INTRODUCTION

The purpose of this project was to create a placement test for the Spanish language program in the Department of Modern Languages and Cultural Studies. In particular, the project investigated whether a measure of vocabulary size could be used as a placement measure. Most placement tests seek to test as many of the four language skills—reading, writing, speaking and listening—as possible, resulting in a procedure that is time-consuming and administration-intensive. This project sought to create a placement test that could measure students’ knowledge quickly yet accurately enough for a placement recommendation to be made. The format chosen was a vocabulary task that estimated the size of the learner’s vocabulary. The underlying assumption was that learners with a larger vocabulary would have had greater exposure to the language in general and consequently, they should perform better in other language skills as well. The implications of this project are twofold: first, it offers a practical placement procedure that can be implemented easily by other language programs, and second, it contributes to existing research that shows a correlation between vocabulary size and overall language proficiency (see below).

RATIONALE AND REVIEW OF LITERATURE

While English-as-a-second-language (ESL) departments benefit from a range of widely-available standardized tests, such as TOEFL or the Cambridge Proficiency examinations, foreign language departments do not share this benefit and are often left to devise their own procedures. Many departments simply place students based on "seat time", that is, how many Spanish classes the students have previously "sat through". However, Spencer and Flaugher (1967), Aleamoni and Spencer (1968), and Hagiwara (1983) found that the results of a seat-time formula equating one year of high school study with one semester of college study correlated poorly with the scores on the standardized tests offered by the Modern Language Association and the College Entrance Examination Board. The lack of correspondence may be due to the fact that such formulae do not take into account factors such as gaps in the years of language study, individual linguistic aptitude, ability and motivation, and the quality of instruction received.

Many foreign language programs also rely heavily on the assessments of instructors, program coordinators and advisors to determine where a student should be placed. However, such subjective evaluations risk inconsistency and therefore lack validity. While some instructors are strict about what a student must know in order to place at a certain level, other instructors are more easily swayed by personal reasons, such as the desire to be in the same class as a friend. Moreover, it is not always possible to determine a student’s level of knowledge through a quick
Students also tend to underestimate their level of knowledge and place themselves in a lower level than they should be (Hagiwara, 1983).

Foreign language programs, including the Spanish program at the University of Alberta, have long recognized the need for an effective placement procedure. The more heterogeneity in a class, the more difficult the instructor’s task is in order to accommodate the needs of the different levels. It is also problematic for students, who may feel intimidated by their more proficient classmates or, conversely, bored by the less proficient ones. Thus, an accurate evaluation of proficiency levels for placement purposes is crucial in order to facilitate teaching and learning. Two surveys of Spanish departments in the United States revealed a general concern about a lack of good tests for placement and a general dissatisfaction with existing procedures (Klee & Rogers, 1989; Wherritt & Cleary, 1990). At the same time, despite their recognition of the importance of improving placement policies, most departments had not yet found an affordable method nor had the time and resources to create their own tests.

Several placement test formats are available for foreign language programs that wish to devise their own tests. One option is the multi-modal test, which strives to test all four language skills (e.g. University of Minnesota, described in Lange, 1987). In their survey of American university Spanish departments, Klee and Rogers (1989) found that of the 42 departments that used a placement exam, all of them tested reading, 60% tested listening, 30% tested writing and 8% tested speaking. The problem with multi-modal tests is that they are difficult to construct and time-consuming to administer because of the numerous components. Another common complaint about multi-modal tests is the difficulty in scoring them. Although it may be possible to limit the responses in some sections (e.g. reading can be tested via multiple choice or true/false questions), there are inevitably some open-ended sections, such as speaking and writing, that require a subjective evaluation on the part of the evaluator, thereby raising issues of rater training and interrater reliability. The labour-intensiveness of the multi-modal test did not make it a viable option for this project.

