Yes/no tests for foreign language placement at the post-secondary level

Yvonne Lam
University of Alberta

Author Note
The creation of the yes/no placement test was funded by the University of Alberta's Teaching Research Fund, Project #03-3. I would like to thank Cam Fraser, Magdalena Stanislawska, and Karl Anvik for their help in creating the online test. I would also like to thank Marilyn Abbott, Nizam Radwan, Todd Rogers, and Patrick Bolger for their help on the statistical analysis. All remaining faults are mine.

Abstract
This study examines the use of a yes/no test as a placement measure in a university Spanish foreign language program. Yes/no tests measure word recognition and have been shown to correlate well with other language proficiency tests. The main advantage of using a yes/no test as a placement tool is its ease of creation and administration. We examined the ability of a yes/no test to discriminate between adjacent placement levels in our program and found that it functions up to the low-intermediate level only. This limitation may be due to the way vocabulary is taught, the stages of development of the lexicon, and the learning plateau that intermediate learners appear to reach. Nonetheless, the yes/no test allows for a quick initial sorting of students into different levels and reduces the need for individual placement by instructors.

Résumé
Cette étude porte sur l'usage d'un test binaire appelant une réponse par oui ou par non-comme instrument de classement pour un programme universitaire d'espagnole langue étrangère. Les tests binaires (oui vs. non) permettent la reconnaissance des mots et ils se sont révélés corrélés de près avec d'autres tests de compétence linguistique. L'avantage principale des tests binaires (oui vs.non) comme outils de classement est qu'ils sont faciles à élaborer et à administrer. On a examiné la capacité d'un tel test à distinguer entre des niveaux de classement contigus et on a trouvé que son efficacité se limite au niveau intermédiaire inférieur. Cette limite peut être attribuée à la manière dont le vocabulaire est enseigné, aux étapes de développement du lexique, et au plateau d'apprentissage que les apprenants intermédiaires semblent avoir atteint. Néanmoins, le test binaire **(**oui vs.non) nous permet de trier les étudiants selon leur niveau de façon préliminaire et rapide et il réduit le besoin de **c**lassement individuel par les instructeurs.

Yes/no tests for foreign language placement at the post-secondary level

Introduction

This study examines the feasibility of using a yes/no test for placement purposes in the Spanish-as-a-foreign-language program at the University of Alberta, Canada. Yes/no tests consist of a list of both real and pseudo words, and learners have to check off the words that they recognize. The underlying assumption for using a yes/no test for placement is that word recognition correlates with overall language proficiency; learners who recognize more words have likely had greater exposure to the language in general, and consequently, they should perform better in other language skills as well (Meara & Jones, 1988). This assumption has been verified in numerous studies that established relatively strong correlations between yes/no tests and more comprehensive measures of language proficiency in a second or foreign language (L2) (Fairclough & Ramírez, 2009; Harrington & Carey, 2009; Lam, Pérez-Leroux, & Ramírez, 2003; Meara & Buxton, 1987; Meara & Jones, 1988). In her review of studies on L2 word recognition, Koda (1996) also concludes that general linguistic knowledge is a necessary (though not sufficient) condition for word recognition competence. While some questions have been raised about the validity and reliability of yes/no tests, as discussed below, this format has clear practical advantages: it is easy to construct, a large number of words can be tested in a short time, the task demand on the learner is low, and the test can easily be administered and scored by computer. For these reasons, we considered using a yes/no test as the placement measure for our program.

This paper begins by discussing some of the issues with placement in L2 programs that have been identified in previous research. It then outlines the rationale for selecting a yes/no test as the placement measure in our program. As few studies have examined the ability of yes/no tests to discriminate between placement levels, the current study seeks to shed light on this issue by seeing if there are statistically significant differences in the test scores of 785 students across five levels of university Spanish. The benefits and limitations of using a yes/no test for placement are then discussed, along with a brief description of how the cut-scores for each level were determined. This study concludes that while the yes/no test does not provide a complete measure of a student's proficiency, it does offer enough of an assessment for an initial sorting of students into placement levels, and it does so in a manner that requires relatively few resources.

Previous Research

Issues in placement testing

Students entering a post-secondary language program often come from a wide variety of backgrounds. Some already have previous knowledge of the language, either through academic study or through travel, while others are true beginners. Even those who studied the language in primary and/or secondary school often reflect a range of proficiency levels that depend, to a large extent, on the quality and type of instruction they received. Such heterogeneity highlights the need for a placement measure for incoming students, as the more variation there is in a class, the more difficult it is for instructors to accommodate learners' needs. A diversity of levels is also

problematic for students, who may feel intimidated by their more competent classmates or, conversely, bored by the less proficient ones. Given the importance of the proper articulation of language programs (Byrnes, 1990), the appropriate placement of students can minimize the heterogeneity within a class so that instructional goals can be achieved in the most efficient and effective way.

Placement testing is, however, much easier said than done. The ideal placement test should provide an accurate measure of a student's linguistic abilities in accordance with the specific curricular objectives of the program (Lange, Prior, & Sims, 1992), and the preferable way to do so is probably via individual assessment by experienced instructors. In actual practice, resource limitations mean that such a procedure is essentially impossible to implement for all but the smallest-sized language programs due to the sheer number of enrollments. Thus, there is a conflict between efficiency and accuracy, with the demands of efficiency almost always taking precedence (Bernhardt, Rivera, & Kamil, 2004). In the end, the key question is not which placement test is the best overall, but which of the feasible tests is the most adequate, even though it may mean adopting a less fine-grained measure that only makes an initial sorting rather than an accurate placement recommendation (Wesche, Paribakht, & Ready, 1996).

