mutilated passage. Success at completing blanks requiring forward reading was related to success at understanding the main idea.

The work on the C-test (Raatz and Klein-Braley, 1981) also involved efforts to interpret how respondents produced responses to deleted material. In this procedure, the second half of every other word was deleted, leaving the first and the last sentence of the passage intact. A given C-test consisted of a number of short passages (a maximum of 100 words per passage) on a variety of topics. This alternative eliminated 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. The introduction of this measure stimulated a series of research studies using self-report and what was referred to as *logical task analysis* to determine just what the measure was assessing (see Klein-Braley and Raatz, 1984; Klein-Braley, 1985; Grotjahn, 1987).

In the review of student research at the Hebrew University of Jerusalem (Cohen, 1984), the closeness-of-fit between the tester's presumptions about what was being tested and the response processes that the test-takers reported was explored. These studies mainly involved comparison of reported strategy use with different test formats. One early study investigating test method effect in EFL reading testing by Israeli high school students (Gordon, 1987) found a relationship between proficiency and strategy use. There were four response formats: multiple-choice questions in English and in Hebrew, and
open-ended questions in English and in Hebrew. A subgroup of respondents were asked to verbalize their thoughts while they sought answers to each question. Low-proficiency students were found to process information at the local (sentence/word) level, without relating isolated bits of information to the whole text. They used individual word-centred strategies such as matching alternative words to text, copying words out of the text, word-for-word translation, formulating global impressions of text content on the basis of key words or isolated lexical items in the text or in test questions. High-proficiency students were seen to comprehend the text at the global level — predicting information accurately in context and using lexical and structural knowledge to cope with linguistic difficulties. As to performance, open-ended questions in the L2 (English) were found to be the most difficult and the best discriminator between the high- and low-proficiency students, since the low-proficiency students had difficulty with them.

A final early period study that deserves mention, one of the few conducted with other languages, looked at strategies used by students taking reading comprehension tests in Hebrew (their L1) and French (their L2) (Nevo, 1989). A multiple-choice format was used to test comprehension of two reading passages each in Hebrew and French by 42 tenth graders. Students answered open-ended questions at the end of each passage to evaluate the test items, completed a checklist of introspective strategies immediately after answering each multiple-choice item, and filled out a more general questionnaire at the end of the test. A transfer of test-taking strategies from the first to second language was noted.

Since verbal report in its various manifestations has been such an important tool in describing test-taking strategies, it has also been important to fine-tune our understanding of verbal report and how to use it in data collection (Cohen, 2000). Until recently, when it would appear that verbal report has finally gained a level of credibility and acceptability, there was not only a need to respond to criticism of verbal report methods but also to make a case for more robust verbal report methods and more complete write-ups.†

Themes in the Test-Taking Strategy Research

Before looking further at specific areas that have received focus in recent years, let us briefly consider five important themes in test-taking strategy research, some of which date from the early studies.

†[Ed. note: For a discussion of new approaches in Verbal Protocol Analysis (VPA), see Lazaraton and Taylor, Chapter 6.]
**Conceptual Frameworks for Classifying Strategies**

The first recurrent theme deals with conceptual frameworks for classifying strategies. While the model for strategies often referred to in the test-taking literature is that of O’Malley and Chamot (1990), the debate continues as to what a language learner strategy is — and by extension, what a test-taking strategy is. A recent survey of strategy experts as to what language learner strategies are yielded both consensus and continuing disagreement (Cohen, in press). For instance, it was a matter of debate as to how conscious of and attentive to their language behaviours learners need to be in order for those behaviours to be considered strategies. In addition, while there was consensus that learners deploy strategies in sequences or clusters, there was some disagreement as to the extent to which a behaviour needs to have a mental component, a goal, an action, a metacognitive component (involving planning, monitoring, and evaluation of the strategy), and a potential that its use will lead to learning in order for it to be considered a strategy. So, in essence, two contrasting views emerged, with each having its merits. On the one hand, there was the view that strategies need to be specific, small, and part of a combination of strategies related to a task; and on the other, there was the view that strategies need to be kept at a more global, flexible, and general level. Notwithstanding these differing views, there was enthusiastic agreement for the view that strategy use and effectiveness will depend on the particular learners, the learning task, and the environment.

As indicated at the outset, this review considers there to be three largely distinct sets of strategies: language learner strategies, test-management strategies, and test-wiseness strategies. Hence, in responding to a reading comprehension item, the respondents may well be drawing from their repertoire of reading strategies (e.g., “looking for markers of meaning in the passage such as definitions, examples, indicators of key ideas, guides to paragraph development” [Cohen and Upton, 2006, p. 34]), test-management strategies (e.g., “selecting options through the elimination of other options as unreasonable based on paragraph/overall passage meaning” [p. 37]) and test-wiseness strategies (e.g., “selecting the option because it appears to have a word or phrase from the passage in it — possibly a key word” [p. 37]).

**L1- and L2-Related Strategies**

A second theme highlighted in a number of studies relates to whether the strategies employed in L2 test-taking are specific to first-language (L1) use, common to L1 and L2 use, or more typical of L2 use. This issue has included, for example, study of how the use of L1 in L2 testing impacts the results — for example, the writing of an L2 essay on a test in the L1 first and then translating vs. writing directly in the L2. Brooks-Carson and I found that, while two-thirds...
of a group of intermediate L2 learners of French at the college level had their essays rated better if they were written directly in French (with only occasional mental translation from English), one-third of the group fared better if they wrote their essay in English first and then translated it into French (see Cohen and Brooks-Carson, 2001, for more details).