Another option for placement testing is the cloze test, which in the last forty years has become a popular alternative to the multi-modal test because of its ease of construction (Davies et al., 1999). Words are deleted from a text passage, either at a fixed rate or according to its lexical category, and the students have to fill in these words based on their understanding of the text. As a result, cloze tests are easy to construct and administer. The underlying rationale is that languages tend to code the same information multiple times in the same text, but this language redundancy only works if the reader is proficient enough to take advantage of it (Aitken, 1977). In addition, reading comprehension requires all aspects of language use and thus reflects overall linguistic proficiency (ibid). Most studies have found a good correlation between a cloze test and standardized placement/proficiency tests (e.g. Heilenman, 1983; Hinofotis, 1980; Stubbs & Tucker, 1974). However, the cloze test is not without problems and is viewed with some suspicion among researchers in language testing. It is far from clear what exactly the cloze test measures, other than the fact that it correlates well with other proficiency tests (Davies et al., 1999). Moreover, changing the word deletion rate, the difficulty level of the text, the scoring method and/or the sample population used can yield divergent results (e.g. Alderson, 1979a, 1979b; Brown, 1983). In addition, the studies that have looked at the predictive value of the
cloze test have cast some doubt on its ability to predict level (e.g. Heilenman, 1983; Oller & Conrad, 1971). For these reasons, a cloze test was not considered for this project.

As an alternative to the multi-modal test and the cloze test, Paul Meara and his colleagues proposed the use of a vocabulary test (Meara, 1992; Meara & Buxton, 1987; Meara & Jones, 1988). Research on first language acquisition, reviewed in Anderson and Freebody (1981), has shown that vocabulary knowledge is a robust predictor of linguistic abilities and general intelligence. In the area of second language acquisition, Meara & Jones (1988) compared a yes/no vocabulary recognition test, where students had to indicate whether they recognized the word or not, with their institution's multi-modal test of English and found medium correlations of .664 and .717 for the two groups of students they tested. Moreover, placement adjustments made later by the instructors tended to correspond to the vocabulary test score rather than to the multi-modal one. An earlier study (Meara & Buxton, 1987) had also established a high correlation between a yes/no vocabulary test and the Cambridge First Certificate in English. Therefore, vocabulary size seems to be a good measure of overall language proficiency.

Vocabulary tests offer numerous advantages over multi-modal and cloze tests. Vocabulary size is a fairly objective measure, independent of all our uncertainties about how much and what type of language structures a learner should know at a particular stage. Vocabulary tests are easier to construct and quicker to administer than multi-modal tests since there is only one component. These tests are also straightforward to score—learners either know the word or not—thus eliminating concerns about rater subjectivity. Many different types of vocabulary tests can be created (cf. Read, 2000; Nation, 2001, for examples and analyses), the most common type being the multiple-choice format. However, multiple-choice questions are difficult to construct because of the need for good distractors (Read, 2000). Moreover, there is no control for random guessing and the student is limited to one particular use of the word that s/he may not happen to know. In addition, because of limitations on test length, the number of words that can be tested is usually quite low, thus casting doubt on the accuracy of the measure (Meara & Buxton, 1987). Other types of vocabulary tests require the students to write in a missing word, match the word with a synonym, translation or definition, or use the word in a sentence. While these tests are easier to construct, they are more difficult to score because there may be more than one acceptable answer. They also take more time to administer and the number of words tested is even lower than with a multiple-choice format since the task load on the learner is greater.

An alternative format is the lexical decision test, also known as the checklist or the yes/no (Y/N) test, as exemplified in the Meara and Buxton (1987) and Meara and Jones (1988) studies cited earlier. The student is given a list of individual words and s/he has to indicate whether s/he knows these words well enough to say what they mean. To control for overestimating, non-words are included in the list as a measure of how truthful the student is. Lexical decision tests are much easier to construct than multiple-choice formats as there is no need for distractors, and they can test a large number of words in a very short time because it only takes seconds for the student to decide whether s/he recognizes a given word. Lexical decision tests are also simple to score by merely counting the number of real words and imaginary words the student claims to know and applying a mathematical formula from signal detection theory: \( \frac{RY - IY}{100 - IY} \), where \( RY \) is the proportion of real words selected and \( IY \) is the proportion of non-words selected
description of lexical decision test

Although the lexical decision test has been used as a placement measure for ESL students (Meara & Buxton, 1988), it has never been applied to Spanish as a foreign language. Therefore, it was necessary to create a Spanish version of the test. Two hundred words were tested: 120 real words and 80 non-words. The 120 real words were drawn from the lexical corpus of the Spanish language LEXESP: Léxico informatizado del español (Gallés, Martí Antonín, Carreiras Valiña, & Cuetos Vega, 2000). These 120 words represented the most frequent 5000 items in the corpus, which vocabulary research has shown to account for approximately 89% of the words in a normal text (Nation, 2001). Moreover, the 5000-word level is situated on the boundary between high- and low-frequency words (Nation, 1990), which is reasonable for testing purposes since there is not much usefulness to testing knowledge of low-frequency words. The 5000 words were sampled at a rate of approximately 1 in 40 words, in descending order of frequency, to ensure equal coverage of more frequent and less frequent words. Proper names, acronyms, interjections and nationalities were omitted from the list of testable words because their frequency of occurrence was inflated due to the high number of newspapers that formed the corpus. All other words—including grammatical words¹, cognates with English², and multi-word lexical units—were included in the list of testable items. It must also be noted that the corpus defined word as a token and not as a lemma (a group of words derived from the same root). In other words, inflectional and derivational forms were counted separately and not as part of the same lemma. The 80 non-words consisted of items that are not part of the Spanish lexicon, although they follow Spanish morphological and phonological patterns. Forty of the words are derived from identifiable Spanish roots, while the other 40 do not contain an identifiable root.