This tension between preferred testing methods and resource constraints is reflected in the various placement measures adopted by foreign language programs. A significant number of colleges and universities simply rely on seat-time equivalencies, the number of language classes that the student has previously taken. For example, in their survey of 58 Spanish departments at American colleges and universities, Klee and Rogers (1989) found that 25% had no separate placement measure, a result confirmed in Wherritt and Cleary's (1990) survey of 126 Spanish programs. Although some correlation has been found between prior instruction and proficiency (Klee & Rogers, 1989; Lange et al., 1992), seat-time equivalencies fail to take into account variation in the quality of previous instruction, differences in individual learning experiences, the time lapse between high school and university, as well as the student's aptitude, ability, and motivation, all of which may affect placement. Students are also expected to self-assess their previous linguistic background, even though some students may not be able—or willing, if placing out of the program is an option—to do so accurately (LeBlanc & Painchaud, 1985; Wesche et al., 1996). Any necessary placement adjustment often cannot be made until the last minute, after classes have started and instructors have met the students for the first time. Moreover, some instructors may be reluctant to make these judgment calls because of non-academic reasons, such as students having already filled their schedules or having friends in the same course, or because the instructors lack the pedagogical experience to make quick assessments of language proficiency.

In order to avoid such problems, some programs have opted to administer commercially available tests, such as the College Board Subject Tests, the Modern Language Association's Proficiency Exams, and Brigham Young University's Computerized Adaptive Placement Exams (CAPE). Although these tests have been standardized and validated, they are not without drawbacks: they are set at pre-determined levels that may not correspond to a particular language program, they do not necessarily reflect the objectives of the program, and they are generally limited in scope (Klee & Rogers, 1989; Wherritt & Cleary, 1990). In fact, Klee and Rogers' (1989) survey found that the institutions that used the College Board's and the Modern Language Association's tests expressed a higher level of dissatisfaction than those who used measures produced in-house. Finally, perhaps the greatest drawback of commercial tests are the fees

charged for their use, which raises the difficult question of whether programs should cover these fees out of their own budget or simply pass them on to the students.

For this reason, numerous programs—approximately one-third of the departments surveyed by Klee and Rogers (1989) and by Wherritt and Cleary (1990)—have chosen to create an in-house test of their own. Such a decision, however, requires a substantial investment of time and labor, as well as expertise in language assessment. For instance, Wherritt, Cleary, and Druva-Roush (1990) reported that at the University of Iowa, the development of a placement and outcome measure required the half-time teaching release of a faculty member, the formation of an advisory committee, and the hiring of a full-time test developer. Not all language programs, unfortunately, have such resources at their disposal. As a result, many in-house tests seem to have been created with efficiency as the main criterion; despite a stated emphasis on communicative language teaching, the majority of these tests limit themselves to the receptive skills of reading and listening using discrete-point items, with far fewer programs assessing writing and even fewer testing speaking (Klee & Rogers, 1989; Wherritt & Cleary, 1990).[1] It is no surprise, then, that despite a general dissatisfaction with the existing in-house placement test, many programs surveyed by Wherritt and Cleary (1990) acknowledged that a lack of further resources prevented them from making the needed improvements.

Such issues suggest that many L2 programs are caught in a catch-22 situation when it comes to placement testing. Using a commercially available test may not correspond to the objectives of the program, but creating an in-house test is resource-intensive. The alternative of not having a placement measure at all places the burden of responsibility on instructors and coordinators at a time when they are busiest preparing for the start of classes. While most students seem to end up in some manner or another at an appropriate level even in the absence of a placement test, there is a clear advantage to having an objective measure as the starting point, both in terms of reducing the stress on instructors and indicating to students that placement is not an issue to be taken lightly.

Yes/no tests as a placement measure

The use of yes/no tests as a placement tool was first developed by Paul Meara and his colleagues, who were given the task of creating a placement test for a language program where there was a large and rapid turnover of students (Meara & Jones, 1988). Due to the high enrolment numbers, neither a test that assessed all four language skills, nor individual face-to-face assessment was feasible because of the time and resources required. A yes/no test, however, could address some of these practical problems (Meara & Buxton, 1987). Yes/no tests are easy to create; while much is still unknown about what types of language structures a learner should know at a particular stage, much more research has been done on how many words learners should know at different levels (Nation, 1990, 2001), so a list of testable words can be quickly compiled, especially if there is an existing lexical corpus for the language. Moreover, unlike other types of lexical tests such as multiple-choice, fill-in-the-blank, translation, or sentence-

---

[1] One could argue that there is no need to evaluate all four skills in a placement test. However, as Klee and Rogers (1989) observe, a test that ignores speaking and listening lacks face validity in light of the current emphasis on oral skills and communication. Programs that had tests that more closely reflected their curriculum and methodology seemed to be more content than those who did not.

writing (for examples, see Nation, 2001; Read, 2000), the yes/no format can test a large number of words in a short time, thus increasing its reliability, and the task demand on the learners is minimal, since they merely read the word and make a quick decision as to whether or not they recognize it. In addition, yes/no tests are easy to score: there are no subjective decisions about what is acceptable and what is not acceptable. A student's score is calculated using the formula (RY - IY) / (1 - IY), where RY is the proportion of real words selected and IY is the proportion of pseudowords selected (Meara & Buxton, 1987).[2] This formula corrects for random guessing, which is reflected by the claim to recognize a pseudoword; the assumption is that a student who selects a large number of pseudowords has a low threshold for recognizing words and, therefore, may have selected real words through lucky guessing, and thus the final score needs to be adjusted downwards accordingly (Meara & Jones, 1988). The simplicity of the yes/no format means that it can be adapted for administration via computer, thereby eliminating any possibility of scoring error, and learners can have instant feedback about their performance, a feature that Meara (1996) reported as being popular with students.

