DOCUMENT RESUME

ED 364 064 FL 021 381

AUTHOR Wesche, Marjorie; And Others
TITLE A Comparative Study of Four Placement Instruments.
PUB DATE Aug 93
NOTE 18p.; Paper presented at the Annual Language Testing Research Colloquium (15th, Cambridge, England, and Arnhem, The Netherlands, August 2-8, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; Comparative Analysis; Dictation; Difficulty Level; *English (Second Language); Foreign Countries; Higher Education; Intensive Language Courses; Language Role; *Language Tests; *Listening Comprehension; Questionnaires; *Reading Comprehension; Self Evaluation (Individuals); *Student Placement; Test Use; *Vocabulary

ABSTRACT

A study compared four tests of English as a Second Language (ESL) used for placement of students of varying language backgrounds and skill levels in an intensive ESL program. The tests were a text-based listening and reading test, a listening dictation, a self-assessment questionnaire, and a self-reported vocabulary size test. All measures were administered to 93 candidates on entry into the program. Results indic that the instruments performed differentially overall, by proficiency level, and by native language background. In general, the first two tests, which required demonstration of proficiency and were text-based, worked best. Neither self-report measure worked well, although the questionnaire worked better than the vocabulary test. Possible explanations for these findings are seen in test design, content, and tasks required. (MSE)

A Comparative Study of Four Placement Instruments[1]

Marjorie Wesche, T. Sima Paribakht, Doreen Ready
University of Ottawa

Abstract

Accurate student placement raises the theoretical issue of method validity while practical concerns often limit potential choice regardless of instructional content.

The present study compared four instruments representing different methods in a multi-skill, intensive ESL program. Placement accuracy for different proficiency levels, for students of different L1 backgrounds, and relationships among measures were considered.

The instruments were:

- *English Placement Test*, a text-based listening and reading test presenting varied tasks and short-answer formats.
- *Listening Dictation*, a cassette-recorded text requiring reconstruction of varied-length chunks of the original text which tax short-term memory, scored for listening precision.
- *Self-Assessment Questionnaire*, which uses Likert scale ability estimates for descriptions of everyday language uses.
- *Eurocentres' Vocabulary Size Test* an adaptive, personal computer administered self-report procedure based on "known" words and including correction for guessing.

All measures were administered to 93 candidates at program entry. Final placement level was the criterion.

The instruments performed differentially overall, by proficiency level and by L1 background.

Key words: placement, method, self assessment, vocabulary, dictation, comprehension

Theoretical Background and Rationale

The accurate placement of students in language programs, like other testing objectives such as certification or achievement, raises the theoretical issue of method validity for given score interpretations and uses. Second language testing instruments are realisations of different methods through which the attempt is made to measure language ability. Tests vary in many ways (see Bachman, 1990, for an analysis of method facets), and although there is little evidence of the effects of specific facets of tests, research has shown that fairly large differences in testing method can lead to systematic variance in test performance apart from ability (Shohamy, 1990). One major cleavage in test methods is between tests which present language processing tasks and those in which testees report on their own knowledge or ability to do things in the second language. The former type of test may, among other things, vary according to the channel, mode and text characteristics of the language input, the nature of the processing tasks that are set, characteristics of the required response, and scoring criteria, as well as features of the testing environment, test format, organization and presentation, scoring procedures, and the interpretation of scores. In placement, the recommended practice is to use a test which reflects the nature and emphasis of instruction in its method. Nonetheless, there is a tendency for large instructional programs to depend upon more easily tested receptive skills and constructed responses when this type of instrument is used.

Self-report procedures usually require candidates to rate their ability to "do" certain things using their L2, or their "knowledge" of particular elements or patterns of the L2. Sometimes, however, the criteria are less precisely defined; e.g., "beginner" to "advanced", or "non-native" to "nativelike". Self-assessments are subject to poor reliabilities when candidates are either unable or unwilling to give an honest appraisal, the first case arising from unclear or unfamiliar criteria or the candidate's inability to analyze his or her own performance; the second case arises when there is a perceived advantage to a high or low rating. (See discussion in Ready, forthcoming.) However, the successful use of such instruments for placement in some settings (LeBlanc & Painchaud, 1985; Meara, 1990) and their ease and limited expense of administration make them worth a second look by L2 instructional programs.

