

# ***Assessing Foreign Language Proficiency of Undergraduates***

***Richard V. Teschner***  
***Editor***



Heinle & Heinle Publishers  
Boston, Massachusetts 02116, U.S.A.

© Copyright 1991 by Heinle & Heinle. No parts of this publication may be reproduced or transmitted in any form or by any means electronic, or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Manufactured in the United States of America.

Heinle & Heinle Publishers is a division of Wadsworth, Inc.

ISBN 08384-39152

10 9 8 7 6 5 4 3 2 1

6

# Self-Assessment and Placement: A Review of the Issues

*L. Kathy Heilenman  
The University of Iowa*

One way of looking at placement is to view it as an attempt on the part of an educational institution to find the best fit between the previous preparation of incoming students and the courses and programs it has to offer them. On the surface, this task appears deceptively easy—one has only to assess students' current level of skill and/or knowledge and then identify the course or sequence that best meets their needs. This apparent simplicity, however, conceals several thorny problems. First, the assumption that a test (or an interview or an equivalency formula) can determine students' present status vis-à-vis a certain program assumes the ability to articulate that program's goals in terms that are measurable. Second, the ability to measure those goals assumes the availability of the time, commitment, and expertise necessary to construct a reliable and valid placement instrument. Third, such an instrument should, insofar as possible, reflect actual classroom tasks that learners will perform. And finally, placement procedures must operate efficiently and effectively under existing budgetary, time, and personnel constraints. The first two requirements—measurable goals and adequate test development—are possible to achieve, albeit with some difficulty. The latter two—essentially face validity/positive backwash and practicality—are frequently contradictory.

Traditionally, language departments have relied on various combinations of formulas equating secondary study to postsecondary study, personal interviews, and locally or externally developed tests to place students (Hagiwara, 1983; Klee & Rogers, 1989). Recently, however, a surge of interest in the use of self-report or self-assessment to assess learners' competencies in a second or foreign language has led to speculation about the suitability of self-assessment as a placement instrument (e.g., Dickinson, 1987; LeBlanc & Painchaud, 1985b; Oskarsson, 1978).

## **Evaluation versus Certification**

Such speculation assumes recognition of the difference between evaluation and certification. Evaluation (here self-assessment) and certification (here the determination that a student has reached a certain ability level or has fulfilled a certain requirement) are not identical (Aleamoni, 1979; Holec, 1979). That is, evaluation can occur without certification, but the reverse is not possible. As a matter of professional integrity, certifying bodies (e.g., language departments) cannot abdicate the responsibility of externally evaluating and thus certifying or not certifying a certain level of competence in students who successfully meet their requirements.

Nevertheless, current practice frequently uses the same test both to certify proficiency or achievement at a certain level (or to exempt students from certain courses) and to place students in courses below that level. It is obvious that self-assessment cannot be used for certification purposes, since it pits students' best interests against their integrity (Dickinson, 1987; Painchaud, 1989; Upshur, 1975). A reasonable proposal, then, would be to use a more extensive and well-developed direct and indirect testing program for those relatively few students presenting themselves for certification (or exemption) and to reserve a self-assessment instrument for the majority of students who need to be placed in courses.

Such a scheme effectively removes the temptation to better one's lot; it does not, however, prevent students from deliberately underestimating their abilities in order to receive credit for a first-semester language course. This problem seems to be one whose solution lies less in test development than it in educational policy and practice. It would be possible, for example, to simply disallow "starting over" as a blanket policy. Students would then have the option of continuing in the language at a more advanced level or "beginning at the beginning" in another language. Alternatively, of course, such students could be

shifted into a special review course designed for false beginners (Klee & Rogers, 1989; Loughrin-Sacco et al., 1990).

## **Self-Assessment and Placement**

Given acceptance, then, of the difference between evaluation and certification, the advantages of self-assessment in the context of placement are evident. First, such tests are efficient and economical. There is little or no concern with test security and, as long as there is no incentive for learners to either over- or underestimate their abilities, dishonesty is not a problem. In addition, a self-assessment questionnaire can efficiently sample more language behavior more quickly than a direct or even indirect test of such behavior can. Thus, areas that are normally left untested because of time constraints (speaking and writing) or lack of facilities (listening) can become part of the placement procedure.

Increased learner involvement is a second advantage. Learners whose opinions and judgments have been sought and valued are less likely to feel manipulated by what often seem arbitrary placement levels decided by machine-mediated procedures. As Canale (1985, p. 250) has pointed out, language testing is all too often "a crude, contrived, confusing, threatening, and above all intrusive event," with learners cast as "obedient examinees" rather than active participants. Self-assessment, on the other hand, has the potential of being both humane and learner-centered, with students being asked to participate in a process rather than being dictated to as the result of a product.

Finally, a self-assessment instrument will communicate a program's goals and expectations much more effectively than a decontextualized test score will. The process of developing an effective self-assessment instrument will, of necessity, involve faculty in an internal discussion of what they expect learners to achieve at various levels. Such discussion, in turn, will cause curricula and tests to be more closely and obviously linked.

Yet before rushing to embrace self-assessment as the magic solution to all problems, several questions must be raised. First, what exactly does it mean to ask learners to self-assess? Second, is it possible to produce self-assessment instruments that are reliable, valid, and instructionally sound within the context of placement? Third, what evidence do we have that learners can indeed accurately assess their capabilities? And fourth, does the existing literature provide guidance in the development of a self-assessment placement test and/or give examples of such instruments?

## Self-Assessment: What Is It?

Self-assessment involves asking learners to make judgments about their own abilities. That is, instead of asking students to write a narrative about last weekend and then evaluating it, holistically or otherwise, students are asked how easily or how well they *could* write such an essay if they were asked to. Or, instead of using multiple-choice questions over a reading passage, students are asked to look at the passage and respond “yes,” “probably,” or “no” to the question, “I can understand the basic ideas.” (See the Appendix for examples of various self-assessment formats.)