An online version of the lexical decision test was created so that it could be accessed 24 hours from any location, thereby eliminating the need for manual administration of the test. Once a student’s login had been recorded, the student was blocked from accessing the test a second time. Students were shown the 200 words in five groups of 40 words each, and they were asked to click in the box that preceded every word they thought they knew. The list of words was randomly generated for each session, so no two tests were identical. Ten minutes were allotted for the test, and a clock at the top of the page indicated the remaining time. At the end of ten minutes, the answers were submitted automatically to the server. If students finished the test before ten minutes had elapsed, they were given the choice of reviewing their answers or submitting them without changes. Students were immediately shown their test score, together

---

¹ It could be argued that grammatical words, such as prepositions and pronouns, should be omitted because they carry little semantic content and thus do not really belong to the lexicon. However, it is difficult to define what is a grammatical word and what is a lexical item: prepositions, while grammatical in function, can contribute semantic content. Moreover, a true beginner with no prior exposure to Spanish is unlikely to recognize a grammatical word, and the chances of many grammatical words being tested were low because we sampled in descending order of frequency and most grammatical words were clustered at the beginning of the list due to their high frequency of use.

² English cognates were retained on the list because they are part of the Spanish vocabulary. It is also difficult to distinguish between a cognate and a non-cognate because some cognates have been more integrated into the Spanish language than others.
with a list of the non-words that they had selected (so that they would not incorporate these words into their vocabulary) and a placement recommendation. As a final step, they were given the option of saving, printing or emailing the results for their records.

Since placement decisions are not always based solely on the test score, but also various administrative policies, a language background questionnaire was administered prior to the lexical decision test. For example, students who have taken the equivalent of Grade 12 Spanish are not permitted to take beginners’ Spanish, even if they do not have sufficient proficiency for intermediate Spanish. Likewise, students who indicate exposure to Spanish in the home are placed in a Spanish for Heritage Speakers class, regardless of their actual proficiency level. Therefore, the final placement recommendation was based both on the test score and the questionnaire answers.

An administrator’s site was also created to allow access to the results and modifications of certain protocols. Administrators are able to see the test scores (including the raw numbers of real words and non-words selected and the time spent on the test), the questionnaire answers and the resulting placement recommendation. They can modify the criteria on which placement decisions are based and add or delete words from the test list.

**RESEARCH QUESTION**

The primary research question of this project was the following:

How well does a lexical decision test distinguish among proficiency levels?

As the purpose of the task was to make placement recommendations, it was crucial that it be able to differentiate among the various proficiency levels so that cut scores could be established. Therefore, before being used as a placement measure, the test had to be administered to the existing levels to determine whether it can distinguish among them and also to establish baseline scores for each level.

**PARTICIPANTS**

The lexical decision task was administered to all students registered in a Spanish language course during the fall and winter terms of the academic year 2004-2005. As the lexical decision task could not be used for placement purposes until it was clear that it could distinguish among proficiency levels, the results were not used to change the students’ placements. Instead, it was assumed that each level was more or less homogenous in proficiency and therefore, the range of scores for that level could be used as a point of comparison for future years. In order to obtain as much data as possible, students were awarded 2% of their class participation mark for taking the test. The number of students per level who took the test is shown in Table 1.

---

3 The test was not administered until after the date for adding and dropping courses had passed so that the advisor and the instructors could make placement changes to maintain a more consistent proficiency level within the class.

4 A previous pilot of the test in the academic year 2003-2004 based solely on volunteers resulted in very few participants and did not reflect the range of students’ abilities in each level. By awarding participation marks, we could ensure that a more representative sample of students would take the test.
Table 1.