The practical question for administrators is, how fine are the ability distinctions required for a given program, and, within the resource constraints under which all programs operate, what is the best feasible solution — be it a standardized, carefully developed in house or ad hoc language test or self-report procedure. The point is usually made that unlike the outcomes of certification tests, changes to poor initial placements are usually possible, and therefore a rough initial sorting is adequate for the purpose. However, poor placements and subsequent changes result in lost instructional time and frustrate students and teachers alike. The method issue thus remains important, even when it has become "Which of the possible tests (methods) is best for this program?" rather than "What is the best possible test (method) for this program?"

Purposes of the Research

The present study, carried out in 1992, sought to compare the accuracy of three alternative placement instruments using different methods with the accuracy of the instrument currently in use in a multi-skill, intensive ESL summer program for Canadian high school graduates. Students came from varied L1 backgrounds, and included a large group of French L1 speakers. The study investigated placement accuracy at seven proficiency levels and overall, relative efficacy for students of the same versus different L1 backgrounds (French versus non-native French speaking) and relationships between the different measures.

Research Questions

The following questions guided the research:

- *How well do the various instruments compare to the English Placement Test and to each other in terms of overall placement accuracy?*
- *Which instruments work best at low, middle and high proficiency levels?* (as defined by criterion groups 1-2, 3-4, and 5, 6, 7).
- *Are the instruments differentially effective in placing homogeneous L1 (i.e., francophone) vs. heterogeneous L1 (L2 French speaking) students?*

- *What are the implications regarding appropriate placement instruments for this and other programs and to what extent does testing method appear to play a role?*

Research Design and Methods

Setting: The six-week summer ESL program is part of a national bursary program to provide intensive L2 exposure and practice to Canadian high school graduates and university students who wish to improve their English or French second language use skills. Official objectives of the program include the strengthening of oral skills and development of knowledge and appreciation of the L2 culture. Since students in the ESL program tend to have strong oral skills already, given the omnipresence of English throughout most of Canada as a language of the wider community and its prominence in the media, a four-skill approach is used at all levels. The program offers a variety of language activities in and out of the classroom, including daily morning classes organized around themes, featuring authentic materials of various kinds, and week-long afternoon workshops (e.g., film interpretation; preparation of a student newspaper).

Subjects: The subjects in the present study were high school graduates and university students from 18-25 years old. The group consisted of 56 francophone students and 37 students from a variety of other linguistic backgrounds, e.g., Arabic, Chinese, Spanish, Polish, Turkish.

Instruments: The main placement instrument was the English Placement Test developed initially for the academic year comprehension-based program for beginners and intermediates at the Second Language Institute (SLI). This test has been carefully validated over the course of several regular academic semesters to provide accurate cut-off scores for these courses. The validation process consisted of comparisons of the scores obtained on the Placement Test with teacher assessments and rankings obtained at the beginning of the semester, with student mid-term marks and with student final marks. At the end of the semester, adjustments were made to the cut-off scores where it was thought to be necessary, and then the process was repeated the following semester to verify any changes that were made. This process continued until the cut-off points were satisfactory. Since the summer bursary program includes students from a

wide ability range, validation at higher proficiency levels was carried out over the past three summers using the same criteria.

A Listening Dictation Test was developed in 1991 to provide supplementary information at lower proficiency levels in the summer bursary and other Institute programs. The other two instruments were more widely used self-report tests which offered considerable logistical appeal, as one was self-administered and scored and the other administered and scored by personal computer. All instruments demonstrated acceptable to very high reliabilities (see below).

- *English Placement Test (EPT)*, SLI, University of Ottawa. The EPT is a text-based listening and reading test presenting varied tasks and short-answer formats. This test assesses the testees' reading and listening comprehension ability. Version II of the test which was used in this study has three listening comprehension sub-tests on different themes (i.e., two students discussing their exam schedules, a radio text on Mother's Day and a biographical sketch on Chopin). Students are given time to read the comprehension questions before listening to the text. After they have listened to the text, they are given time to answer the questions. The text is then played a second time and at the end students are given time to check their answers. The listening test takes about 20 minutes and students answer a total of 32 questions in a variety of formats (multiple choice, fill-in the blank, chart).