Vocabulary tests in general have been shown to be a robust predictor of overall language proficiency, both in first and second language acquisition (e.g. Anderson & Freebody, 1981; Staehr, 2008), as effective communication depends in large part on having the adequate words to express oneself (Meara, 1996). Yes/no tests, in particular, have correlated well with other measures of language proficiency. For L2 English, Meara and Buxton (1987) reported a statistically significant degree of association between their yes/no test and the Cambridge First Certificate Examination, with the yes/no test correctly predicting the results of all but five of the 26 candidates. Similarly, in Meara and Jones (1988), their yes/no test showed a moderate correlation with the Eurocentres Joint Entrance Test of listening comprehension, grammar, reading, and speaking. Moreover, when placement adjustments were needed, the majority were made in accordance with the yes/no test result rather than the entrance test result, confirming that the former is a valid measure of overall proficiency. Wesche, Paribakht, and Ready (1996) also observed moderate correlations between their yes/no test and their program's entrance test of reading and listening comprehension. More recently, Harrington and Carey (2009) found a consistent strength of association between a yes/no test and the school's battery of listening, writing, grammar, and speaking tests. They noted that while the yes/no test was less sensitive to differences between placement levels than the test battery, the differences were modest.

Likewise, in the case of L2 Spanish, Lam, Pérez-Leroux, and Ramírez (2003) and Fairclough and Ramírez (2009) found high correlations between a cloze test—a widely-used, though not uncontroversial, measure of general language proficiency (e.g. Heilenman, 1983; Hinofotis, 1980)— and a yes/no test. Finally, the yes/no format has also shown itself to be highly reliable, where reported in the above-mentioned studies. These results all suggest that the yes/no test is a feasible alternative to other more established measures of language proficiency, especially in light of its practical advantages over these test formats.

---

[2] Various scoring formulae for the yes/no test have been discussed and compared (Huibregtse, Admiraal, & Meara, 2002). The actual differences among them seem to be fairly small, however (Mochida & Harrington, 2006), and so we have chosen to continue using the formula originally proposed in Meara and Buxton (1987). Note that with this scoring formula, it is possible to obtain a negative score if the proportion of pseudowords selected is greater than the proportion of real words selected.

As with any test, the yes/no format has its critics (for a more detailed discussion, see Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001; Mochida & Harrington, 2006; Read, 2000). The primary concern is with the type of vocabulary knowledge that it measures. Yes/no tests deal only with passive word recognition rather than active vocabulary skills involving the appropriate use of the words (Meara, 1996). Words are presented in isolation, with no context, and it is possible for students to recognize a word without really knowing what it means or how to use it. As such, yes/no tests reveal little about the extent of underlying word knowledge (Mochida & Harrington, 2006). This limitation, however, does not necessarily detract from the usefulness of the test. Before a word can be understood and used appropriately in context, learners must first recognize the form; in fact, if learners are to extend the use of the word to novel situations, they must be able to recognize it in isolation (Cameron, 2002). Meara (1996) goes so far as to argue that it is unlikely that learners only recognize forms without having at least some idea of how these words might be used, since most people acquire words from exposure to the language, not from memorizing lists in the abstract. Thus, a measure of word exposure can be indicative of general vocabulary knowledge and overall language proficiency.

Another criticism of the yes/no test resides in how students interpret "knowing" a word; for some, this may mean merely recognizing the form, while others may interpret knowledge as being able to use the word effectively. This variable interpretation of the instructions may lead to differing degrees of conservativeness in the judgments; some students may check off a word simply because the form looks somewhat familiar, while others refuse to select it unless they can articulate a definition. For this reason, Mochida and Harrington (2006) suggest that the test instructions include an explicit explanation of what type of knowledge is being sought, such as being able to say the word's basic meaning in the form of a declaration. This variability in the threshold for acceptance is also reflected in student reaction to pseudowords. While the presence of pseudowords is to control for guessing, in that students with more liberal criteria for accepting words are more likely to select pseudowords than students who only select words they feel certain about, factors other than conservativeness of judgment, such as the language background of learners and their proficiency level (Meara, 1996; Mochida & Harrington, 2006), seem to influence the acceptability of pseudowords in ways that are not yet well understood. Given the central role of pseudowords in scoring a yes/no test, it is important to conduct further research on the basis for students' decisions before determining its reliability and validity.

Despite these problems, the stakes in placement testing are relatively low, as Read (2000) concedes, so the issues mentioned above do not necessarily have to be resolved for the yes/no test to be useful for making initial placement recommendations. Indeed, the correlation of the yes/no format with more elaborate tests of language proficiency and its considerably greater ease of creation and administration make it worthwhile to consider for large-scale placement testing.

Research Question

In order for the yes/no test to be a useful placement tool, it needs to discriminate between adjacent placement levels. There have been few studies on this question, and the results have been mixed. For instance, Lam et al. (2003) found a statistically significant effect for level ($p < .05$) in all three levels tested, and a plot of the individual raw scores revealed little overlap between adjacent levels. Fairclough and Ramírez (2009) also found a statistically significant difference between most adjacent levels of L2 students, but not between first and second year. Harrington and Carey (2009) found that, while there was a systematic increase in the mean score

across the five placement groups, the yes/no test was only able to distinguish between the adjacent groups of Upper Intermediate and Advanced at a level of statistical significance. When the five groups were conflated to three, however, the yes/no test could discriminate between both sets of adjacent groups at a level of statistical significance. In contrast, Wesche et al. (1996) did not even find a consistent increase in the mean from level to level, and none of the differences between contiguous pairs of means was statistically significant.

