G161-1

Practical Considerations in Receptive Skills Testing

Judith E. Liskin-Gasparro
Educational Testing Service

Introduction

Margaret Mead has spoken of the need for an instrument of inquiry called a "macroscope." Unlike either the microscope, which makes the extraordinarily small seem large, or the telescope, which makes the extraordinarily distant seem close, a macroscope would be designed to make the extraordinarily large seem manageable.

Although Margaret Mead wanted a macroscope to serve as a way of understanding and thinking about the world, in our context such a device would be a miraculous aid to understanding and thinking about the extraordinarily complex area of language proficiency. The concept of proficiency indeed looms large and complex, including skill areas, linguistic functions, reasoning abilities, strategies for comprehension and production, and more, the investigation of each of which can and has occupied untold hours of creative thought and reams of paper.

Although we don't have a macroscope *per se,* I would like to propose the idea that testing can serve something of the same function by acting as a kind of window into the world of language learning and language acquisition. If tests are thoughtfully conceived and well constructed, they can not only assess student ability, but can also reveal a great deal about language instruction and how to make it more effective, thereby helping us to understand and manage the language learning process.

This reaction paper will focus primarily on Canale's third section, "Suggestions for improvements in receptive language testing." With the assumption that our ultimate goal will be the construction of tests measuring listening and reading proficiency, the sections that follow will address the knowledge and skills that we already possess that will be of help in that task. Specifically, we will be assisted by our understanding of the receptive skills by our knowledge about foreign language testing and test development, and by our experience with the assessment of language proficiency.

The Receptive Skills

Canale points out that much of what we know and infer about receptive language proficiency comes from research in the area of native language skills acquisition. In reading, for example, he cites the research of Curtis and Glaser (1) and Singer (2) on the stages of receptive language proficiency: students begin by decoding, reading in order to learn how to read, and only later move to a stage in which they read for information or recreation. At this later stage they bring to the reading process not only decoding skills but also world knowledge and reasoning and comprehension strategies.

This notion of stages of receptive language proficiency has implications for both teaching and testing. The reading and listening materials to which beginning foreign language students are exposed are designed to assist them in the decoding process; the language base is purposely limited to that which students have learned in order not to make too frustrating the experience of integrating the discrete pieces of language they have been taught into the comprehension of a printed or oral text. The negative side of the coin, however, is that

0161-2

all too often textbook material, particularly reading material, is presented with an ulterior motive—to reinforce the grammar topics introduced in the chapter. Even if the stated purpose of the reading selection is to present cultural information or to engage the students' attention with a bit of fiction, all too often the material is primarily a vessel for vocabulary items and grammatical structures. An analysis of the content of the material reveals it to be rather empty, not something that students would read in their native language if left to their own devices.

If students are to move in receptive language proficiency from decoding to higher-level comprehension strategies, it is important to include early on natural or minimally edited texts on topics of high interest. The subject matter would provide motivation to understand the material in the foreign language, and the fact that the text is presented much as it would be for a comparable native language audience means that students would be pushed to exercise their knowledge and the comprehension strategies that they have acquired in the native language as a result of the whole educational process.

The notion that comprehension strategies can be taught has come later to foreign language education than to the teaching of reading in the native language. The work of Swaffar, Arens, and Morgan (3) on the teaching of reading, for example, demonstrates that foreign language readers can learn to apply prereading strategies to the comprehension of a text. The authors write particularly about identifying the focus and information categories of a text before beginning to read, but clearly this idea can be extended to other prereading strategies that native language readers are taught to employ: analyzing the title, perusing the questions about content that usually follow the text, analyzing and interpreting text format (e.g. placement of the elements, headings, use of boldface type, etc.), making use of visuals, such as illustrations, tables, charts, and in longer texts, looking through the preface, table of contents, notes to the reader, etc.

The fact that there are identifiable stages of receptive language proficiency has interesting implications for testing as well. Item types can be classified according to the reading or listening skill they measure (e.g. decoding, word knowledge, grasping the main idea, following the author's argument, etc.). Students can be tested on their use of prereading comprehension strategies; in some cases these items will be identical or very similar to those that would be constructed to test comprehension of the text itself. In addition, the testing of prereading comprehension strategies can have a positive effect on instruction.

Knowledge about Testing

In addition to our understanding about the nature of the receptive skills, we also bring to the task of receptive language proficiency assessment our knowledge about testing in general. Canale cites two guiding principles that are often underemphasized in test development: acceptability and feedback potential. While large-scale standardized tests, such as the College Board Foreign Language Achievement tests (taken by some 50,000 candidates yearly) have to give priority to such features as efficiency of administration and scoring and predictive validity, smaller-scale tests can emphasize the teachers' and students' acceptance of a test as "fair, important, and interesting" as well as the ability of the test to provide feedback to teachers and students.