In other words, a self-assessment instrument is an overtly indirect measure of language ability. In addition, it is a self-report of a belief or judgment concerning one’s own language behavior. As such, self-assessment can deal with potentially observable behaviors (e.g., how well students can narrate last weekend’s activities) or with constructs that for all intents and purposes are difficult to observe (e.g., the ability to speak German in various contexts).

Such self-reports or assessments have long been common in educational, psychological, and social research. Self-reports have been used to facilitate data collection (e.g., asking people their age), to measure behavior (e.g., the number of times a doctor was visited within the past month), to assess attitudes (e.g., anxiety, satisfaction with life), and to solicit opinions (e.g., for or against animal rights). Given the effort, reports of behavior can be verified by objective observation: birth certificates can be checked or an observer can be posted in the doctor’s office. Attitudes and opinions, however, cannot—one cannot directly observe “anxiety” or “pro-animal rights.” It is possible, though, to measure and compare other, equally indirect indices. Thus, people who rate themselves as “very anxious” will be expected to behave in ways that are commonly assumed to denote anxiety (worried expressions, increased heart rate, observed anxiety attacks, etc.); people who declare themselves pro-animal rights will be more likely to pat than to kick stray dogs; and so forth. In other words, reports of private states and beliefs will be borne out by public and observable behaviors (Evans, 1986).

In terms of language, such self-report measures as “experience in a Spanish-speaking country” are common. Likewise, the use of self-report to gather data on such attitudes and opinions as “foreign language anxiety” or “attitude toward speakers of Japanese” is also well accepted. Less familiar, though, are self-assessment measures of language ability. Such measures do, however, hold promise in terms of sampling power, efficiency, and student involvement; they should be seriously evaluated.

In general, researchers and testing and evaluation experts have been suspicious of the "subjective" nature of self-report data and have proceeded to validate such measures against extrospective, observable, direct, and indirect measures. But there is no logical reason for self-report data to be any more subject to experimental contamination than data gathered by the "objective" sampling of a particular domain (Howard, 1981). In fact, the case can be made that self-report data, which effectively sample respondents' total experience, should be used as the criterion for less face- and content-valid indirect measures (Barrows et al., 1981). Finally, as Howard and his colleagues (1980) argue, the fact that in many cases self-report measures are found to correlate rather modestly with a behavioral criterion measure casts doubt on both, since error variance in either or both measures could be contributing to the low correlation.

## **Characteristics of Self-Assessment as a Test**

It is important to realize that self-assessment refers to a type of measurement rather than to a particular testing instrument. There are many possible ways to gather learners' perceptions of their abilities, some of which are discussed later (see the Appendix for examples). In addition, the items included in a self-assessment instrument will differ according to situational constraints and demands. Thus, it is not possible to talk about the reliability or validity of self-assessment instruments in a general sense. There are likely to be valid and invalid self-assessment instruments, reliable and unreliable self-assessment instruments, just as is the case in other testing formats. It is possible, however, to discuss the advantages and disadvantages of self-assessment in regard to reliability and validity and to survey the existing literature for evidence of test characteristics of particular self-assessment tests.

### **Reliability**

Reliability is the degree to which a measuring instrument is accurate and consistent (Aleamoni, 1979; Hughes, 1989). Thus far, reliabilities reported for self-assessment instruments have been high, ranging from .54 to .96, with the majority being greater than .80 (see Bachman & Palmer, 1981, 1989; Davidson & Henning, 1985; Heilenman, 1990; Hilton et al., 1985; Weltens et al., 1989).

### **Validity**

Validity, or how well a test does what it is supposed to do (Oller, 1979), can be viewed in several ways. Face validity, the acceptability of the test

to those involved, seems to pose little or no problem. In general, learners react positively to self-assessment items (e.g., von Elek, 1985), although there has been one report of a small number of students who felt that "someone" should place them (LeBlanc & Painchaud, 1985b).

Content validity, in reference to placement tests, requires evidence of the extent to which a test reflects the content and task types found in the curriculum it serves. This aspect is obviously a function of test development but does not seem to pose any particular difficulties. In fact, self-assessment alleviates the problem of not being able to test certain areas because of time or equipment constraints, thereby increasing the number of potential domains that can be sampled.

Construct validity focuses on the test score as a measure of an underlying construct, and, all else being equal, is less important within the pragmatic context of a placement test than it might be elsewhere (Schaefer, 1982). Instead, for a placement test it is more important to consider predictive validity (how well a test predicts later performance) and concurrent validity (how well a test agrees with the results of other, accepted measures). The issue of predictive validity is addressed by several studies. LeBlanc and Painchaud (1985b) and Painchaud (1989) describe an ongoing placement program at the University of Ottawa that depends on self-assessment for placement. They report that changes in placement are less frequent using self-assessment than previously. In addition, Jannssen-van Dieten (1989) describes pilot work indicating that had self-assessment scores been used for placement, only 3 of the 25 persons involved would have been placed in a group different from that indicated by the criterion measure. Finally, Heilenman (1989), using four 11-point scales, asked university students of French to rate their global ability in listening, reading, writing, and speaking ( $n = 327$ ). Overall differences among levels (first semester, second semester, second year, above second year) were significant ( $F = 27.24, p < .001$ ). Post hoc tests, however, indicated no significant differences between judgments of students enrolled in the second semester and those of students enrolled in the second year. In other words, although overall student judgments could differentiate among levels, this effect was primarily due to differences between judgments of students at beginning levels and those of students at more advanced levels.