Given the inconsistent results, it is necessary to confirm the discriminatory ability of the yes/no format in order to determine if we can draw appropriate and meaningful inferences from it. Therefore, this study seeks to answer the following question: *Is a yes/no test able to discriminate between adjacent placement levels in our program?*

One reason for the mixed findings in previous studies may have been the small numbers of participants, leading to greater variance in the scores: Fairclough and Ramírez (2009) only tested 55 L2 students over four levels, Harrington and Carey (2009) tested 88 students over five levels, Lam et al. (2003) tested 96 students over three levels, and Wesche et al. (1996) tested 93 students over six levels. In contrast, we administered the lexical decision test to 785 students across five levels, as described in the following section.

Method

The yes/no test

To ensure a sufficient sampling of words, we tested a total of 200 items: 120 real words and 80 pseudowords, keeping with the proportions originally proposed by Meara and Buxton (1987). The real words are drawn from the computerized Spanish lexical corpus *LEXESP: Léxico informatizado del español*, which contains five million words drawn from Spanish-language newspapers, novels, scientific texts, and essays published between 1978 and 1995 (Sebastián, Martí, Carreiras, & Cuetos, 2000). The test items are taken from the most frequent 5,000 words in the corpus, as the 5,000-word level is situated at the boundary between high-frequency and common low-frequency words and covers approximately 89% of the words in a normal English text (Nation, 1990, 2001). Davies (2005) confirmed that Nation's estimates for English are comparable in Spanish, with the 6,000-word level translating to approximately 85% to 90% coverage of a typical Spanish text in his *Corpus del Español* (Davies did not calculate for the 5,000-word level). As text coverage figures vary according to the makeup of the corpus used for calculation, we do not have a precise estimate of what the 5,000 most frequent words from the *LEXESP* corpus represent, but we can hypothesize that they likely correspond to at least 80% coverage of a typical Spanish text, which is reasonable for placement testing.

The list of testable words considers all the inflected forms of a word as one entry: for example, *corrió* "he or she ran" and *corrimos* "we run/ran" are part of the word *correr* "to run"; therefore, students are only tested on their ability to recognize *correr*. Proper names (e.g. *Madrid*), acronyms (e.g. *ETA*, the Basque separatist movement), and interjections (e.g. *¡ay!*) were omitted. Multi-word lexical expressions that act as one unit, such as *por ejemplo* "for example," were retained, as were grammatical words like prepositions. While it could be argued that grammatical words should be omitted because they carry no lexical meaning but rather express grammatical relationships, in practice it can be difficult to determine what is a grammatical word and what is a content word. As an example, the Spanish preposition *a* has a purely grammatical function when it is used to indicate a definite, animate direct object, as in

Conozco a Pedro "I know Pedro," but it can also contribute lexical information when it expresses the direction in which an object is moving, as in *Voy a la tienda* "I'm going to the store." Moreover, a true beginner with no prior exposure to Spanish is unlikely to recognize grammatical words, and the chances of many grammatical words appearing on the test are low because we sample from groups of words in descending order of frequency, and most grammatical words are clustered in the same part of the list due to their high frequency of occurrence.

Spanish-English cognates—words that have a similar form in both languages (Carroll, 1992)[3]—also remain on the list because of the difficulties in deciding what constitutes a cognate. For example, while one can easily say that the English *hotel* and the Spanish *hotel* are cognates, it is less clear whether *suggest* and *sugerir* should also be cognates. As for the English word *number*, it looks more similar to the Spanish *nombre* "name" than to the actual Spanish equivalent *número*, and so it is unclear which word should be omitted if cognates are deleted. In addition, there have been conflicting findings about whether L2 learners are actually sensitive to cognates in a word recognition task: although controlled laboratory studies have shown that cognates are recognized more quickly than non-cognates (e.g. de Groot, Borgwaldt, Bos, & van den Eijnden, 2002; Lemhöfer, Dijkstra, Schriefers, Baayen, Gainger, & Zwitserlood, 2008), classroom studies have found that learners do not always recognize cognates consistently (e.g. Lightbown & Libben, 1984; Moss, 1992). Moreover, the recognition and use of cognates is part of communicative proficiency, and so knowledge of cognates should not be ignored in testing.

As for the 80 pseudowords, they consist of items that are not part of the Spanish or English lexicon. Forty of the words contain identifiable Spanish morphemes, such as *perroso* which contains the word *perro* "dog" and the derivational suffix *–oso* found in other denominal adjectives such as *horroroso* "horrific." The other 40 pseudowords contain no identifiable Spanish morphemes, such as *guflaitas*. The pseudowords were checked carefully with a native speaker of Spanish to ensure that they looked and sounded like possible Spanish words.

The placement test was administered online. The 200 test words were displayed 40 words at a time so as not to overwhelm the test taker. In order to ensure a balanced sampling of higher and lower frequency real words, the 5,000-word master list was divided into groups of 40, in descending order of frequency, and the computer randomly chose one word from each group of 40. As for the 80 pseudowords, the same ones were used on each test, but they were interspersed in random positions by the computer. Thus, it was unlikely—though still possible—that two students received the same 200 words in the same order, as no two tests should be identical in content or in format.