**Proficiency Level and Test-Taking Strategies**

A third theme in the literature has been that of the influence of the respondents' proficiency level on the test-taking strategies that they employ, with a focus on the frequency with which certain strategies are employed by respondents at different proficiency levels and the relative success of their use. For example, it may be expected that weaker respondents indulge more in the use of test-wiseness strategies as a way of compensating for lack of language proficiency.

**Strategies as a Function of Testing Method**

A fourth recurrent theme is the question of how the strategies deployed in responding to a given language assessment measure are in part a function of the testing method. For example, in the case of C-tests (where the second half of every other word is deleted, leaving the first and the last sentence of the passage intact), it has been found that response strategies predominantly involve micro-level processing. Since half the word is given, students who do not understand the macro-context have been observed to mobilize their vocabulary skills in order to fill in the appropriate discourse connector without indulging in higher-level processing (see, for example, Cohen, Segal, and Weiss Bar-Siman-Tov, 1985; Stemmer, 1991).

**Appropriateness of the Research Methods**

A fifth theme is that of the appropriateness of the research methodology for the study of test-taking strategies. The above description of the earlier studies alluded to some of these methodologies, especially the use of verbal report, whether obtained through interviews or other forms of face-to-face interaction, or through questionnaires, checklists, or other means. More recently, verbal report has involved technologically sophisticated approaches (such as the software application Morae mentioned below, TechSmith, 2004). Whereas a robust discussion of these varying methods is beyond the scope of this chapter, nonetheless it is important to point out that efforts are always being made to refine the measures. That said, some of the earlier techniques are still the most insightful.

While the prime vehicle for test-taking strategy research continues to be verbal report, there have been a few changes in procedures for conducting such verbal reports over the years in an effort to improve the reliability and
validity of the results. One has been to model for the respondents the kinds of responses that are considered appropriate (see, for example, Cohen and Upton, 2006), rather than to simply let them respond however they wish, which often failed to produce enough relevant data. In addition, in the collection of think-aloud and introspective self-observational data, researchers now may intrude and ask probing questions during data collection (something they tended not to do in the past), in order to make sure, for instance, that the respondents indicate not just their rationale for selection “b” as the correct alternative in multiple-choice, but also their reasons for rejecting “a,” “c,” and “d.” With regard to all forms of verbal report, respondents have also been asked to listen to a tape-recording or read a transcript of their verbal report session in order to complement those data with any further insights they may have (Nyhus, 1994). These innovations have helped to improve the quality of the data to a certain extent. Nonetheless, certain issues have not been resolved even to this day, such as the impact of intrusive measures on test-taking performance, or the fact that for that very reason, strategy data are usually not collected in actual high-stakes testing situations. Hence, the strategies actually used in responding to tests in high-stakes situations may differ from those identified under research conditions.

The following section deals with developments in test-taking strategy research in more recent years, in an effort to illustrate the areas of concern that researchers have focused on with regard to test-taking strategies, the kinds of test-taking strategies investigated, and the purposes for these investigations.

Research Related to Test-Taking Strategies from 1990 to 2005

The last fifteen years have seen a modest but steady increase in the number of studies dealing with test-taking strategies, with a decided increase in the number of related areas that have been included in the research focus.

These areas may be grouped according to three main emphases of test-taking strategy research: as a contribution to the validation of language tests, to investigate the relationship between respondents’ language proficiency and their test-taking strategies, and to evaluate the effectiveness of strategy instruction for improving respondents’ performance on high-stakes standardized tests.

Test-Taking Strategy Research for Test Validation Purposes

As Bachman (1990) pointed out, “A ... critical limitation to correlational and experimental approaches to construct validation ... is that these examine only the products of the test taking process, the test scores, and provide no means for investigating the processes of test taking themselves” (p. 269). Findings from test-taking strategy research on how learners arrive at their test responses in different contexts have increasingly been seen to provide insights for test
validation, complementing those obtained by correlational and experimental means. Such research has been used in construct validation studies, providing a new source of data for convergent validation of the construct being assessed. It has also provided insight into how given test methods, formats, and item types may affect learner responses, and how these may interact with proficiency and other contextual factors.

For example, the relationship among test-taking strategies, item content, and item performance was explored in a construct validity study of a reading test used in a doctoral study on reading strategies (Anderson, 1989, 1991). The study consisted of reanalyzing the think-aloud protocols from the Anderson doctoral thesis, adding additional strategy response categories to the 47 original ones in the thesis (Anderson, Bachman, Perkins, and Cohen, 1991). A content analysis of the reading comprehension passages and questions carried out by the test designer and an analysis based on an outside taxonomy were compared with item performance data (item difficulty and discrimination). This triangulation approach to examining construct validation indicated that strategies are used differently depending on the type of question being asked. For example, the strategies of trying to match the stem with the text and guessing were reported more frequently for inference type questions than for the other question types such as direct statement or main idea. The strategy of paraphrasing was reported to occur more in responding to direct statement items than with inference and main idea question types. 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 to examine the validation of the test in a convergent manner.