The concurrent validity of self-assessment instruments has been addressed by a number of studies comparing performance on a self-assessment instrument with performance on another measure of language ability. Given the variety of learners tested, the wide range of self-assessment instruments used, and the various criterion measures chosen, it is difficult to draw any general conclusions. Nevertheless, correlation coefficients between self-assessment instruments and other

measures seem to cluster between .30 and .60, with a few reaching around .90 and a few being very low (for reviews, see Blanche, 1988; Blanche & Merino, 1989; Heilenman, manuscript). According to Aleamoni (1979) concurrent validity coefficients frequently lie between .40 and .68, and so this range is not surprising. Another way of looking at the issue, however, is to view these correlations as neither (1) low enough to reject completely various self-assessment instruments as substitutes for other measures nor (2) consistently large enough to accept certain self-assessment instruments as totally satisfactory alternatives. Nevertheless, as Oller (1979) has pointed out, low correlations in and of themselves do not indicate that two tests are not measuring the same thing; nor do they in and of themselves establish low concurrent validity. Such correlations may result from, among other possibilities, inherently poor measures, low reliability, or a restricted range of scores (Hatch & Farhady, 1982). It is encouraging, however, that the most consistently high correlations with behavioral criterion measures are found for those self-assessment instruments which have a history of conscientious test development (e.g., see Barrows et al., 1981; Hilton et al., 1985; LeBlanc & Painchaud, 1985b).

### **Instructional and Curricular Concerns**

Beyond a concern with reliability and validity, test developers need also to consider a test's effects in terms of utility, feasibility, and fairness (Shohamy, 1990). Utility involves ascertaining that a test is indeed useful in a practical sense. For a placement test, this element would involve the prompt and useful reporting of scores, the backwash effect of the test on teaching and curriculum, and perhaps the possibility of acquiring diagnostic information. Here there seem to be no real disadvantages and several actual advantages in the use of self-assessment. If the self-assessment placement test is constructed so as to faithfully reflect the goals of a program, then instructors can be supplied with students' assessments in various areas in order to plan instruction better. Similarly, by having program goals and objectives that are clearly defined and delineated as part of an open placement procedure, teachers at the secondary level can make more informed decisions regarding the articulation of their programs with those of postsecondary institutions.

Both feasibility and fairness are also issues that self-assessment procedures can address satisfactorily. As already outlined, self-assessment instruments are easier to administer, take less time, and are more cost-effective than comparable direct and indirect measures of language ability. Finally, fairness—whether a test is ethical, legal, and in the best interests of the learner—would not seem to be a major issue, except

perhaps in the case of students who feel that such assessment is more properly and correctly done by those in charge of instruction.

## **Accuracy of Learner Self-Assessments**

The question of whether learners can accurately assess their language abilities depends on the manner in which accuracy is defined. There is, as Barrows et al. (1981) point out, no logical reason to think that the evaluation provided by a particular extrospective test is more accurate than the assessment given by learners themselves. Nevertheless, if the assumption is made that instructor ratings and/or standardized test scores represent a closer approximation of the so-called true value of learners' language proficiency, then it makes sense to assess such accuracy by comparing learner self-assessment with instructor judgments or test scores. To the extent, then, that learners' self-assessments approximate these measures, learners will be said to be accurate. To the extent to which they diverge, learner self-assessments will be said to be under- or overestimations of ability.

Evidence provided in several studies is mixed, with some pointing to quite good matches between self-assessment and other measures (e.g., Barrows et al., 1981; Hilton et al., 1985; LeBlanc & Painchaud, 1985b; Oskarsson, 1981; von Elek, 1985), while others find that learners, particularly less proficient ones, tend to overestimate their abilities (Anderson, 1982; Blue, 1988; Janssen-van Dieten, 1989; Oller & Perkins, 1978; Wesche et al., 1990). Heilenman (1990) found clear evidence of overestimation among less proficient learners, thought because of the research design it was not possible to measure the degree of overestimation. It should be noted here that a tendency for beginners to overestimate is a not-uncommon finding in other areas (Garhart & Hannafin, 1986; Olson & Martin, 1980; Pohl, 1982; Reed, 1988) and may simply reflect such learners' inexperience with what it means to be proficient. Note also that, for placement tests, a tendency toward overestimation may be reduced since, as Martin (1984) points out, tendencies toward favorable self-presentation in surveys (comparable here to tendencies toward overestimation) are minimized when subjects are aware that their claims will be verified by external evidence. That is, learners who are aware that their judgments will help to determine the class into which they will be placed may tend to be less likely to overestimate their abilities.

In summary, it seems that overestimation on the part of less proficient learners may serve to mitigate the usefulness of self-assessment as a placement procedure. Still, the fact that such overestimation is not a general finding (as well as the possibility of controlling or minimizing it

through instructions and question selection/wording) should encourage test developers to proceed with some caution but to proceed nevertheless in investigating self-assessment as a useful placement procedure.

## **Developing a Self-Assessment Placement Instrument**

Many of the steps for developing a self-assessment placement instrument will be quite similar to those for developing any language test. The following outline of test development is based on materials provided in Aleamoni (1979) and Hughes (1989), with specific suggestions regarding self-assessment.

### **Step 1: State the Problem/Define the Purpose**

#### *Placement/Certification*

Will this be a placement test or will it serve as both a placement and an exemption instrument? If the latter, self-assessment is not a viable option, since students will find themselves forced to weigh integrity against self-interest. To paraphrase Upshur (1975, p. 58), it is neither fair nor reasonable to ask people to in effect cut their own throats.

#### *Results*

How detailed and accurate must results be? It will in all likelihood be easier to develop a self-assessment instrument (or other test) that will divide students into three groups rather than into four or into four groups rather than eight. Will a global score suffice or is it necessary to sort students according to abilities in particular areas (conversation, writing, etc.)? How difficult will it be for students to change their placement? It is probably unrealistic to expect any test to place all students accurately, and in fact there is likely to be a significant amount of overlap in score distributions between contiguous quarters or semesters (Dizney & Gromen, 1967; Heilenman, 1983). Thus, it will be important to establish and monitor the degree of change or number of placement misses that are acceptable. For example, if it is relatively easy to move from level to level (e.g., classes are not full), then less accuracy is needed than if the reverse is the case.