Students were instructed, in English, to click the box that preceded every item that they thought was a valid Spanish word or phrase. While there was no limit on the amount of time a student could spend on a particular word, a total of ten minutes was allotted for all 200 words, with a clock at the top of the screen showing 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 elapsed, they were given the choice of reviewing their answers or submitting them without changes. In this way, students were encouraged to balance the demands of speed and

---

[3] Cognates are generally defined as word that share the same etymology. As Carroll (1992) argues, however, L2 learners are unlikely to be aware of the etymology of words; rather, they tend to assume that if two words have a similar form, they must mean the same, even if not true, as in the case of "false friends."

accuracy: we did not want students to think too much about whether they recognized a word, nor did we want them to make overly hasty judgments about what they recognized, since both these behaviors would affect the accuracy of their self-assessment.[4]

Participants

We administered the yes/no test to all students registered in a Spanish language course during the fall and winter terms of one academic year. There were 323 students from SPAN 111 (Beginners' Spanish I); from SPAN 112 (Beginners' Spanish II), there were 254 students; from SPAN 211 (Intermediate Spanish I), 96 students; from SPAN 212 (Intermediate Spanish II), 74 students; and from SPAN 300 (Advanced Spanish), 38 students. The notably higher number of enrolments in SPAN 111 and 112 is due to the fact that all students in the Faculty of Arts are required to take two foreign language courses. Because of the large numbers, we did not control for student background, such as age, gender, knowledge of another language, or study abroad experience. However, heritage speakers were excluded, given that they do not seem to perform in a comparable manner to L2 students on yes/no tests (Fairclough & Ramírez, 2009), and because they are not permitted to take the standard sequence of language courses but rather are streamed into a separate course, regardless of proficiency level.

Procedure and data analysis

The test was administered in the third and fourth week of classes in each term, immediately after the deadline for changing courses had passed but before students had advanced too far into the course. We assumed that the levels were distinct, based on our then placement method of seat-time equivalencies plus individual adjustments made by instructors and coordinators. Thus, if the yes/no test is a valid measure of proficiency, there should be statistically significant differences in the scores between adjacent levels.

As an additional measure of validity, we compared the continuing students who took the test in both the fall and the winter terms. We assumed that the number of words that these students recognized increased between the two courses; therefore, if the yes/no test is able to capture differences in proficiency level, the same student should score significantly higher on the test from one level to the next. Although there was a possibility of a test effect, the fact that each test contained different words in different orders mitigated such an effect. There were 163 students from SPAN 111 who continued directly to SPAN 112, and 44 students from SPAN 211 who carried on to SPAN 212 in the following term.

Finally, we conducted a reliability analysis. Given that each test was generated randomly by the computer, and there was no record kept of which specific words were selected for a student, it was not possible to conduct a standard reliability analysis using Cronbach's alpha. The studies cited earlier all established high degrees of reliability for their yes/no test, regardless of

---

[4] The decision to impose a ten-minute limit—an average of three seconds per word— was an arbitrary decision on our part. The use of yes/no tests in psychology generally involves measuring response time as well, and so there is no need for a time limit. Meara's version of the test (Meara & Buxton, 1987; Meara & Jones, 1988) simply asks the student to keep track of how long it takes them to do the test rather than setting a limit. For our purposes, a time limit was necessary as students were taking the test under unsupervised conditions.

the precise format, and so there does not seem to be a reliability issue with yes/no tests in general. It was possible, however, to examine the false alarm rate, the proportion of "yes" responses to the 80 pseudowords that are supposed to control for guessing and provide a more meaningful score. These false alarm rates can provide an indirect indication of the reliability and interpretability of the yes/no test results (Harrington & Carey, 2009).

Results

Test validity

In order to determine whether the yes/no test was able to measure differences in proficiency between adjacent levels, we first calculated the mean, the standard deviation, and the median score for each level (Table 1). For this part of the analysis, we did not control for the same student taking the test in the fall term at one level and then again in the winter term at the next level because the objective was not to track student development over time, but rather to obtain an idea of the typical performance of students at each level. The results show that there was indeed an increase in the mean score across levels, especially up to and including SPAN 211; the size of the increase diminished beyond SPAN 211, although there was still a gain in the mean score (Table 2). There was also a fair amount of variation within each group (Table 1), but the assumption of homogeneity of variance was met (Levene's test, $p > .05$), and the distribution of scores within each group was largely normal, and so the raw scores were not screened for possible outliers.

Table 1

Yes/No Test Scores

Level	**N**	**Mean**	**SD**	**Median**
SPAN 111 (Beginners' Spanish I)	323	13.71	18.92	12.24
SPAN 112 (Beginners' Spanish II)	254	26.79	15.96	26.39
SPAN 211 (Intermediate Spanish I)	96	46.07	16.85	46.79
SPAN 212 (Intermediate Spanish II)	74	51.13	17.65	48.71
SPAN 300 (Advanced Spanish)	38	57.57	17.97	60.66

Table 2

Gains in Mean Test Scores Between Levels (differences between consecutive levels in bold)

	SPAN 111	**SPAN 112**	**SPAN 211**	**SPAN 212**	**SPAN 300**
SPAN 111		***+13.08**	*+32.36	*+37.42	*+43.86
SPAN 112			***+19.28**	*+24.34	*+30.78
SPAN 211				**+5.06**	*+11.50
SPAN 212					**+6.44**

Note. * $p < .05$

A one-way analysis of variance revealed an effect for group, $F(4,780) = 137.63$, $p < .05$, $\eta^2 = .414$, which indicated that the scores between at least two groups were significantly different. Post-hoc comparisons using Tukey HSD showed that, at the .05 level of significance, the test distinguished well between SPAN 111 and all other levels, between SPAN 112 and all other levels, and between SPAN 211 and SPAN 300, but it was unable to distinguish between SPAN 211 and SPAN 212, nor between SPAN 212 and SPAN 300 (Table 2). Therefore, our yes/no test can only discriminate between adjacent levels up to and including SPAN 211.