**Number of Participants per Level**

<table>
<thead>
<tr>
<th>Level</th>
<th>Term</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>111 (Beginners’ Spanish I)</td>
<td>Fall 2004</td>
<td>327</td>
</tr>
<tr>
<td>112 (Beginners’ Spanish II)</td>
<td>Winter 2005</td>
<td>256</td>
</tr>
<tr>
<td>211 (Intermediate Spanish I)</td>
<td>Fall 2004</td>
<td>97</td>
</tr>
<tr>
<td>212 (Intermediate Spanish II)</td>
<td>Winter 2005</td>
<td>74</td>
</tr>
<tr>
<td>300 (Conversation and Composition in Spanish)</td>
<td>Fall 2004</td>
<td>39</td>
</tr>
</tbody>
</table>

The test was administered online outside of class time. At login, students indicated which section they were enrolled in and upon completion of the test, their instructors were sent an email message confirming that the student had taken the test and that 2% should be awarded for their participation. At no time were the test scores released to the instructors. Students were given a two-week period to take the test.\(^5\)

The results were analyzed descriptively per level and comparatively across levels. We did not control for the same student taking the test at the beginning of the fall term at one level and again at the beginning of the winter term at the next level because the objective was not to track student development over time, but rather to get an idea of the typical performance of a SPAN 111 or SPAN 112 class. It was hypothesized that the differences among levels would be statistically significant.

**ANALYSIS OF RESULTS**

Table 2 and Figure 1 show the results of the test per level. Note that it was possible to obtain a negative score on the test if the number of non-words checked was greater than the number of real words checked.

Table 2.

**Descriptive Statistics of Test Scores per Level (max score = 100)**

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Mode</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>327</td>
<td>12.19</td>
<td>28.41</td>
<td>12.17</td>
<td>0.00</td>
<td>88.70</td>
<td>-322.22</td>
</tr>
<tr>
<td>112</td>
<td>256</td>
<td>26.69</td>
<td>17.02</td>
<td>26.39</td>
<td>33.33</td>
<td>81.48</td>
<td>-53.33</td>
</tr>
<tr>
<td>211</td>
<td>97</td>
<td>46.61</td>
<td>17.61</td>
<td>46.91</td>
<td>32.22(+)</td>
<td>99.16</td>
<td>4.48</td>
</tr>
<tr>
<td>212</td>
<td>74</td>
<td>51.13</td>
<td>17.65</td>
<td>48.71</td>
<td>45.24</td>
<td>97.47</td>
<td>14.81</td>
</tr>
<tr>
<td>300</td>
<td>39</td>
<td>56.09</td>
<td>19.98</td>
<td>60.22</td>
<td>0.00(+)</td>
<td>90.91</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\(^{(+)}\) multiple modes exist; smallest value is given

---

\(^5\) It was not possible to have students do the test in the first two weeks of classes because of movement between courses and sections. Therefore, the test could only be administered following the last day to add or drop courses. Because of the large number of students involved, the testing had to be spread over two weeks, but as the purpose of a placement test is to gauge the amount of knowledge a student enters a course with, not their knowledge one month into the course, students were not given any later opportunity to take the test after these two weeks had passed.
Figure 1. Boxplot of test scores per level (circles and asterisks represent extreme values)

An ANOVA test revealed a main effect for group, $F(1,1584)=655, p=.000$. However, as this statistic only shows an overall difference and does not compare between two levels, pairwise comparisons were calculated, yielding the results shown in Table 3.

Table 3. Pairwise Comparisons of Levels (* indicates $p<.05$)

<table>
<thead>
<tr>
<th></th>
<th>111</th>
<th>112</th>
<th>211</th>
<th>212</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>n/a</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>112</td>
<td></td>
<td>n/a</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>211</td>
<td></td>
<td></td>
<td>n/a</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>212</td>
<td></td>
<td></td>
<td></td>
<td>n/a</td>
<td>*</td>
</tr>
<tr>
<td>300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>n/a</td>
</tr>
</tbody>
</table>

The results show that the test distinguishes well between SPAN 111 and all other levels, between SPAN 112 and all other levels, but it does not distinguish among the SPAN 211, SPAN 212 and SPAN 300 levels. This result is not surprising if we consider what is taught in the various levels. Students in SPAN 111 start with no prior knowledge of Spanish; therefore, their learning would increase significantly between the beginning of SPAN 111 and the beginning of SPAN 112. Similarly, anything that students learn in SPAN 112 would be new; therefore, we would expect a significant difference between the beginning of SPAN 112 and the beginning of SPAN 211. In SPAN 211, SPAN 212 and SPAN 300, the objective is not so much the teaching of new material as the review and elaboration of preexisting knowledge. Consequently, it is
conceivable that students reach a “plateau” in their learning, where they do not accumulate new information as much as practice what they already know, both in terms of vocabulary and grammar. As a result, it is not surprising that the placement test does not distinguish between these levels as it assesses the quantity of knowledge (number of words the students purport to know) and not the quality (how well they know the words).