On the other hand, if movement among levels is administratively feasible, then one might consider a scheme something along the lines proposed by Shaw (1980), whereby students are provided with a package of materials and encouraged to visit classes before settling on the appropriate level.

***Backwash***

How important is the backwash effect? Self-assessment instruments lend themselves well to positive backwash effects because there is little concern for test security and since they can efficiently sample domains normally difficult to test. As long as care is taken in the selection of content, a positive backwash effect is to be expected.

***Constraints***

Are there constraints resulting from lack of time (for construction of the test as well as for administration and scoring), expertise, or personnel? Valid self-assessment instruments are no less time-consuming to construct than more traditional measures are. They are, however, less mystifying. As a result, faculty input and effort may be easier to procure than is usually the case. From an administrative point of view, once concern with test security is eliminated, several possibilities open up. Self-assessment placement tests could be sent to students before they enroll (cf. Painchaud, 1989). Similarly, secondary students could complete such instruments during their last language class rather than waiting to arrive at the college or university. In addition, it seems feasible to adapt such instruments to computer administration and scoring so that students could be evaluated on a walk-in basis.

**Step 2: Determine Instructional Objectives and Curricular Goals**

This may well be the most problematic step in the process. As Byrnes (1990, p. 75) has pointed out, foreign language curricula are at present in an untidy state of "unarticulation" and/or "disarticulation" that makes it difficult to clearly specify the content of courses and programs in terms useful for either testing or placement. This situation is less the fault of language faculty than it is a reflection of a lack of consensus on a conceptual plane. Jakobovits' assessment of the situation in 1970 bears repeating today:

The question of what it is to know a language is not yet well understood and consequently the language proficiency tests now available and universally used are inadequate because they attempt to measure something that has not been well-defined (p. 75).

(See also Hagen, 1990, and Stevenson, 1981, for similar views.)

Luckily for those involved in placement testing, this state of affairs, although troublesome, is not fatal. As Schaefer (1982, pp. 75-76) has put it:

A [placement] test used in the real world to make practical decisions is primarily justified not by its theoretical foundations but by the

degree to which it improves the decision-making process, making it more effective or more efficient.

This does not mean that theory can be ignored. Developing a placement test, however, does not necessarily entail validating a construct; nor does it necessarily imply the development of a curriculum. It does mean identifying those features which differentiate among various course levels. Such information may be gathered by looking at written course objectives and course achievement tests, as well as by polling instructors and inspecting materials used. Of course, insofar as the people involved in test development are the same people involved in instruction, test development and curriculum development are likely to proceed in tandem. The advantage in such a case is that many unspoken assumptions on the part of the teaching staff are likely to be made explicit and become subject to discussion, thus leading to better-articulated goals and objectives as well as to closer ties between testing and curriculum.

### **Step 3: Determine the Content for the Self-Assessment Instrument**

A set of test specifications is needed such that many possible tests could be developed. Actual test items will be samples of these specifications. At a minimum, the following should be specified: tasks students are expected to carry out, types of texts (oral and written), and people to whom student output is to be addressed (Hughes, 1989). An example of how this task might be done is given by LeBlanc and Painchaud (1985a, 1985b) in their description of the development of a placement test for which reading and listening were the skills to be evaluated. LeBlanc and Painchaud asked instructors to write descriptors they felt were representative of each of the six existing levels of reading and listening comprehension courses. This effort produced more than 1,000 descriptors that varied quite widely. It was nevertheless possible to establish rough levels. Subsequently, a matrix was developed, taking into account types of texts, genres, registers, length, and type of presentation (speed, number of repetitions for oral texts). The result was formal descriptions of each level that could be used to develop actual self-assessment items. Reproduced below is the description for level 2 listening (LeBlanc & Painchaud, 1985b, p. 683):

Is able to understand, in a dialogue of the social conversation genre between two students at the familiar register and spoken at regular speed, the topic of the conversation with one or two details.

At this point, it will also be necessary to establish test format (types of items, scales, multiple-choice versus forced choice or free response,

etc.), time to be allowed, difficulty level, and scoring procedures. Here, choices made may have consequences for the test's reliability or validity. Outlined below are suggestions based on the literature concerning the construction of survey, interview, and self-report instruments (see also Molenaar, 1982). In the case of self-assessment of language skills, however, much work remains to be done as to how these issues impinge on the reliability, validity, and accuracy of self-assessment of second language abilities.

### *Item Type*

Items may consist of global evaluations of ability or may focus on discrete behaviors. Global questions may ask for an overall evaluation of ability or may query distinct skill areas. Questions focusing on discrete behaviors may ask learners to assess their abilities by describing the task or may give learners actual examples (see the Appendix). Although studies asking for global judgments have produced acceptable correlations with other measures (Oskarsson, 1981; Wangsotorn, 1981), studies that have directly compared the use of global scales with that of behavioral descriptions indicate that the latter are preferable (Janssen-van Dieten, 1989; LeBlanc & Painchaud, 1985b). Taken in conjunction with the advice commonly given survey writers to avoid overly general, vague questions (Turner, 1984), the exclusive use of global questions would seem inadvisable.

Learners have been asked to assess their abilities using descriptions of tasks (e.g., LeBlanc & Painchaud, 1985b), as well as to give their judgments after looking at actual samples of the task (e.g., Janssen-van Dieten, 1989; von Elek, 1985). To date, there has been no direct comparison of these two formats. Learners asked to judge their abilities based on a concrete task sample may well be more accurate than learners whose judgments are based on what they interpret the task to be. On the other hand, by using actual texts (written and oral), test developers run the risk of having learners' judgments focus on the characteristics of that one task *sample*, rather than on the more general question of how they deal with that *type* of task.