Other test-taking strategy studies related to test validation have dealt with the effects of various aspects of test method or format. One approach has been to investigate whether tests using the same language material but different formats are, in fact, assessing the same thing. For example, one study compared the effects of two test formats (free response and multiple-choice) on ESL learners' reading comprehension of the same expository science text with high lexical density (Tsagari, 1994). The tests were administered to 57 Greek ESL graduate students, together with a checklist of test-taking strategies and retrospective questionnaires concerning more general reading strategies. Two students also participated in a verbal report interview dealing with the tests. Results indicated that the two tests, with identical content but different formats, did not yield measures of the same trait. This result was further evidenced by the frequency with which students selected strategies from the checklist to describe their ways of processing the same items in the two formats.

Another kind of test-taking strategy research related to test method has been to investigate some aspect of a given test format. So, for example, a study was conducted to determine the impact of using authentic vs. inauthentic texts
in reading tests (i.e., real-life texts vs. texts written or simplified for language learners) (Abanomey, 2002). The study investigated whether the authenticity of texts may impact the way in which examinees use test-taking strategies with multiple-choice and open-ended tests, and whether there are significant differences between examinees who are reading authentic and inauthentic texts with regard to their use of bottom-up (text-based) and top-down (knowledge-based) strategies. A group of 216 adult, male Saudi Arabian EFL students were asked to respond to questions that were multiple-choice, open-ended, or a combination of both on either authentic or inauthentic texts. While text authenticity was not found to influence the number of strategies that were utilized with authentic as opposed to inauthentic texts, it did affect the manner in which examinees used the test-taking strategies. Whereas all readers used bottom-up strategies in a similar manner in reading both authentic and inauthentic texts, those responding to questions on the inauthentic texts reported using more top-down strategies (e.g., for multiple-choice: “choosing the multiple-choice alternative through deductive reasoning”; for open-ended, “writing a general answer”). The interpretation offered was that the modifications which produced the inauthentic texts also disturbed the conventional organization which made the bottom-up strategies less effective, and called for top-down strategies that called on previous knowledge (Abanomey, 2002, pp. 194–196).

Research on test-taking strategies in test validation has also looked at what particular test methods measure. One such study looked at cloze testing, examining the processes employed by subjects in Hong Kong engaged in a multiple-choice, rational-deletion ESL discourse cloze test (Storey, 1997). Verbal report protocols were obtained from 25 female Chinese students in a teacher education course. The students, working in a language lab, provided both concurrent introspection and immediate retrospection. This approach yielded data on the reasoning that the respondents employed in selecting items to complete gaps in the cloze passage, and the strategies that they used to do so. The cloze test began as follows, with the first blank calling for a discourse marker and the second for a cohesive tie:

Teachers are often unaware of the reasons for problem behavior exhibited by pupils in their classes, and, ___(1)___, inadvertently respond in ways which prolong it. Insights into the motivation behind disruptive behavior could benefit ___(2)___ in practical ways.

   a. moreover   b. nevertheless    c. as a result    d. thereafter

1. such teachers    such pupils    schools    us

The findings provided a picture of the subjects’ test-taking behaviour, as well as a cognitive perspective on the question of what cloze measures. It was
found that the multiple-choice cloze items from deleted discourse markers encouraged respondents to decompose the associated arguments and analyze the rhetorical structure of the text in some depth. The items from deleted cohesive ties were less successful in this sense, since these items could be answered locally without reference to the macro-structure of the text, although some respondents still did deeper processing of such items.

Three recent validation studies involving TOEFL iBT\(^2\) together illustrate the important role strategy studies can play in test development. One of these studies looked at test-takers' understandings of and responses to what were referred to as Integrated Reading/Writing Tasks (as these interacted with their English proficiency levels), and at related issues faced by raters. The tasks in question, on the prototype LanguEdge Courseware (Educational Testing Service, 2002) not only required comprehension of the reading text but also synthesis, summary, and/or evaluation of its content, all considered typical requirements of writing in academic settings. The study investigated how test-takers at different writing proficiency levels interpreted the demands of these integrated tasks and identified the strategies they employed in responding to them. The study drew on verbal report data from tests taken by students in Beijing, Hong Kong, and Melbourne (30 Mandarin, 15 Cantonese, and 15 Koreans), along with verbal reports by four raters in the act of rating. Score data and the student texts were also analyzed. The study described characteristics of students' responses to tasks, together with their descriptions of strategies used and raters' reactions to the student's texts. The study uncovered numerous problems with this subtest. For example, raters were found to have a major problem identifying copied language vs. students' own wordings. Furthermore, whereas the researchers had initially hoped to elicit detailed information on how students went about selecting information from the reading text, or in transforming the language of the input text into their own words, respondents had difficulty providing accurate insights as to how they had gone about producing their texts. The results of this study along with other in-house research apparently led to the removal of the subtest from the exam.