When we used a paired-samples *t*-test to compare the continuing students who took the test in both the fall and the winter terms, our hypothesis was again only partially confirmed.

Students from SPAN 111 who continued to SPAN 112 did indeed have a significantly higher score when taking the yes/no test a second time, $t(162) = 5.96$, $p < .05$, $\eta^2 = .179$. In contrast, students from SPAN 211 who continued to SPAN 212 did not show a significant improvement in their test score, $t(43) = 1.34$, $p > .05$, $\eta^2 = .039$, which corroborated our findings that the test could not distinguish between adjacent levels beyond SPAN 211. Possible explanations for this limitation will be discussed later.

Test reliability

Table 3 shows the mean false alarm rate for each level as a percentage, together with the standard deviation. All levels showed similar false alarm rates, unlike previous studies where lower proficiency learners showed higher rates because they had less of a basis on which to distinguish real words from pseudowords (Mochida & Harrington, 2006). These rates indicate that the students were claiming to know almost one out of every three pseudowords, which suggests a liberal response strategy and a relatively low threshold for acceptance of a word. The implications of this finding will be discussed below.

Table 3

False Alarm Rate

Level	N	Mean	SD
SPAN 111	323	32%	23%
SPAN 112	254	27%	20%
SPAN 211	96	26%	19%
SPAN 212	74	28%	21%
SPAN 300	38	34%	21%
Overall	785	29%	21%

Discussion

Use of yes/no tests for foreign language placement

In response to our research question about whether a yes/no test can distinguish between adjacent placement levels, the findings reported above, like previous studies, provide a mixed answer. Our test functions as desired only up to the low-intermediate level in our program. This finding is in contrast to those in Lam et al. (2003), who were able to discriminate between all adjacent levels, and in Fairclough and Ramírez (2009), who could distinguish between intermediate and advanced levels, but not between beginner and intermediate, which they attributed to a lack of correspondence between the vocabulary tested and the vocabulary taught at the beginner level. Nonetheless, our findings do correspond to the results reported in Wesche et al. (1996), where the authors found that their yes/no test correlated significantly with placement at the intermediate level only, but not at the advanced level (they did not administer the test to the beginner level). Harrington and Carey (2009) also noted that their yes/no test had the greatest difficulty discriminating between the elementary, lower intermediate, and upper intermediate levels. Neither, however, offered any explanation of why yes/no tests seem to fail to distinguish changes in proficiency once students reach the intermediate levels.

The fault may lie with the test itself. The yes/no test is a measure of vocabulary recognition, and while the number of words one recognizes can be indicative of overall language proficiency, it is possible for proficiency to develop without a substantial increase in the number of words recognized. If we consider how vocabulary is taught in our program, in SPAN 111 and SPAN 112, students have little or no prior exposure to Spanish; therefore, the number of words they recognize should increase rapidly. At the intermediate levels, however, the focus is not so much on introducing new material as it is on reviewing, consolidating, and expanding on previous knowledge, so students are exposed to fewer new words and concentrate instead on practicing those they already know. As a result, even though vocabulary and general language proficiency continue to develop, the yes/no test would not be ideal to distinguish changes at these levels, since it assesses simple word recognition as opposed to how well students actually know these words and are able to use them. Moreover, Meara (1996) suggests that as the lexicon grows

larger, the number of words is of less importance than the way in which the items are organized. In addition, the yes/no test is limited to the first 5,000 most frequent words given in the corpus, which are most likely the words encountered by students at the lower levels as they deal with simple, everyday situations and texts. In contrast, at the higher levels, the texts become more specialized in preparation for content courses; thus, the new vocabulary that students are learning come from the lower frequencies, which are not targeted by the yes/no test.

It is also possible that it may simply be difficult to measure gains in proficiency at the intermediate levels. Richards (2008) noted that once learners arrive at the intermediate levels, they appear to reach a plateau in their learning and make little observable progress. Such an observation does not mean that there is no improvement in proficiency, but that these improvements may be such that it is difficult to find a test capable of measuring them reliably. This learning plateau is acknowledged implicitly by Brigham Young University's popular CAPE test, which states that it is designed primarily for placement into the first three semesters of college-level study only (Perpetual Technology Group, 2008). Thus, it is possible that placement at the different intermediate levels cannot be accomplished reliably via testing.

From a pedagogical and administrative perspective, limiting initial program placement to the low-intermediate level is not necessarily problematic. Most students who enter the program with some knowledge of Spanish and no prior coursework have learned it through travel and informal contact with native speakers, and the diversity of their previous experiences makes face-to-face assessment preferable to online testing. Moreover, although these students may recognize more words than classroom learners due to their greater exposure to the language, they may lack some of the explicit, academic knowledge of Spanish that is required of upper-level foreign language students in a university setting. If these students skipped ahead too many levels, any gaps in their knowledge could be compounded and adversely affect their overall performance, especially in third- and fourth-year content courses where they are expected to read and write academic texts. By restricting initial placement to no higher than SPAN 211, adjustments can be made more easily if it turns out that the student should indeed be promoted to a higher level, as opposed to placing them at a higher level first and then demoting them.