### *Type of Scale Used*

Learners may be asked to respond using labeled rating scales calibrated with varying numbers of points. Technically, such scales produce ordinal- rather than interval-level measurement, thus limiting the statistical analyses that can be performed (for discussion of this general issue, see Bass et al., 1974; Borgatta & Bohrnstedt, 1980; Newstead & Collis, 1987). From a practical standpoint, however, the majority of scales can be treated as if they produced interval-level data (Guilford, 1985; see

Bass et al., 1974, for examples of statistically optimal scales). On the other hand, scales for which the underlying construct is not logically measurable on an interval scale (e.g., the ACTFL-ETS proficiency ratings), scales that are restricted to dichotomous answers (“yes” or “no”), and scales that deliberately have more positive labels than negative ones (or vice versa) should be treated as producing ordinal rather than interval data.

### *Order of Items*

Items may be presented in a random order or they may be ordered in a Guttman-like scale ranging from least to most difficult (see Barrows et al., 1981, and Hilton et al., 1985, for examples). The use of the latter may help learners to assess their abilities more realistically, since in effect it provides a sort of macroscale along which learners can align their judgments (Schwarz & Hippler, 1987). In other words, beginning learners will be less likely to assess themselves as highly capable on items toward the difficult end of the scale simply because they will realize that beginners should not be able to do things toward the top of the scale. On the other hand, the use of such an ordered scale implies the ability to establish a satisfactory rank order of difficulty prior to and independent of the self-assessment instrument.

### *Instructions*

Within the context of survey/interview methodology, Cannell et al. (1981) emphasize the importance of clarifying the purpose of the test as well as attempting to increase the respondents’ commitment to give accurate and well-thought-out answers. For a self-assessment placement instrument, students should be given a rationale for the use of self-assessment, told that items represent a range of abilities, and encouraged to use the complete range of the scale. Practice items, perhaps including one or two filled out by fictitious learners at various levels, could be provided.

### *Language of Test*

If at all possible, the language of the test should be the students’ first language. Asking learners to rate themselves using a language in which they are not proficient increases the risk of inaccurate self-assessment (Oller & Perkins, 1978). If this arrangement is not possible, questions should be extensively pretested, including having actual students read and revise them. Reid (1990) reports that using small numbers of structural and vocabulary patterns seems to produce a more understandable, if rather repetitious, instrument.

**Step 4: Item Writing**

Writing the actual items is demanding. Test developers will probably want to write many more items than they think will be needed in order to eliminate those which are unclear or otherwise troublesome. Here, developers of self-assessment instruments will want to take heed of experience incurred in writing survey, interview, and other self-report measures.

*Avoid Negatively Worded Items*

Such items are often found to introduce error owing to respondents' difficulties in processing (Schmitt & Stults, 1985). Bachman and Palmer (1989) may be right in their contention that learners are better able to discern what they cannot do than to report what they can do. Nevertheless, inquiring about a lack of skills confounds this issue with that of negatively worded questions and should be avoided. (For further discussion, see Heilenman, 1990.)

*Attempt to Write Items That Are Directly Relevant to Learners' Experiences*

If at all possible, questions asking students to speculate about what they *might* be able to do but have never before done (e.g., order a meal in a restaurant, buy a train ticket) should be avoided. Learners whose experience is largely limited to the classroom are unlikely to be able to give accurate or even well-informed judgments in these cases. (For further discussion and examples, see Heilenman, 1990.)

*Allow Sufficient Time for Revision*

The development of a self-assessment instrument appears simple. In reality, however, it may prove difficult to write items that consistently elicit the expected responses. Wording effects (changes in response caused by changes in wording) can present unpredictable and frustrating challenges. It is also dangerous to assume that the frame of reference of the test developers (usually experienced language learners and teachers) is shared by learners. Students may legitimately judge themselves able to "hold a conversation" based on their ability to do so within the sheltered confines of their classrooms. On the other hand, test developers may be quite aware of these same students' limitations (see Heilenman, 1990, for discussion and illustrations). In sum, the use of self-assessment in placement is not a substitute for the time and effort required to develop a valid and useful testing instrument. Many of the problems are the same; some are different. The time involved, however, is likely to be substantial.

### **Step 5: Pretest**

This is essential for self-assessment instruments in which ambiguities, differing frames of reference, and question vagueness may combine to produce error and confusion. It may be extremely helpful to discuss a preliminary draft of the test with a stratified, random sample of the students for whom it is destined. Open discussion of questions among colleagues may also reveal unforeseen problems. In addition, Converse and Presser (1986) suggest the following:

- 1) Create split-sample (split-ballot) comparisons. That is, ask the "same" question twice using different wording. Then compare learners' performance on the supposedly identical questions.
- 2) Use follow-ups to closed questions. Ask learners to explain their answers.
- 3) Use multiple indicators. If the ability to read literary works is an important goal, ask students several questions in this area.

Of course, it is assumed that appropriate statistical analyses will be performed (item analysis, reliability, and validity), either at this point, if sufficient data is available, or later, as part of the first administration (Hughes, 1989). Here, too, decision scores can be determined (Aleamoni, 1979).

### **Step 6: Develop a Plan for Evaluation, Review, and Modification**

Are students actually being placed satisfactorily? How many students were misplaced and had to be moved? Were any students obviously underplaced? It may be advantageous to look more closely at those cases of misfit with an eye toward later test modification. Finally, all tests should be reviewed periodically and formally revalidated on a regular basis. Aleamoni (1979) suggests a three-year cycle of yearly review and revalidation.

## **Conclusion**

Two final questions remain. The first returns to the problem of construct validity, or "the extent to which scores are consistent with theoretical expectations" (Yaremko et al., 1986, p. 40). Although, as already pointed out, placement tests are most appropriately validated against either the accurate assignment of students to courses (Hughes, 1989) or perhaps against another test whose validity is satisfactory (i.e., predictive and concurrent validation), concern with construct validity is not misplaced.