The second of these studies consisted of a process-oriented effort to describe the reading and test-taking strategies that test-takers used with different item types on the Reading section of the LanguEdge Courseware (ETS, 2002) materials developed to familiarize prospective respondents with the TOEFL iBT (Cohen and Upton, 2006). The investigation focused on strategies used to respond to more traditional, single-selection multiple-choice formats (i.e., basic comprehension and inferencing questions) vs. the new selected-response (multiple selection, drag-and-drop) reading to learn items. The latter were designed to simulate the academic task of forming a comprehensive and coherent representation of an entire text, rather than focusing on discrete points in the text.
Thus, the study set out to determine whether the TOEFL iBT is actually measuring what it purports to measure, as revealed through verbal reports. In a test claiming to evaluate academic reading ability, it was felt that the emphasis needed to be on designing tasks calling for test-takers to actually use academic reading skills, rather than being able to rely on test-wiseness tricks. Verbal report data were collected from 32 students, representing four language groups (Chinese, Japanese, Korean, and Other) as they did the Reading section tasks from the LanguEdge Courseware materials. Students were randomly assigned to two of the six reading subtests, each consisting of a 600- to 700-word text with 12–13 items, and subjects’ verbal reports accompanying items representing each of the ten item types were evaluated to determine strategy use.

The findings indicated that, as a whole, the Reading section of the TOEFL iBT does, in fact, call for the use of academic reading skills for passage comprehension — at least for respondents whose language proficiency was sufficiently advanced so that they not only took the test successfully but could also tell us how they did it. Nevertheless, it was also clear that subjects generally approached the TOEFL iBT reading section as a test-taking task that required that they perform reading tasks in order to complete it. Thus, working through the Reading sections of the LanguEdge test did not fully constitute an academic reading task for these respondents but rather a test-taking task with academic-like aspects to it. Two reading strategies found to be common to all subtests were: reads a portion of the passage carefully and repeats, paraphrases, or translates words, phrases, or sentences (or summarizing paragraphs/passage) to aid or improve understanding.

While the respondents were found to use an array of test-taking strategies, these were primarily test-management strategies. The six common test-management strategies were:

- Goes back to the question for clarification: rereads the question.
- Goes back to the question for clarification: paraphrases (or confirms) the question or task (except for basic comprehension — vocabulary and pronoun reference items).
- Reads the question and then reads the passage/portion to look for clues to the answer either before or while considering options (except in the case of reading to learn — prose summary and schematic table items).
- Considers the options and postpones consideration of the option (except for inferencing — insert text$^4$ items).
- Selects options through vocabulary, sentence, paragraph, or passage overall meaning.
- Discards options based on vocabulary, sentence, paragraph, or passage overall meaning as well as discourse structure.
In addition, the reading to learn and the inferencing items were not found to require different, more academic-like approaches to reading than the basic comprehension items. Because they now required examinees to consider words and sentences in the context of larger chunks of text and even whole passages, basic comprehension item types were found to make them reflect more academic-like tasks and elicit comparable strategies to those required by the inferencing and reading to learn tasks. It was also found that there were no significant differences across the different L1 groups (Chinese, Japanese, Korean, and Other) in terms of the use of test-taking strategies. The findings from this study on learners’ test-taking strategies would ideally lead to more precise selection and refinement of item types, so that the subtest would better approximate the construct to be tested.

Finally, an ongoing test-validation study, which has as its central focus the strategies and sources of knowledge test-takers use to respond to TOEFL iBT listening test tasks (Douglas and Hegelheimer, 2005), also illustrates an innovation in gathering test-taking strategy data. The research interest is in identifying the strategies that test-takers use to respond to the tasks on the subtest, and in identifying the linguistic and content knowledge that they use to do so. The procedures involve the use of a software application Morae (TechSmith, 2004), which allows for remotely monitoring, recording, and analyzing the data produced by users in front of a monitor, including audio recording of the verbal protocol, screen-capturing (recording everything the participants do on the computer, namely, selecting and changing answers or attempting to proceed without having completed a question), and video-capturing (recording facial expressions and note-taking behaviour).5

While the participants are working on the test, the researchers are able to watch what is happening on- and off-screen (e.g., when students take notes, refer to notes, or hesitate) and to insert comments for the post-completion interviews that they will conduct a few minutes after the participants have finished the verbal protocol. These interviews consist of the participant and the researcher together viewing the video- and screen-capture and talking about the comments that the researcher has inserted. This research marks a decided methodological refinement in the collection of verbal report data, allowing for a new level of precision and comprehensiveness, undoubtedly improving on the reliability of such data collection. In addition, subsequent coding of the think aloud protocols is assisted by the use of a qualitative analysis program, NVivo (QSR International, 2005), which allows for the insertion and then the analysis and cross-tabulation of codes reflecting categories of strategies and knowledge revealed in the verbal protocol.

While results from the study are still forthcoming, preliminary results are yielding robust descriptions of reported strategies and sources of knowledge.
for responding to the TOEFL iBT listening subtest. For example, the analysis has revealed four types of strategies for approaching the response task:

1. recalling elements of the test input including the instructions, the question, the input text, or a previous question;
2. working with the response options by reviewing them in order, narrowing the options to the two or three most plausible, and stopping the review of options without considering the rest when one is deemed correct;
3. making an hypothesis about the likely answer; and
4. referring to notes before reviewing options.