For placement purposes, the lexical decision test would therefore only be able to place students up to and including SPAN 211. From a pedagogical perspective, such a situation is not necessarily disadvantageous. Most students who enter the program with some knowledge of Spanish have learned it through travel and informal contact with native speakers. Therefore, they do not have the same type of knowledge as the other students who learned Spanish in a classroom. If these students skip ahead too many levels, any gaps in their knowledge will be compounded and could adversely affect their overall performance in a classroom situation. By limiting initial placement to no higher than SPAN 211, adjustments can be made more easily if it turns out that the student should be promoted to a higher level, rather than placing them higher and then “demoting” them to a lower level.

**ESTABLISHING CUT SCORES**

The final step involved establishing the minimum score a student needed to enter the SPAN 112 and SPAN 211 levels. As there are no objective criteria for determining a cut score, it was decided that the best solution was to have a panel of experienced instructors set the minimum score (known as the Angoff method). A version of the lexical decision test was generated randomly and four Spanish instructors with experience teaching a variety of levels were asked to determine how many of the 200 words a minimally competent student entering SPAN 112 and SPAN 211 should know. Their judgments were then averaged to determine the cut score for that level. Two rounds of judging were needed, as the first round yielded overly high cut scores that would have eliminated the majority of the students currently registered at that level (62% for SPAN 112 and 52% for SPAN 211). In contrast, the second round resulted in 81% of the current students in SPAN 112 and 91% of the students in SPAN 211 being accurately placed. The divergence between the first and second round of judgments suggested that it was likely that the judges underestimated minimal competency the second time knowing that they overestimated the first time. Therefore, the average of the first and second rounds of judging was considered to be a more representative cut score for the level.

**IMPLICATIONS AND CONCLUSION**

The lexical decision test appears able to distinguish well between lower levels of proficiency, although it did not seem to distinguish between higher levels of proficiency, thereby limiting its usefulness. The correlation between vocabulary size and overall language proficiency that was found in previous research only partially holds true in this study. More research is needed to determine which variables—such as the test format (online as opposed to on paper),

---

6 Students who have taken previous courses in Spanish are placed according to the prerequisite equivalent that they have completed, not according to their results on the lexical decision test.

7 SPAN 111 did not require a cut score as any student scoring below the minimum for SPAN 112 would automatically be placed in SPAN 111.
the time limit of ten minutes, or the sample of participants—may have led to a partial correlation instead of a full correlation.

Despite this limitation, there are numerous advantages to the lexical decision test that make it a good choice for a placement measure: it can be administered quickly and efficiently (in this case, via the Internet); it can test a large sample of words in a short time, thereby increasing the validity of the test; and it avoids tricky decisions about what students should know at each level. The lexical decision test can also be easily adapted for any language as long as a frequency list of words is available. Even though the test, as operationalized in this study, fails to distinguish beyond the low-intermediate level of proficiency, it does distinguish very well among the different beginner levels that have the largest enrolment numbers and therefore the greatest need for a placement test. Moreover, the final decision about a placement level is, and should always be, left in the hands of the instructor and/or the program coordinator as no single test is perfect and there are always exceptional cases that must be dealt with individually. However, as a general placement measure, the lexical decision test fulfills its purpose by allowing a “quick-and-dirty” decision about a student’s proficiency level with minimal administrative cost.

ACCESS TO LEXICAL DECISION TASK

The test is available at http://www.arts.ualberta.ca/~span/. Please email the researcher, yvonne.lam@ualberta.ca, for a temporary login ID so that we know not to take your record into consideration. We can also show you the administrator’s site if you are interested.

ACKNOWLEDGEMENTS

I would like to thank Cam Fraser, Magdalena Stanislawksa, and Karl Anvik for their help in creating the online version of the placement test. I would also like to thank Marilyn Abbott and Nizam Radwan for their advice on how to determine cut scores.

REFERENCES