What exactly is a self-assessment or self-report instrument measuring? A commonsense approach would be to say that it is measuring respondents' perception of their own abilities based on self-observation in a variety of situations. But as Evans (1986) points out, perceived competence also reflects self-image and will not necessarily be congruent with another measure of reality. Thus, self-assessment of language ability may reflect learners' self-confidence, experience, ability to judge, or other factors not necessarily directly related to their language ability (Oskarsson, 1978). Wesche et al. (1990), for example, found a substantial negative correlation between scores on a scale constructed to measure anxiety in using French and self-assessment scores. This finding, however, is preliminary and should not be exaggerated; it could result from the fact that both the self-assessment instrument and the anxiety scale were Likert-type instruments, with the correlation being, then, at least partly, the result of a method effect (cf. Campbell & Fiske, 1959). Another possibility is that lack of anxiety in language use is actually related in predictable ways to success in using that language, a supposition supported by MacIntyre and Gardner's (1989, pp. 272-73) demonstration of "a clear relationship . . . between foreign-language anxiety and foreign-language proficiency."

The second question is more practical and concerns the use of self-assessment within ongoing language programs. Dickinson and Carver (1980) point out the many advantages of helping students learn how to learn. To become truly self-directed learners, however, students have also to become successful monitors and evaluators of their own progress (Dickinson, 1987). In this sense, the development of self-assessment instruments for use in placement represents a step toward a larger goal.

## Works Cited

- Aleamoni, Lawrence M. *Methods of Implementing College Placement and Exemption Programs*. Princeton, NJ: College Entrance Examination Board, 1979.
- Anderson, Pamela L. "Self-Esteem in the Foreign Language: A Preliminary Investigation." *Foreign Language Annals* 15 (1982): 109-14.
- Bachman, Lyle F. & Adrian S. Palmer. "The Construct Validation of Self-Ratings of Communicative Ability." *Language Testing* 6 (1989): 14-29.
- \_\_\_\_\_. "The Construct Validation of the FSI Oral Interview." *Language Learning* 31 (1981): 67-86.
- Barrows, T., S. M. Ager, M. F. Bennett, H. I. Braun, J. L. D. Clark, L. G. Harris & S. F. Klein. "College Students' Knowledge and Beliefs: A Survey of Global Understanding. The Final Report of the Global Understanding Project." New Rochelle, NY: Change Press, 1981. ERIC ED 215 653.

- Bass, Bernard M., Wayne F. Cascio & Edward J. O'Connor. "Magnitude Estimations of Expressions of Frequency and Amount." *Journal of Applied Psychology* 59 (1974): 313-20.
- Blanche, Patrick. "Self-Assessment of Foreign Language Skills: Implications for Teachers and Researchers." *RELC Journal* 19 (1988): 75-93.
- \_\_\_\_\_ & Barbara J. Merino. "Self-Assessment of Foreign-Language Skills: Implications for Teachers and Researchers." *Language Learning* 39 (1989): 313-40.
- Blue, George M. "Self-Assessment: The Limits of Learner Independence." *ELT Documents* 131 (1988): 101-18.
- Borgatta, Edgar F. & George W. Bohrnstedt. "Level of Measurement Once Over Again." *Sociological Methods and Research* 9 (1980): 147-60.
- Byrnes, Heidi. "Addressing Curriculum Articulation in the Nineties: A Proposal." *Foreign Language Annals* 23 (1990): 281-92.
- Campbell, Donald T. & Donald W. Fiske. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56 (1959): 81-105.
- Canale, Michael. "Language Assessment: The Method is the Message." *Georgetown University Round Table on Languages and Linguistics 1985*. Ed. Deborah Tannen & James E. Alatis. Washington, DC: Georgetown University Press, 1985: 249-62.
- Cannell, Charles F., Peter V. Miller & Lois Oksenberg. "Research on Interviewing Techniques." *Sociological Methodology*. Ed. Samuel Leinhardt. San Francisco: Jossey-Bass, 1981: 389-437.
- Clark, John L. D. & Eleanor H. Jordan. "A Study of Language Attrition in Former U.S. Students of Japanese and Implications for Design of Curriculum and Teaching Materials." 1984. ERIC ED 243 317.
- Converse, Jean M. & Stanley Presser. *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills, CA: Sage, 1986.
- Davidson, Fred & Grant A. Henning. "A Self-Rating Scale of English Difficulty: Rasch Scalar Analysis of Items and Rating Categories." *Language Testing* 2 (1985): 164-79.
- Dickinson, Leslie. *Self-Instruction in Language Learning*. Cambridge: Cambridge University Press, 1987.
- \_\_\_\_\_ & David Carver. "Learning How to Learn: Steps towards Self-Direction in Foreign Language Learning in Schools." *ELT Journal* 35 (1980): 1-7.
- Dizney, Henry F. & Lauren Gromen. "Predictive Validity and Differential Achievement on Three MLA-Cooperative Foreign Language Tests." *Educational and Psychological Measurement* 27 (1967): 1127-30.
- Evans, Ian M. "Response Structure and the Triple-Response-Mode Concept." *Conceptual Foundations of Behavioral Assessment*. Ed. R. O. Nelson & S. C. Hayes. New York: Guilford Press, 1986. 131-55.
- Garhart, Casey & Michael Hannafin. "The Accuracy of Cognitive Monitoring during Computer-Based Instruction." 1986. ERIC ED 267 768.
- Guilford, J. P. "A Sixty-Year Perspective on Psychological Measurement." *Applied Psychological Measurement* 9 (1985): 341-49.