The reading comprehension part consists of three sub-tests on a variety of themes (i.e., a letter to a magazine editor, an announcement of a contest honouring the founding of a city and fitness levels in Canada). A variety of task formats (e.g., multiple choice, true or false, summary cloze) are used. There are a total of 32 questions in this part of the test, and students are given one hour to complete it. Both the listening and reading questions cover a range of comprehension tasks, ranging from identification of main ideas to finding specific information. The results have always been quite consistent and very few changes have had to be made to the initial placement levels. The main weakness that has been observed is that for some students the listening part of the test gives an underestimation of student ability at lower ranges of proficiency

because of the unfamiliarity of some of the tasks. That is why the Listening Dictation was developed in order to provide an additional measure of listening ability.

- *Listening Dictation*, SLI, University of Ottawa. This test presents a listening text, based on a short biography of a youthful Canadian hero, which must be understood and written down by the student. Presented on cassette, it requires reconstruction of varied-length chunks of the original text which tax short-term memory. Testees are given one point for every identifiable word in the correct order (total of 147 words). They are not penalized for spelling errors, verbs with the wrong endings, singular instead of plural, etc., unless the word is unrecognizable and far from the original meaning. Extra words are ignored. Sentence or word inversions are scored as correct if the sentence and/or word still makes sense. The Listening Dictation is best described as testing precision in listening comprehension.

- *Self-Assessment Questionnaire*, University of Ottawa (cf. Ready, forthcoming). This instrument uses Likert scale ability estimates for descriptions of everyday language uses in an academic environment. The self-assessment instrument, which is administered in the student's L1 (either English or French), has been used for initial placement purposes in academic credit courses at the SLI, University of Ottawa since 1985. It consists of a series of 60 statements which briefly outline situations in which students might find themselves having to use their second language receptively. They are asked to respond using a five-point scale ranging from "I cannot do the task at all" to "I can do it all the time." The tasks are related either to listening or reading and are sequenced according to increasing difficulty. An example of a low level task is:

> *"I can understand a notice announcing a class cancellation when it is only written in French."*

An example of a more difficult task is:

"I can read a French newspaper and understand the gist of the stories on the front page."

Experience has shown that there is a sufficient variety of tasks included in the self-assessment questionnaire to allow differentiation among the seven levels.

• *Eurocentres' Vocabulary Size Test (EVST)* (Eurocentres, 1990). The EVST, developed by Meara and his colleagues (cf. Meara & Jones, 1990), belongs to a family of self-report checklist tests. Using words sampled from a word frequency list plus a set of imaginary words which would be possible in the given language, these tests ask students whether they "know" sample words, and provide an overall estimate of ESL learners' vocabulary size in the target language. The English language version is used by the British Council Eurocentres for purposes of placement, on the authors' rationale that "vocabulary knowledge is heavily implicated in all practical language skills" (Meara & Jones, 1988, 80). An example from a French pencil and paper version is given below (Meara & Jones, 1988, 81):

Look through the French words listed below. Cross out words that you do not know well enough to say what they mean. Keep a record of how long it takes you to do the test.

VIVANT	TROUVER	MAGIR	ROMPANT
MELANGE	LIVRER	IVRE	FOMBE
MOUP	VION	LAGUE	INONDATION
SOUTENIR	SIECLE	TORVEAU	PRETRE
REPOS	GANAL	HARTON	TOULE
GOUTER	FOULARD	EXIGER	AVARE
ETOULAGE	ECARTER	MIGNETTE	JAMBONNANT
DEMENAGER	POIGNEE	EQUIPE	MISSONNEUR
AJURER	BARRON	CLAGE	TOUTEFOIS
LEUSSE	CRUYER	HESITER	SURPRENDRE
LAVIRE	SID	ROMAN	CHIC
ORNIR	CERISE	PAPIMENT	CONFITURE
GOTER	PONTE		