There is a growing interest in United States education generally for small-scale tests that can profitably be used by the classroom teacher. Teachers who understand the impact that the ACTFL Language Proficiency Guidelines can have on the curriculum and who begin to design proficiency based courses have a need for tests that will measure students' attainment of the proficiency goals. Such tests are potentially as valuable for teachers' evaluation of themselves and their programs as they are for the evaluation of student progress.

Computer-assisted instruction may well be a cause or a reflection (or perhaps both) of the current interest in classroom tests. The software packages that test, correct errors, and then teach, blur in a very healthy way the distinction between teaching and testing. Criterion-referenced classroom tests that are designed with this same potential for feedback can let students know how well they have done in meeting a particular objective and where their errors lie, and indicate to the teacher where instructional time and attention might best be invested.

Leaving aside for a moment the special needs of classroom testing and focusing on large-scale evaluation, Canale's discussion of thematic organization as an important feature of test design merits further comment. Most, if not all, national-level foreign language tests currently in use in the U.S. tend to be organized by linguistic criteria (e.g. vocabulary, structure, and reading comprehension sections in a reading test) or item types and stimulus materials (e.g. dialogues, rejoinders, visual stimuli,

0161-3

and the like in a listening comprehension test), and there is no continuity or thematic relationship between one item or stimulus and the next. Students are, unfortunately, trained to cope with and even expect this shotgun approach by their textbooks, most of which have traditionally organized exercises and activities as though the target language were composed of a series of non-sequiturs. Some encouraging change is beginning to be seen in text materials, but by and large testing continues to lag behind. There would be clear improvement in the affective impact of tests, and possibly in their measurement characteristics as well, if language were presented in larger contextual segments.

Experience with Language Proficiency Assessment

The development of assessment instruments in receptive language proficiency will clearly benefit from our experience with the testing of oral proficiency. Although there are major differences between the receptive and productive skills that affect test design, the fact that we are interested in constructing tests of *proficiency* means that we can draw on much of what we have already learned about proficiency levels and the assessment of performance. In the section on test design features, Canale discusses the four-phase approach, computer-adaptive testing, and criterion-referencing, all of which are concepts that emerge from our experience with the oral interview. One might well summarize Canale's remarks in this section by saying that the assessment of proficiency in the receptive skills requires the elicitation of a ratable sample of reading or listening.

In terms of test design, the elicitation of a ratable sample means that the test will have to allow the candidate to demonstrate the highest sustained level of his or her receptive language ability. This requires, first of all, descriptions of reading and listening proficiency levels. The ILR Testing Committee has recently completed a set of level descriptions for reading, and it is safe to assume that this work will serve as the basis for drafting listening proficiency definitions as well. The ILR Testing Committee has also proceeded with the assignment of levels to texts, a task that has significant implications for teaching as well as testing. Many teachers have only the most subjective sense of a hierarchy of difficulty levels of reading materials and item types. Having a sense of what the level descriptions mean and being able to rate the difficulty of stimulus material and the questions based on it will allow teachers to assess more accurately where their students are and how best to help them increase their proficiency.

Canale recommends applying the four-phase approach used in the oral interview to tests of receptive language proficiency. This makes a great deal of sense, and corroborates the experience of ETS in developing standardized achievement tests. It is always good to begin with easy material to counteract students' anxiety in the testing situation. The hardest material is never left for the end of a test, for experience has shown that many students faced with a long or difficult final reading comprehension selection will simply give up and not attempt it, particularly if the test is at all speeded. It works much better to flow from easy to medium difficulty to hard items, and then to end with material in the middle difficulty range.

Computer-adaptive testing, which will create a tailor-made test for each student, is the technological analogue of the human tester in the oral proficiency interview. The tester rates the student's performance throughout the interview, adapting the content of questions, the question types, and the difficulty level according to the student's responses. In computer-adaptive testing of receptive language proficiency, the computer will be programmed to do the same kind of evaluating and selecting that the human tester does in the oral interview.

Canale's sample test design, which allows students to choose a humanities or sciences context for their test, raises the intriguing possibility of a computer-adaptive test program that would select stimulus material of interest to a particular student. The student could respond to a short questionnaire about his or her areas of interest before beginning the test, and those responses would signal the computer to present a particular set of stimulus materials. As Canale has mentioned, knowledge of the subject matter is an important component of comprehension, and, depending on the circumstances, it might be important to test receptive language proficiency in familiar as well as unfamiliar subject matter.