In addition, there were five main categories of reasons or sources of knowledge given by study participants for selecting/rejecting options and changing a response, plus a category for no reason given or discernable:

1. the option did (or did not) match elements of the listening text or the question, in terms of keywords, specific details, inferences about details, level of specificity, or not understanding an option;
2. they drew on knowledge outside the test context, from their own life experience;
3. they referred to their notes during the response process;
4. they referred to prior experience with multiple-choice tests, or to prior questions or part of a prior question as a guide to selecting a response; and
5. they resorted to a best guess when the participant was uncertain about the correct answer.

As in the Cohen and Upton (2006) study, the researchers will be interested in comparing the strategies hypothesized in the TOEFL framework document (Jamieson, Jones, Kirsch, Mosenthal, and Taylor, 2000) — namely, locating, cycling, integrating, and generating — to the strategies that test takers report using in responding to the tasks.

The picture that emerges from these test validation studies is that the field has progressed beyond the days when tests were validated simply by statistical analysis of correct and incorrect responses. We have progressed to a point at which we are asking crucial questions about what these tests are actually measuring and taking impressive strides to determine what it actually entails for respondents to arrive at answers to language assessment measures. The results have had an impact on the tests, even to the extent whereby the results help convince test constructors to eliminate a given test, as in the case of the Lumley and Brown (2004a, 2004b) study on an innovative subtest proposed for the TOEFL iBT.
A Focus on Test-Wiseness to Validate Tests

As a complement to the more conventional approaches to test validation, there have also been several studies looking specifically at whether it is possible to arrive at correct answers on the basis of test-wiseness rather than knowledge of the language material. A major study along these lines involved the development and validation of a test of test-wiseness for ESL students (Allan, 1992). The test that was developed included stem-option cues (where matching is possible), grammatical cues (where only one alternative matches the stem grammatically), similar option (where several distractors can be eliminated because they essentially say the same thing), and item giveaway (where another item already gives away the information). There were 33 items, each having only one test-wiseness cue in it and none intended to be answerable on the basis of prior knowledge. The students were warned that they would encounter vocabulary they had not seen before and that they could still answer the questions using skill and initiative. There were three groups of students (N = 51), with one group writing a brief explanation of how they selected their answers. The fact that the mean was well above chance (18.3, whereas 8 would be chance) suggested that the respondents did not merely guess randomly and that the test was at least to some extent measuring test-wiseness.

A more recent study (Yang, 2000) investigated the impact of test-wiseness (identifying and using the cues related to absurd options, similar options, and opposite options) in taking the test (TOEFL PBT; see note 2). First, 390 Chinese TOEFL candidates responded to a modified version of Rogers and Bate son's (1991) Test of Test-Wiseness (TTW) (see Yang, 2000, pp. 58-61) and the TOEFL Practice Test B (ETS, 1995). An item analysis of the TTW results for a subsample of 40 led to the selection of 23 respondents who were considered “test-wise” and another 17 who were deemed “test-naive.” These students were asked to provide a verbal report about the strategies that they were using while responding to a series of test-wiseness-susceptible items selected from the TTW and TOEFL. It was found that 48% to 64% of the items across the Listening and Reading Comprehension subtests of the TOEFL Practice Test B were identified as susceptible to test-wiseness. It was also found that the test-wise students had a more meaningful, thoughtful, logical, and less random approach to the items than did the test-naive students. In addition, they were more academically knowledgeable and used that knowledge to assist them in figuring out answers to questions. Finally, they extended greater effort and were more persistent in looking for test-wiseness cues, even when it involved subtle distinctions.

We need to keep performing test-wiseness studies as a means of checking whether we are giving away the answers to items more readily than we would imagine. I still remember the surprising results of a student study I reported on in my 1984 paper, where the EFL respondents received just the title of an
English passage and had to respond to multiple-choice questions about it. The more proficient students, in particular, did far too well on the items to have had it be by chance. Even some of the less proficient students almost passed the test. The items were simply too guessable.

**Language Proficiency Related to Test-Taking Strategies**

A growing body of literature has investigated the relationship between the proficiency level of the respondents, their reported use of strategies in test-taking, and their performance on the L2 tests. For example, Purpura (1997, 1998) had a total of 1,382 test-takers from 17 language centers in Spain, Turkey, and the Czech Republic answer an 80-item cognitive and metacognitive strategy questionnaire (based on the work of Oxford, 1990; O’Malley and Chamot, 1990; and others), then take a 70-item standardized language test. Purpura used structural equation modeling to examine the relationships between strategy use and second language test performance (SLTP) with high- and low-proficiency test-takers. Whereas the metacognitive strategy use and SLTP models were found to produce almost identical factorial structures for the two proficiency groups, the use of monitoring, self-evaluating, and self-testing served as significantly stronger indicators of metacognitive strategy use for the low-proficiency group than they did for the high-proficiency group (Purpura, 1999, p. 182). In addition, it was found that high- and low-proficiency test-takers, while often using the same strategies or clusters of strategies, experienced differing results when using them.

In further analysis of the data looking across proficiency levels, the researcher found that there was a continuum ranging from product-oriented to process-oriented test-takers, where the more product-oriented test-takers were seen to be able to answer questions quickly and efficiently by retrieving information from long-term memory, while the more process-oriented test-takers might be more prone to spending time trying to comprehend or remember test input, rather than simply answering the question being asked. (Purpura, 1999, p. 181)

In the appraisal of the researcher, process-oriented test-takers, regardless of their proficiency level, would be disadvantaged in timed testing situations.