Meara & Jones (1990) have produced computer-administered versions of the earlier pencil and paper vocabulary tests in a number of languages which make the test even more practical for some settings. The computerized EVST has a "Yes/No" format and consists of a bank of vocabulary items drawn from different frequency bands (up to a ceiling of 10,000 words in version E1.1/K10, MSDOS), as well as non-existent words which conform to English wo formation rules as a correction for guessing. The test begins with the easiest words and gets progressively more difficult and stops once it finds a sufficiently low level of performance and then does a detailed analysis at that level. Target words appear on the screen one at a time and the testee is asked to indicate if (s)he knows the word well enough to be able to give its meaning. The imaginary words act as a built-in mechanism for adjusting scores for false claims and overestimates, and a correction factor based on the percentage of these is calculated into the final score (Meara & Buxton, 1987). Meara & Jones (1988) noted the possibility that the test overestimates true vocabulary knowledge but Meara (1990) has subsequently revised this position based on experience with the test, to the effect that most people probably underestimate their knowledge, due possibly to inherent conservatism or to the inability to access little-known words presented in this way. In any case, we do not know what individuals do, and other studies of self-assessment of L1 proficiency suggest considerable inter-subject variability (Ready, forthcoming). The EVST shows good test-retest reliability (Meara personal communication). Part of its attractiveness is that it is very easily administered, requiring approximately 10 minutes on a personal computer, and is automated and self-scoring.

Procedures

All measures except the vocabulary test were administered to over 100 candidates at program entry. The vocabulary test was subsequently administered to those in levels 2-7 (N=93). Final placement level (based on information from the first three tests, in-class measures and teacher observation) was the criterion. The placement procedure was the following. Students were ranked in ascending order of their scores on the English Placement Test and were then divided into seven approximately equal groups. The teachers administered both an oral exercise (each student interviewed and presented a classmate to the rest of the class) and a composition

task to their initial groups during the first and second day of classes, and then received the scored Listening Dictation papers.

Based on this information and their own observations, the teachers reconsidered the appropriateness of the initial placement, particularly of the most and the least proficient students in each class. The teachers met together on the third day to compare information and decide upon placement changes, maintaining approximately equal groups. Approximately 2% of the students were changed from their initial group, mainly in cases where their oral proficiency was markedly different from the rest of their group. This percentage was particularly low compared to recent years. Teachers reported that, due to the relatively large classes, they were reluctant to add students to their groups or to ask others to do so.

Analyses

The following analyses were carried out.

Descriptive statistics were calculated for all four instruments overall and at each final placement level (Table 1). Correlations were calculated among the four instruments plus the reading and listening sub-tests of the EPT and with final placement level for the overall population, for low middle and high proficiency segments of the population and for all francophone and non-francophone subjects (Tables 2, 3 and 4).

Results

The results are reported in terms of research questions.

How well do the various instruments compare to the English Placement Test and to each other in terms of overall placement accuracy?

Table 1 **Descriptive Statistics for the Four Placement Instruments (N=93)**

	Total EPT		
Level	Range	X	S.D.
1	15 - 30	22.2	4.3
2	26 - 33	31.5	3.0
3	36 - 46	42.1	3.3
4	46 - 53	48.6	2.6
5	53 - 62	55.6	2.3
6	58 - 63	60.9	1.8
7	63 - 69	65.9	1.9
Overall			

	Listening Dictation		
Level	Range	X	S.D.
1	34 - 114	73	21
2	77 - 122	97	15
3	67 - 142	111	20
4	85 - 142	120	15
5	89 - 140	127	13
6	17 - 145	138	7
7	35 - 147	141	3.5
Overall			

	Self-Assessment		
Level	Range	X	S.D.
1	60 - 283	161	59
2	148 - 250	193	31
3	115 - 262	196	35
4	158 - 251	210	30
5	171 - 269	227	27
6	202 - 292	241	26
7	169 - 295	236	34
Overall			

	Vocabulary		
Level	Range	X	S.D.
1	—	—	—
2	2393 - 6398	4107	1165
3	2633 - 6504	3980	1115
4	1720 - 6518	5142	1409
5	3578 - 7600	5430	1289
6	3458 - 7714	5628	1420
7	3470 - 8616	6122	1247
Overall			

Table 2 Correlation of Placement Instruments with Final Placement Level (All Subjects)