One could argue that since our yes/no test can only sort students into the first three placement levels and not into the remaining two levels, we should not bother with using it at all; the previous method of seat-time equivalencies plus instructor modifications accomplished the same goal of dividing students according to their proficiency level, and placement was not restricted to the first three levels. There are also other types of placement tests which, though more expensive and time-consuming, may work better at discriminating between adjacent placement levels. Despite its limitations, the use of the yes/no test does still confer several practical advantages that make it worthy of consideration.

First, the test does work well for placement into the beginner and low-intermediate levels in our program, which have the largest number of enrolments. Thus, it reduces the number of individual assessments that instructors need to deal with. Indeed, there have been fewer straight-forward placement queries, which means that instructors and coordinators spend less time counseling individual students and can focus their attention on true problematic cases, such as students who have lived abroad for an extended period of time. Second, the test results provide a concrete starting point from which instructors and coordinators can make adjustments. In this sense, the purpose of the test is not to take the place of instructors but to “help decision makers make the best judgments they can regarding human performance under a set of constraints” (Bernhardt et al., 2004, p. 357). In fact, in Klee and Roger’s (1989) survey of Spanish

departments, almost all programs (97%) made adjustments after the placement test was given, regardless of the type of test used, and our students are explicitly told that the placement is a recommendation only. Third, rather than relying on instructors to flag misplaced students after the start of classes, some of these students now identify themselves ahead of time because they disagree with their placement recommendation, thus allowing us to make adjustments well before the beginning of the semester. Finally, the implementation of a test has lent legitimacy to the placement process. Course selection is no longer done at the whim of the student who registers in whatever level he or she feels is appropriate or convenient; rather, there is an objective measure of proficiency on which placement decisions can be based.

In light of these arguments, the yes/no test does serve a purpose for our program, and its ease of administration as well as the limited resources that it demands give it an advantage over other test formats. It is not a perfect placement test, but one could say that such a test does not exist because proficiency levels constitute a continuum, and there are always grey areas or borderline cases that a placement test simply cannot be sensitive to and that require the intervention of a human being (Teschner, 1990). Indeed, it is not unreasonable for a placement test to be more effective at some levels than others when dealing with a large group of students with a wide range of abilities (LeBlanc & Lally, 1997). Harrington and Carey (2009) arrive at a similar conclusion regarding their yes/no test:

> Given the speed of the format and the advantages of computer-driven testing, the test warrants further attention as an alternative or complement to existing measures. The results do not suggest that the Yes/No test can completely replace the global assessment of the student's proficiency available through spoken and written production, though it can complement this testing and possibly reduce the need for extensive testing in other areas. (p. 623)

Issues with test design

Since the implementation of the test, we have noticed that a few students who have studied another Romance language scored highly on the test simply by recognizing cognates. Meara (1996) has made the same observation regarding the performance of French native speakers on a yes/no test of English. At the moment, it is not clear how this issue can be resolved. Given the previously mentioned difficulty of determining what is a cognate, the fact that languages naturally borrow words from each other, and the diverse language backgrounds of our students, the elimination from the test of any possible cognate with English or another

Romance language would leave us with very few words to test in Spanish. Moreover, the recognition and appropriate use of cognates is part of overall language proficiency, so we should not discount the role of cognates altogether. Thus, rather than remove possible cognates from the test, we should leave these students—who usually identify themselves because they placed higher than expected—for instructors to place on an individual basis.

Another issue lies with the high false alarm rates. While the inclusion of pseudowords allows scores to be adjusted to reflect lax criteria for selecting words, such high rates still cast some doubt on the reliability of students to self-assess their knowledge and suggest that they are overestimating their knowledge, which could lead to higher placements than appropriate. The wording of the instructions may have been a contributing factor: students were asked which words they thought were valid in Spanish, and as all the pseudowords are potentially possible

words, with half of them even containing identifiable Spanish morphemes, students may have been confused between identifying words that actually exist and words that could exist, both of which could fit the definition of *valid* and result in broader criteria for selecting words. It is unclear, however, whether rewording the instructions more stringently affects the false alarm rate, and encouraging a conservative response strategy may have the opposite effect by causing learners to underestimate their knowledge (van Ee, 2007). Nonetheless, given the relatively low-stakes nature of placement testing and the opportunity for adjustments to be made afterwards, high false alarm rates do not necessarily invalidate the use of yes/no tests for an initial sorting of students into placement levels.

Setting cut scores

Once we confirmed that our test was capable of discriminating between adjacent placement levels in an acceptable manner, the final step was to determine the cut-scores for SPAN 112 and SPAN 211; that is, the minimum score that a student needs in order to be placed at that level. (SPAN 111 does not require a cut-score as it includes, by default, all students who place below the minimum for SPAN 112.) There have been several methods used to set cut-scores on placement tests. For example, Aleamoni and Spencer (1968) and Hagiwara (1983) used the median score for each level as cut-off points for the next course. Wherritt et al. (1990) used a loss function approach, in which they set a cut-score that minimized false negatives (students predicted to fail when they actually succeeded) and false positives (students predicted to succeed when they actually failed), where success was measured as a grade of C+ or higher in the course. They also suggested examining the mean grade across individuals within a course for each test score and setting the cut-score at the point where the mean grade becomes unacceptable. Livingston and Zieky (1982) and Zieky and Perie (2006) summarize several other methods of establishing cut-scores on tests of educational achievement, based on judgments about the test questions or about the test-takers.