A second study comparing proficiency level of respondents investigated L2 learners’ test-taking strategies in taking a listening comprehension test (again based on the work of Oxford, 1990; O’Malley and Chamot, 1990; and others). Fifty-four Japanese college EFL students took an English listening test and completed a strategy questionnaire immediately after the test (Faguchi, 2001). The questionnaire, consisting of 42 Likert-scaled items and four open-ended questions, addressed the students’ perceptions of listening strategies.
used for recovering from comprehension breakdown, compensating for non-comprehension, and reducing testing anxiety. The questionnaire also asked about the elements that caused comprehension difficulty for the students. The results of the Likert-scaled item section revealed a statistically significant difference between more-proficient and less-proficient listeners in their perceived use of top-down strategies and in their reported elements of listening difficulty, but no difference in their reported use of bottom-up strategies, repair strategies, or affective strategies. Analyses of the open-ended responses showed that proficient listeners also identified a greater range of strategies.

A third proficiency-related study was conducted in order to determine the kinds of communication strategies L2 learners use in oral interactional situations and the relationship between their use of communication strategies and their proficiency levels (Yoshida-Morise, 1998). Sample oral proficiency interviews designed by the Educational Testing Service (1982) were analyzed, focusing on the nature and number of communication strategies in the speech production of native-Japanese-speaking adult learners of English as a foreign language in Japan (N = 12). It was observed that in general the lower-proficiency respondents used more strategies and a greater variety of strategies than the higher-proficiency respondents in order to compensate for their insufficient knowledge of the target language. Nonetheless, the higher-proficiency respondents were seen to use certain strategies more, such as paraphrase, interlingual strategies, and repair strategies.

A fourth study that involved the same structural equation modeling approach used by Purpura (1997) examined the nature of text-processing strategy use and the relationships among strategy use, levels of proficiency, and levels of foreign language aptitude of Japanese university students learning English as a foreign language (Yoshizawa, 2002). The study looked at the text-processing strategies that learners reported using when they were engaged in reading or listening tasks in second language use situations, typically classrooms and testing situations. Instruments included reading and listening strategy questionnaires, the Language Aptitude Battery for the Japanese (The Psychological Corporation, 1997), and the TOEFL. Three factors emerged from the factor analysis of the test-taking strategy data:

1. comprehension and monitoring strategies,
2. compensatory strategies (translation and repair in reading, and elaboration strategies in listening), and
3. strategies related to attention and task assessment.

A fifth study relating test-taking strategies to the respondents' proficiency level involved a large-scale investigation into the relationship between use of cognitive and metacognitive strategies on an EFL reading test and success on the test (Phakiti, 2003). The study employed both quantitative and
qualitative data analyses. The 384 students enrolled in a fundamental English course at a Thai university took an 85-item reading achievement test (with both multiple-choice cloze and reading comprehension questions), followed by a cognitive-metacognitive questionnaire on what they had been thinking while responding to test items. The questionnaire was similar to that of Purpura (1999), but adjusted to suit a reading test. Eight of these students (four highly successful and four unsuccessful) were selected for retrospective interviews, which also included a 10-minute reading test (a short passage and six multiple-choice questions), to help remind them how they reported thinking through issues while performing such tests. The results suggested that the use of cognitive and metacognitive strategies had a weak but positive relationship to the reading test performance, with the metacognitive strategies reportedly playing a more significant role. In addition, the highly successful test-takers reported significantly higher metacognitive strategy use than the moderately successful ones, who in turn reported higher use of these strategies than the unsuccessful test-takers. Strategy patterns that were related to success on the reading test included reading a passage by translating it into Thai to see if it made sense and making efforts to summarize the passage as a check for comprehension.

Lest the impression be left that the use of test-taking strategies tends to have a positive impact on test results, it is important to include mention of studies that have identified strategies that may be counter-productive. So, for example, a recent study by Song (2004) with 179 ESL respondents on the Michigan English Language Assessment Battery (MELAB) revealed that while strategies such as synthesizing what was learned and linking it with prior knowledge were positively related to performance, strategies such as mechanically repeating/confirming information were not. Again, in one of the earlier studies I conducted, in this case with Aphek (Cohen and Aphek, 1979), we found one learner of Hebrew on a test of reading comprehension insisted on writing out a full translation of the Hebrew passage before he was willing to answer the open-ended questions. Not so surprisingly, he did not have enough time to answer the questions. So this translation strategy had an apparent negative impact in this instance.

**Strategy Instruction for Performance on High-Stakes Standardized Tests**

Finally, there is a limited literature addressing the issue of strategy instruction for prospective respondents on high-stakes standardized tests such as the TOEFL and the Test of English for International Communication (TOEIC). Such strategy instruction usually includes guidance in both test-management and test-wiseness strategies. One approach, for example, is to provide a set of “should do” strategies intended to help respondents perform better on such standardized tests (Forster and Karn, 1998). The strategies presented in this
document were not intended to be specific to any one section of the tests, but were intended to be applied throughout both tests. Here are just a few examples among many:

- You should not try to understand every word in a sentence; instead, you should do your best to guess the meaning of a word, and if unable to make a guess, you should skip the question. (p. 41)
- You should always be prepared to use the process of elimination to arrive at the correct answer. (p. 43)
- When considering possible answers, you must not get side-tracked by answers which are not logical inferences. (p. 44)

The existence of such documents and of institutes that provide test preparation training has also prompted studies that take a close look at the outcomes of such programs with regard to preparing students to take these standardized tests. One such study focused on the strategies used by Taiwanese students with coaching-school training when attempting a set of TOEFL reading comprehension items (Tian, 2000). Data were collected from 43 students at a coaching school in Taiwan while they did three tasks:

1. thinking aloud while attempting a set of TOEFL reading comprehension items,
2. writing down what they recalled of the passage, and
3. answering interview questions regarding their preparation for the test and their perceptions of the training received.