- Hagen, L. Kirk. "Logic, Linguistics, and Proficiency Testing." *ADFL Bulletin* 21, 2 (1990): 46-51.
- Hagiwara, M. Peter. "Student Placement in French: Results and Implications." *Modern Language Journal* 67 (1983): 23-32.
- Hatch, Evelyn & Hossein Farhady. *Research Design and Statistics for Applied Linguistics*. Rowley, MA: Newbury House, 1982.
- Heilenman, L. Kathy. Self-Assessment of Second Language Ability. Manuscript.
- \_\_\_\_\_. "Self-Assessment of Second Language Ability: The Role of Response Effects." *Language Testing* 7 (1990): 172-98.
- \_\_\_\_\_. "The Use of a Cloze Procedure in Foreign Language Placement." *Modern Language Journal* 67 (1983): 121-26.
- \_\_\_\_\_. "Use of Self-Assessment in Placement Testing." Paper presented at the Modern Language Association, Washington, DC, 1989.
- Hilton, Thomas L., Jerilee Grandy, Roberta Green Kline & Judy E. Liskin-Gasparro. *Final Report: The Oral Language Proficiency of Teachers in the United States in the 1980s—An Empirical Study*. Princeton, NJ: Educational Testing Service, 1985.
- Hippler, Hans-J. "Response Effects in Surveys." *Social Information Processing and Survey Methodology*. Ed. Hans-J. Hippler, Norbert Schwarz & Seymour Sudman. New York: Springer-Verlag, 1987: 102-22.
- Holec, Henri. *Autonomy and Foreign Language Learning*. Oxford: Pergamon Press, 1979.
- Howard, George S. "On Validity." *Evaluation Review* 5 (1981): 567-76.
- \_\_\_\_\_, Scott E. Maxwell, Richard L. Weiner, Kathy S. Boynton & William M. Rooney. "Is a Behavioral Measure the Best Estimate of Behavioral Parameters? Perhaps Not." *Applied Psychological Measurement* 4 (1980): 293-311.
- Hughes, Arthur. *Testing for Language Teachers*. Cambridge: Cambridge University Press, 1989.
- Jakobovits, L. A. *Foreign Language Learning: A Psycholinguistic Analysis of the Issues*. Rowley, MA: Newbury House, 1970.
- Janssen-van Dielen, Anne-Mieke. "The Validity of Self-Assessment by Inexperienced Subjects." *Language Testing* 6 (1989): 30-46.
- Klee, Carol A. & Elizabeth S. Rogers. "Status of Articulation: Placement, Advanced Placement Credit, and Course Options." *Hispania* 72 (1989): 763-73.
- LeBlanc, Raymond & Gisèle Painchaud. "Self-Assessment as a Placement Test." 1985a. ERIC ED 259 584.
- \_\_\_\_\_. "Self-Assessment as a Second Language Placement Instrument." *TESOL Quarterly* 19 (1985b): 673-87.
- Loughrin-Sacco, Steven J., Sylvia A. Matthews, Wendy M. Sweet & Jan A. Miner. "Reviving Language Skills: A Description and Evaluation of Michigan Tech's Summer Intensive French Course." *ADFL Bulletin* 21, 2 (1990): 34-40.
- MacIntyre, P. D. & R. C. Gardner. "Anxiety and Second-Language Learning: Toward a Theoretical Clarification." *Language Learning* 39 (1989): 251-75.

- Martin, Elizabeth. "Appendix H: Scheme for Classifying Survey Questions According to Their Subjective Properties." *Surveying Objective Phenomena*. Ed. Charles F. Turner & Elizabeth Martin. New York: Russell Sage Foundation, 1984: 1: 407-31.
- Molenaar, Nico J. "Response Effects of 'Formal' Characteristics of Questions." *Response Behaviour in the Survey-Interview*. Ed. W. Dijkstra & J. van der Zouwen. New York: Academic Press, 1982: 49-89.
- Newstead, Stephen E. & Janet M. Collis. "Context and Interpretation of Quantifiers of Frequency." *Ergonomics* 30 (1987): 1447-62.
- Oller, John W. *Language Tests at School*. London: Longman, 1979.
- \_\_\_\_\_ & Kyle Perkins. "Language Proficiency as a Source of Variance in Self-Reported Affective Variables." *Language in Education: Testing the Tests*. Ed. John J. Oller Jr. & Kyle Perkins. Rowley, MA: Newbury House, 1978.
- Olson, Margot A. & Diane Martin. "Assessment of Entering Student Writing Skill in the Community College." 1980. ERIC ED 235 845.
- Oskarsson, Mats. *Approaches to Self-Assessment in Foreign Language Learning*. Oxford: Pergamon Press, 1978.
- \_\_\_\_\_. "Self-Assessment of Language Proficiency: Rationale and Applications." *Language Testing* 6 (1989): 1-13.
- \_\_\_\_\_. "Subjective and Objective Assessment of Foreign Language Performance." *Directions in Language Testing: Selected Papers from the RELC Seminar on "Evaluation and Measurement of Language Competence and Performance"*. Ed. John A. S. Read. Singapore: Singapore University Press, 1981.
- Painchaud, Gisèle. Personal communication. 1989.
- Pohl, Norval F. "Using Retrospective Pre-Ratings to Counteract Response-Shift Confounding." *Journal of Experimental Education* 50 (1982): 211-14.
- Reed, Keflyn X. "Expectation vs. Ability: Junior College Reading Skills." 1988. ERIC ED 295 706.
- Reid, Joy. "The Dirty Laundry of ESL Survey Research." *TESOL Quarterly*. 24 (1990): 323-38.
- Schaefer, Carl F. "The Cloze Procedure for Placement Testing." *Glottodidactica*. 15 (1982): 75-82.
- Schmitt, Neal & Daniel M. Stults. "Factors Defined by Negatively Keyed Items: The Result of Careless Respondents?" *Applied Psychological Measurement* 9 (1985): 367-73.
- Schwarz, Norbert & Hans-J. Hippler. "What Response Scales May Tell Your Respondents: Informative Functions of Response Alternatives." *Social Information Processing and Survey Methodology*. Ed. Hans-J. Hippler, Norbert Schwarz & Seymour Sudman. New York: Springer-Verlag, 1987: 163-78.
- Shaw, Peter A. "Comments on the Concept and Implementation of Self-Placement." *TESOL Quarterly* 14 (1980): 261-62.
- Shohamy, Elana. "Language Testing Priorities: A Different Perspective." *Foreign Language Annals* 23 (1990): 385-94.