	Placement Level
Total EPT	.96
Reading EPT	.91
Listening EPT	.90
Listening Dictation	.82
Self-Assessment	.58
Vocabulary Size	.52

Table 1 shows the descriptive statistics for each instrument at each final placement level. While there is little overlap of scores between levels for the EPT, all four comparison instruments show a wide range of scores at each level with considerable overlap. The Listening Dictation shows a steady increase in the mean at each level, although the only difference between contiguous pairs that is statistically significant is that between levels 1 and 2. The Self-Assessment test and the Vocabulary test do not consistently show increases in the mean from level to level and none of the contiguous pairs of means are statistically significantly different from each other.

Table 2 shows the relationship of each test with the final placement level for all subjects. As might have been expected, the EPT total score correlates most highly with the final placement level, followed closely by the EPT reading and listening sub-test scores. Of the other three instruments, the Listening Dictation score is the highest (.82) while the Self-Assessment and Vocabulary self-report scores are both quite low (.58 and .52 respectively).

Which instrument or combination works best at low, intermediate and high proficiency levels?

Table 3 **Correlations of Placements Test with Final Placement Level for Low, Intermediate and Advanced Groups**

Group	Total EPT	Listening EPT	Reading EPT	Listening Dictation	Self-Assessment	Vocabulary
Low	.80	.45	.66	.56	n.s.	—
Intermediate	.74	n.s.	.69	n.s.	n.s.	.43
Advanced	.91	.64	.74	.54	n.s.	n.s.

Table 3 shows the correlations of the scores on various instruments and part scores with final placement grouped as low, intermediate and advanced proficiency levels. Only the total English Placement Test score and the Reading EPT sub-score correlate with final placement level across all three proficiency levels. In both cases, the correlation is highest at the advanced proficiency level. At low levels of proficiency, Listening Dictation was somewhat better correlated with Final Placement level than the Listening EPT sub-score but in both cases, the correlations are in the moderate range. The Self-Assessment score is not significar at any of the three proficiency levels. (The Vocabulary test was not administered at the low proficiency level.) At intermediate levels of proficiency, the only other score besides EPT Total and EPT Reading that correlates with final placement level is that of the Vocabulary test. At advanced levels of proficiency both the Listening Dictation and the Listening EPT sub-score also correlate with Final Placement Level but Self-Assessment and the Vocabulary test do not.

The correlation between the EPT and final level placement is almost certainly an overestimate of the relationship at lower levels but not at higher levels (Table 3). If initial student placement had been consistently changed on the basis of their Listening Dictation scores, approximately 8% of them at lower levels (1-3) would have been moved. This was not done, however, for the reasons previously indicated. At higher levels (high intermediate to advanced) the correlation is .91, at low intermediate levels .74 and at high beginner levels .80. It appears that this test works particularly well at higher proficiency levels and the listening part of the test works best with advanced students. This may be partially due to a method effect, in that the

novelty of listening item formats — including a variety of fill-in, matching, chart and multiple choice items — may create added difficulties for some lower level students. No such effect is seen in the reading part of the test, where students are not constrained by time.

Are the instruments differentially effective in placing homogeneous L1 (francophone) and heterogeneous L1 (non-native French speaking) students?

Table 4 shows the correlations of the various instruments and sub-scores with each other and with the final placement level for francophone students (N = 56) and non-native French speaking students (N = 37). The pattern of correlations of the various instruments and sub-scores with final placement level is quite similar for the two populations except in the case of Self-Assessment. That correlation is moderate for francophones but not significant in the case of non-francophones.

Table 4 **Correlations Among Test Scores and With Final Placement Level for Francophone and Non-Native French Speaking Students**
(correlations for francophone students are given first followed by correlations for non-francophone students (italicized)

	EPT-T	EPT-L	EPT-R	L-Dict	Self-A	Vocab
English Placement Test Total	— —					
English Placement Test Listening	.91 *.90*	— —				
English Placement Test Reading	.88 *.91*	.60 *.62*	— —			
Listening Dictation	.82 *.77*	.73 *.75*	.75 *.64*	— —		
Self-Assessment	.67 *(.22)°*	.65 *(.21)°*	.54 *(.19)°*	.72 *.31*	— —	
Vocabulary Size	.52 *.52*	.43 *.39*	.51 *.55*	.49 *.56*	.48 *(.25)*	— —
Final Placement	.98 *.96*	.88 *.85*	.87 *.88*	.79 *.77*	.66 *(.27)°*	.51 *.56*

° not significant

Discussion and Conclusions

The final question leads into our discussion of results and conclusions:

What are the implications regarding appropriate instruments for this and other programs and to what extent does testing method appear to play a role?