The verbal report data were transcribed and coded to build a taxonomy of strategies. The participants were categorized into three groups according to their scores on the set of items and then compared in their performance on the recall task, their use of strategies, and their perceptions of the coaching-school training. The taxonomy developed from these data included 42 strategies distributed in three categories — technical strategies, reasoning strategies, and self-adjustment strategies (based largely on Cohen, 1984; Sarig, 1987; Nevo, 1989).

Comparison of the high and low scorers indicated that the high scorers tended to focus on their understanding of the passages, to use the strategies taught by the coaching school only as an auxiliary to comprehension, and to stress the need to personalize these strategies. The low scorers tended to focus on word-level strategies, to use the suggested strategies as a way of circumventing comprehension, and to follow the coaching-school training mechanically. The findings from the Tian study should serve as a warning that strategy training materials may not necessarily help those who need it the most, and perhaps
most benefit those who least need assistance. This review would suggest that there are advantages in making sure that any materials developed are based to a large extent on empirical findings from process-oriented studies, rather than on the hunches of the test constructors and their associates. But this review would also suggest that even if the materials reflect honestly on the respondents’ true behaviours, it may not be easy to pass these insights on to respondents with more limited language proficiency.

Conclusion

The following then is a recap of key insights gained from twenty-five years of research on test-taking strategies.

Test Validation

- Research on test-taking strategies can serve as a valuable tool for validating and refining notions about the test-taking process. It can help us, for example, more rigorously distinguish language learner strategies on the one hand from test-taking (test-management and test-wiseness) strategies on the other.

- Empirical research on test-taking strategies can provide valuable information about what tests are actually measuring.

- Such research can also help to determine how comparable the results from different test methods and item types are — with regard to level of difficulty, the strategies elicited, and the abilities actually assessed, depending on the characteristics of the individual respondents or cultural groups.

- Research can help to determine whether performance on a given assessment measure is reflective of L2 language behaviour in the area assessed or rather represents behaviours employed for the sake of getting through the test.

Research Methodology

- Think-alouds and retrospective verbal report through interviews or other forms of face-to-face interaction, through questionnaires, through checklists, or most recently, through technologically sophisticated approaches (e.g., the advent of the software application Morae), have helped us gain a better understanding of the testing process. With regard to the collection of verbal report data for listening and speaking assessment tasks, the trade-offs between think-alouds and retrospective verbal report need to be considered — i.e., the advantages of obtaining data close to the
completion of the task vs. the threat of adversely influencing the performance by being too intrusive.

- It is beneficial to model for respondents the kinds of verbal report responses that are considered appropriate, and it may be necessary for researchers to ask probing questions during data collection to ensure the collection of fine-tuned test-taking strategy information and even have the respondents review their own verbal report for the sake of clarifying or complementing their responses.

- Test-taking strategy studies have provided insights concerning the retrospective verbal report and its advantages and disadvantages compared with think-alouds. I think this is particularly important, given the difficulty of obtaining think-alouds in tests of listening and speaking.

**Research Findings**

- Test-taking strategy studies have successfully used a variety of statistical analyses (e.g., chi-square, ANOVA, MANOVA, and structural equation modeling) to examine the relationships between strategy use and second language test performance (SLTP) with high- and low-proficiency test-takers.

- Test-taking strategy research has provided insights concerning
  - low-level vs. higher-level processing on a test;
  - the impact of using authentic vs. inauthentic texts in reading tests;
  - whether the strategies employed in L2 test-taking are more typical of first-language (L1) use, common to L1 and L2 use, or more typical of L2 use;
  - the more effective strategies for success on tests as well as the less effective ones;
  - test-takers’ vs. raters’ understanding of and responses to integrated language tasks; and
  - the items on a test that would be susceptible to the use of test-wiseness strategies.

As this review of the literature would suggest, test-taking strategy research has at present assumed a level of respectability as a viable source of information contributing to a more comprehensive understanding of test validity. While differences in test-taking strategy frameworks have resulted in some fragmentation of efforts, there is growing consensus on the importance of metacognitive strategies in test-taking, as well as the need for more fine-tuning as to their nature. Also clear is the need for a distinction between strategies for language use vs. strategies for responding to a test, since the former generally
focus on making sense out of the language material, while the latter may simply focus on getting the right answer. Researchers are increasingly aware that theory building in the area of test-taking strategies is called for if we are to develop a coherent body of knowledge.

So what possible directions can researchers take to move the debate forward at a basic conceptual level? While it was not the purpose of this chapter to refine the definitions of strategy, other efforts are afoot to do just that. An edited volume slated to appear in 2007 (Cohen and Macaro, in press) brought together leading international researchers in the L2 strategy field to provide an introspective and self-critical account of three decades of research on language learner strategies. The volume deals with definitions of language learner strategies and relates strategies to individual, group, and situational differences with regard to strategy use.