- Stevenson, Douglas K. "Language Testing and Academic Accountability: On Redefining the Role of Language Testing in Language Teaching." *IRAL* 19 (1981): 15-30.
- Turner, Charles F. "Why Do Surveys Disagree? Some Preliminary Hypotheses and Some Disagreeable Examples." *Surveying Subjective Phenomena*. Ed. Charles F. Turner & Elizabeth Martin. New York: Sage Foundation, 1984, 2:159-214.
- Upshur, John A. "Objective Evaluation of Oral Proficiency in the TESOL Classroom." *Papers on Language Testing 1967-1974*. Ed. Leslie Palmer & Bernard Spolsky. Washington, DC: TESOL, 1975: 52-65.
- von Elek, Tibor. "A Test of Swedish as a Second Language: An Experiment in Self-Assessment." *New Directions in Language Testing*. Ed. Y. Lee, A. Fok, R. Lord & G. Low. Oxford: Oxford University Press, 1985: 47-57.
- Wangsotorn, Achara. "Self-Assessment in English Skills by Undergraduate and Graduate Students in Thai Universities." *Directions in Language Testing: Selected Papers from the RELC Seminar on "Evaluation and Measurement of Language Competence and Performance"*. Ed. John A. S. Read. Singapore: Singapore University Press, 1981: 240-60.
- Weltens, Bert, Theo J. M. Van Els & Erik Schils. "The Long-Term Retention of French by Dutch Students." *Studies in Second Language Acquisition* 11 (1989): 205-16.
- Wesche, M. B., F. Morrison, D. Ready & C. Pawley. "French Immersion: Post-secondary Consequences for Individuals and Universities." *Canadian Modern Language Review* 46 (1990): 430-51.
- Yaremko, R. M., Herbert Harari, Robert C. Harrison & Elizabeth Lynn. *Handbook of Research and Quantitative Methods in Psychology: For Students and Professionals*. Hillsdale, NJ: Erlbaum, 1986.

## Appendix

### Sample Self-Assessment Items

Following are actual items used in various reported studies and projects.

■ Global, writing (Heilenman, 1989).

Imagine that you have been given the job of writing a detailed (3 to 4 page) account about yourself in French. It would include your life from when you were small up to the present day (childhood memories, games, school, friends, youth, present occupation, interests, personal qualities, relatives and friends, memorable experiences, future plans, etc.). You have plenty of time but there's no dictionary available. How well would you manage?

- \_\_\_\_\_ I wouldn't be able to write anything at all.
- \_\_\_\_\_ I would be able to write only a few simple sentences about my life.

- \_\_\_\_\_ I could give a bare-bones account of my life, but it wouldn't be very long or sophisticated.
- \_\_\_\_\_ I could write about most of the major events of my life but I would probably have to leave out some things that I wouldn't know how to say.
- \_\_\_\_\_ I would be able to give a fairly accurate account of my life but I probably would have some trouble being very elaborate (giving lots of details, etc.) I might have some trouble with particular vocabulary or grammar.
- \_\_\_\_\_ I would be able to write a fairly complete and coherent account of my life with only minimal difficulty.

- Global, difficulty (Davidson & Henning, 1985, p. 178).  
Speaking fluently  
none / very little / some / average / more than avg. / much / extreme
- Can-do, speaking, 3-point scale, rough difficulty ordering (Barrows et al., 1981, p. 164)  
Give simple biographical information about myself (place of birth, composition of family, early schooling, etc.).  
quite easily / with some difficulty / with great difficulty or not at all
- Can-do, various skills, 4-point scale (Clark & Jordan, 1984, p. 77)  
Teaching a class using Japanese as the language of instruction.  
extreme difficulty / considerable difficulty / some difficulty / little or no difficulty
- Can-do, listening comprehension, 4-point scale (Hilton et al., 1985)  
Understand movies without subtitles.  
quite easily / with some difficulty / with great difficulty / not at all
- Can-do, listening comprehension, 5-point scale, rough difficulty ordering (Painchaud, 1989)  
*Je peux comprendre des directives données en anglais dans un cours sans avoir à les faire répéter.*  
*jamais / rarement / moitié / souvent / toujours*
- Can-do, task provided (von Elek, 1985)  
Do you understand this sentence?  
yes, absolutely / I think so / no
- Difficulty, grammar, 4-point scale (Bachman & Palmer, 1989, p. 27)

How many different kinds of grammar mistakes do you make in English?

(BAD) I make grammar mistakes in almost everything / many kinds / only a few kinds / I almost never make grammar mistakes (GOOD)

■ **Descriptor matching, reading (Barrows et al., 1981, p. 268)**

This question asks you to judge your own level of reading ability in your MPL [most proficient language]. Please read each of the six paragraphs below and decide which paragraph best describes your ability to read the MPL. Circle the number preceding only one of the paragraphs below.

- 1) I cannot really read anything in the language, or can read only a few words that I have "memorized."
- 2) I can recognize the letters of the alphabet or the very common characters or printed syllables of the language. I can read some personal and place names, street signs, office and shop designations, numbers, and some isolated words and phrases.
- ...
- 6) I can read extremely difficult and abstract prose, as well as highly colloquial writing and the classic literary forms of the language.