The results of this study lead to the not surprising conclusion that tests which have been shown to work well in other seemingly similar contexts cannot be assumed to be appropriate in a new context. Overall, the tests requiring a demonstration of proficiency on the part of students

worked best (EPT Reading & EPT Listening and Listening Dictation). These were, furthermore, text-based tasks. EPT content and tasks conformed most closely to the communicative instructional objectives and content of the summer ESL program (although it did not test productive skills). The EPT uses authentic (non-contrived) texts of general interest to university-age students, and tasks require global understanding through listening and reading of the kinds of information voluntary listeners and readers would be expected to retain. The texts are varied in subject matter, genre and tasks, unlike the Listening Dictation or Vocabulary Test. The Listening Dictation is also based on an interesting extended text, but tests only listening comprehension and a threshold level of writing. The findings suggest that content validity is important in placement testing.

Neither self-report measure worked well, although the Self-Assessment based on functional descriptions of language uses worked better than the vocabulary measure overall, particularly for francophone L1 students. Although all students taking the test had French as their first language of study, many were allophones, and for them, placement via the Self-Assessment was extremely unreliable. There are two possible explanations for this; one would be the language factor in the instrument itself. This seems unlikely, however, as these students have done their high school work in French. The other possibility is that of cultural differences in English learning experiences and/or in ability and readiness to self-report one's language knowledge. The Vocabulary Size Test also did not work well overall. Since the summer bursary course does not specifically aim to teach vocabulary, a vocabulary test may be less appropriate here than in other situations. Still, it should be remembered that the rationale for using this test for placement is that it is viewed as an indicator of language proficiency. In spite of its general ineffectiveness in this context, this test was reasonably effective at intermediate proficiency levels. An interesting question would be whether the relationship between vocabulary and general proficiency is strongest at this level, but this study provides no further evidence on this issue. Unlike the Self-Assessment, this test was presented in the target language, English. However, the language of presentation and task was very straightforward, and what was required more than language knowledge was, probably a threshold comfort level with computers.

Finally, it should be noted that method factors do appear to influence language test performance in this study as in others, and that, for adequate placement in courses, tests developed for local needs and normed on representative populations are required.

Note

1. We are grateful to Sandra Burger, Michael Massey, Paul Meara and to the 1992 E.S.L. Summer School teachers and their students for their help with this study, and to Trixi Magyar for graphics and word-processing.

References

Bachman, L.F. 1990. *Fundamentals of Language Testing*. Oxford: Oxford University Press.

LeBlanc, R. and Painchaud, G. 1985. Self-assessment as a second language placement instrument. *TESOL Quarterly* 19(4), 673-687.

Meara, P. 1990. Matrix models of vocabulary acquisition. *AILA Review*, 66-74.

Meara, P. and Buxton, B. 1987. An alternative to multiple choice vocabulary tests. *Language Testing* 4(2), 142-154.

Meara, P. and Jones, G. 1988. Vocabulary size as a placement indicator. In Grimwell, P., editor, *Applied Linguistics in Society*. London: CILT, 80-87.

---- 1990. *Eurocentres Vocabulary Size Test* (version E1.1/K10,MSDOS). Zurich: Eurocentres Learning Service.

Ready, D. forthcoming. The role and limitations of self-assessment in testing and research. *Proceedings LTRC 1991*. Princeton, N.J.: Educational Testing Service.

Second Language Institute 1986. *Self-Assessment Questionnaire*. Ottawa: Second Language Institute, University of Ottawa.

Shohamy, E. 1990. The effect of contextual variables on test takers' scores on language tests. Internal document, University of Tel Aviv.