Aside from conceptual issues in strategy research, there is also the matter of how learner characteristics such as L2 proficiency level, task or method characteristics, and strategy use interact, and how all of these impact test performance. One issue in this regard that still seems unresolved is the directionality— that is, the extent to which test takers who adopt strategies to fit the demands of the assessment tasks perform better on the assessment (see Bachman and Cohen, 1998).

With regard to methods for investigating test-taking strategies, we have observed verbal report methods to emerge as a crucial tool in the process of understanding what tests actually measure. We have gone from a research situation where the very use of verbal report measures needed to be justified to the current situation where such measures are accepted as a matter of course, and the researchers can focus on how best to employ them. In addition, the picture is emerging that more proficient learners are better able to utilize test-taking strategies to their advantage than are the less proficient students. Sometimes the two groups of respondents may be using the same strategies, but there is a qualitative difference in how they use them. In addition, the work by Purpura would remind us that there will be differences among respondents across given proficiency levels according to the manner in which they approach the test (e.g., more process-oriented or more product-oriented). Finally, it would appear that strategy training might have a somewhat differential impact on prospective test-takers of high-stakes standardized tests, depending in part on their language proficiency as well as on a variety of other factors. The findings from Tian’s (2000) study would warrant follow-up research concerning this matter.

This review has demonstrated among other things that if test constructors have the extra knowledge as to what respondents actually do to produce answers to testing tasks, they are able to perform a crucial form of validation—
i.e., verifying the extent to which this behaviour is consistent with their expectations as to what was going to be assessed.\textsuperscript{7} A lingering question is whether the findings from such research on test-taking strategies have actually contributed in some way to making such tests more valid. In other words, are changes made consistent with the findings? This remains an open issue for further investigation: the impact of test-taking strategy research on the refinement of assessment measures. One would like to think that the research has a direct impact but test construction and revision is dependent on numerous factors aside from the results of research.

Whereas the research reported on in this review has focused primarily on test-taker strategies used to respond to formal L2 tests, the principles and methods are also relevant to language assessment more generally. For instance, in contexts where English language learners (or indeed any other language learners) are mainstreamed in public schooling, they are often assessed by teachers in the classroom. Given that L2 students often find teacher-made tests challenging, it could be beneficial to apply test-taking strategy research to such contexts of language assessment as well.

It is gratifying for those of us who have watched the field develop to note that it is now acceptable to include a process-oriented study of respondents' test-taking strategies when attempting to validate new tests, whether they are local, in-house measures, or standardized tests such as the TOEFL. So, test-taking strategy research has indeed come of age over the last twenty-five years. Yet there still remain numerous challenges ahead, such as in arriving at a more unified theory for test-taking strategies. Another challenge is to continue finding ways to make the research effort as unobtrusive as possible while at the same time tapping the test-taking processes. This is a particularly difficult task in the case of speaking tests, since respondents cannot simultaneously speak and provide verbal report on their speaking. Fortunately the world of technology continues to produce new, less intrusive means for collecting data, such as in the Douglas and Hegelheimer study of test-taking strategies on the Listening section of the TOEFL iBT. Such advances are likely to lead to further exciting developments in this valuable line of testing research. Stay tuned.

Notes

\textsuperscript{1} In reality, the level of conscious attention in the selection of strategies can be on a continuum from high focus to some attention to just general awareness.

\textsuperscript{2} The TOEFL Internet-based test (TOEFL iBT) assesses all four language skills: speaking, listening, reading, and writing. In comparing the TOEFL iBT with the previous versions of the TOEFL, the new version emphasizes integrated skills and provides better information to institutions about students' ability to communicate in
an academic setting and their readiness for academic coursework than does the previous TOEFL (both the paper-based test, PBT and the computer-based test, CBT). In addition it is a longer test, taking more than four hours to complete. The TOEFL iBT has a new Speaking section, which includes independent and integrated tasks. There is no longer a Structure section. Grammar is tested on questions and tasks in each section. Lectures and conversations in the Listening section are longer, but note-taking is allowed. The Reading section has new questions that ask test takers to categorize information and fill in a chart or complete a summary. The Writing section requires keyboarding.

3 The respondents were perhaps reluctant to use test-wiseness strategies because they knew we were observing their behaviour closely.

4 Items intended to measure the respondents' ability to understand the lexical, grammatical, and logical links between successive sentences by determining where to insert a new sentence into a section of the reading that is displayed to them.

5 To interject a caveat: While the use of Morae software to capture test-taker activities and actions represents a major advance in data collection techniques and facilities, the extensive data collected can apparently be very complex, and consequently difficult to interpret.

6 In the case of the Abanomey (2002) study, authentic texts were defined as those that were texts not written for a language learner audience. In this context, inauthentic texts would be those written specifically for language learners, including simplification of vocabulary and grammatical structures.

7 A tangentially related issue here is the extent to which judgments by language assessment experts as to what items are testing are reliable and valid (see, for example, concerns voiced by Alderson, 1993). Having respondents describe the processes that they actually use is a way to corroborate or refute these expert predictive judgments.
This page intentionally left blank