The Challenges Ahead

As we move toward the development of receptive skills tests, there are a number of considerations and questions that bear discussion:

(1) *Level descriptions for academic use*

Building on the ETS Common Yardstick Project, the ACTFL Proficiency Guidelines used the expanded lower end of the oral proficiency scale

in developing descriptions of reading and listening proficiency. As we move to test development in the receptive skills, we may well discover that these intra-level descriptions will not be either measurable or useful for the receptive skills. If students move as quickly through the 0 and 1 ranges as we believe they do, then there will be no need to use the ACTFL/ETS subdivisions as benchmarks of progress. Simlarly, we may find that we need to reinstate some or all of the level descriptions within the Superior range to assess adequately the proficiency of high-level students.

The ACTFL/ETS nomenclature of Novice, Intermediate, Advanced, and Superior may also not prove useful in reporting proficiency levels for the receptive skills. These verbal tags correlate well with teachers' experience of students' oral proficiency at the various stages of instruction (e.g. one can reasonably expect students in intermediate courses to speak at the Intermediate level, etc.), but this same correlation does not seem to exist for the receptive skills. Most students in intermediate courses (i.e. third- or fourth-year high school or third- or fourth-semester college) can read at the Advanced level, and the denominations give a false impression of where students are at the various stages of instruction.

(2) *Item and item types*

The question of how actually to test receptive language proficiency is one that has not yet been adequately addressed. Before we get down to the practical task of item writing, we will have to answer at least these questions and probably some others as well. They can be summarized under the general rubric of "setting specifications."

—*Are we limited to multiple-choice items?*
Practical considerations would indicate that we are.

—*Will the items be in English or in the target language?*
Assessment considerations argue for English; pedagogical considerations argue for the target language.

—*What item types will we use?*
We will have to draw up a list of item types, e.g. main idea, supporting idea, inference, application, evaluation of author's logic, etc. If we are committed to a multiple-choice format, can we make that format serve our need to test "process" features such as skimming, scanning, and gisting?

—*How will we judge the difficulty level of the item types and of the individual items themselves? Can we attach some item types to particular proficiency levels?*
This is a particularly thorny question. One possible procedure would be to tentatively assign levels to a stimulus and to each of its accompanying items by matching them with the definitions. It is to be expected that items based on a particular stimulus will not all be at the same level. We would follow this up with traditional pretesting, which would give us a norm-referenced indication, based on percent correct, of the difficulty level of each item to compare with our previous criterion-referenced judgment before assigning a proficiency level to the item. Pretesting would also allow us to flag items for revision or elimination.

(3) *Validation*

Any multiple-choice test of receptive language proficiency will be to some extent an indirect measure. How will we correlate it with a direct measure of these skills?

Conclusion

Although academic and government assessment needs in receptive language proficiency have a great deal in common, and although it is important above all to focus on these commonalities, there are some differences that should not be overlooked. Government agencies most often talk about language *training;* in schools and colleges we speak of language *teaching.* In the government context one assumes that the program need provide only the language instruction, and the students will provide the rest—the reasoning ability, the native-language comprehension strategies, and the knowledge and world experience of an educated adult. In academe, we can make no such assumptions. When our students learn to read and understand a foreign language, they are often learning for the first time to observe, compare, analyze, infer, and deduce. When we produce a Level 3 or Level 4 reader, in many cases the student is developing the component skills for the first time.

In this context, the question of content for the teaching and testing of listening and reading is most important. As teachers of the humanities, we want not only to teach students to read and listen, but to guide them in deciding what to read about and what to listen for. We will want to include literature, history, culture, and the arts in the subject matter for the teaching of listening and reading. It will follow naturally that at the higher levels we

will also want to test the ability to do such things as interpret literary texts, or understand and apply the notion of the relativity of cultures. We may need to follow a somewhat different path from that of our government colleagues, even while joining with them in the testing of most aspects of listening and reading proficiency.

REFERENCES

1. Curtis, M. and R. Glaser. "Reading Theory and the Assessment of Reading Achievement." *Journal of Educational Measurement* 20 (1983): 133-48.
2. Singer, H. "IQ Is Not Related to Reading," in S.M. Wanat, ed. *Issues in Evaluative Reading*. Washington, DC: Center for Applied Linguistics, 1977.
3. Swaffar, Janet King, Katherine Arens, and Martha Morgan. "Teacher Classroom Practices: Redefining Method as Task Hierarchy." *The Modern Language Journal* 66 (1982): 24-33.