After consultation with a psychometrician, we chose the Modified Angoff method because of its simplicity and its well-established history, especially for tests like the yes/no format that have dichotomously scored items (Morgan & Michaelides, 2005; Ricker, 2006). This method involved generating a version of the yes/no test and asking experienced Spanish instructors to decide which of the words a minimally competent, or "borderline" (Livingston & Zieky, 1982), student at the given level would be likely to recognize. The judgments of the individual instructors were then averaged to produce a cut-score. Two rounds of judging were needed in our case, as the first yielded overly high cut-scores that more than half of the students then registered at that level did not meet. Therefore, the judges were asked to revise their notion of a minimally competent student, with the second set of judgments yielding lower cut-scores that over four-fifths of the students registered at that level met. As it was likely that the judges underestimated minimal competency in the second round based on the knowledge that they overestimated it in the first round, we averaged both rounds of scores to produce the final cut-scores.

While the validation of these cut-scores is beyond the scope of this paper, we will present a cursory analysis of their viability by examining the final marks of students who scored at or above the cut-scores with those who scored below them. We recognize that final marks are not necessarily a reliable measure of appropriate placement, as students who are poorly placed still have the possibility of performing well in a course, while students who are appropriately placed

may do badly due to other variables such as differences in motivation, learning styles, and the amount of time dedicated to coursework. There may also be variation in grading practices among instructors; that is, performance that one instructor considers an A grade may be assigned a B grade by another. Nonetheless, if our cut-scores are tenable, we should see a general pattern where students who scored at or above the cut-score had higher mean final marks than those who scored below it. This was indeed the case: students in SPAN 112 and SPAN 211 who did not have the minimum score performed an average of 5% poorer than those who did score above the cut; the difference is significant for SPAN 112 ($p < .05$) and almost significant for SPAN 211 ($p = .06$). Therefore, it appears that our cut-scores are workable for the moment, although they will certainly need further empirical validation as well as revision in order to take into account changes in the course curricula and student performance over the years.

Conclusion

Placement testing seems to be one of the banes of post-secondary L2 programs. Most programs recognize the importance of having some sort of proficiency measure for incoming students; yet, limited resources prevent them from implementing a satisfactory placement test. The solution that we chose was an online yes/no test that manages to balance the demands of both accuracy and efficiency. The yes/no test is by no means a perfect placement test, but it does fulfill its purpose of providing a preliminary placement recommendation that discourages students from registering themselves in an uninformed way, and that gives instructors and coordinators a basis on which to make further adjustments if needed. The relative ease with which a yes/no test can be created and administered makes it a viable option for programs that do not have many resources to dedicate to placement testing.

References

Aleamoni, L., & Spencer, R. (1968). Development of the University of Illinois Foreign Language Placement and Proficiency System and its results for Fall, 1966 and 1967. *Modern Language Journal, 52*, 355-359.

Anderson, R., & Freebody, P. (1981). Vocabulary knowledge. In J. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77-117). Newark, Delaware: International Reading Association.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the Yes/No vocabulary test: Some methodological issues in theory and practice. *Language Testing, 18*(3), 235-274.

Bernhardt, E., Rivera, R. J., & Kamil, M. L. (2004). The practicality and efficiency of Web-based placement testing for college-level language programs. *Foreign Language Annals, 37*(3), 356-366.

Byrnes, H. (1990). Priority: Curriculum articulation. *Foreign Language Annals, 23*(4), 281-292.

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research, 6*(2), 145-173.

Carroll, S. E. (1992). On cognates. *Second Language Research, 8*(2), 93-119.

Davies, M. (2005). Vocabulary range and text coverage: Insights from the forthcoming *Routledge Frequency Dictionary of Spanish*. In D. Eddington (Ed.), *Selected proceedings of the 7th*

Hispanic Linguistics Symposium (pp. 106-115). Somerville, MA: Cascadilla Proceedings Project.

de Groot, A. M. B., Borgwaldt, S., Bos, M., & van den Eijnden, E. (2002). Lexical decision and word naming in bilinguals: Language effects and task effects. *Journal of Memory and Language, 47*(1), 91-124.

Fairclough, M., & Ramirez, C. J. (2009). La prueba de decisión léxica como herramienta para ubicar al estudiante de español en los programas universitarios [The lexical decision test as a tool for placing Spanish students in university programs]. *Íkala, revista de lenguaje y cultura [Íkala, journal of language and culture], 14*(21), 85-99. Retrieved from http://aprendeenlinea.udea.edu.co/revistas/index.php/ikala/article/view/2666

Hagiwara, P. (1983). Student placement in French: Results and implications. *Modern Language Journal, 67*, 23-32.

Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System, 37*, 614-626.

Heilenman, L. (1983). The use of a cloze procedure in foreign language placement. *Modern Language Journal, 67*(2), 121-126.

Hinofotis, F. (1980). Cloze as an alternative method of ESL placement and proficiency testing. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 121-128). Rowley, MA: Newbury House.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing, 19*(3), 227-245.

Klee, C., & Rogers, C. (1989). Status of articulation: Placement, advanced placement credit, and course options. *Hispania, 72*(3), 763-773.

Koda, K. (1996). L2 word recognition research: A critical review. *Modern Language Journal, 80*(4), 450-460.

Lam, Y., Pérez-Leroux, A. T., & Ramírez, C. (2003). *Using lexical decision for Spanish language placement testing*. Paper presented at the 2003 conference of the American Association for Applied Linguistics.

Lange, D., Prior, P., & Sims, W. (1992). Prior instructions, equivalency formulas, and functional proficiency: Examining the problem of secondary school-college articulation. *Modern Language Journal, 76*(3), 284-294.

LeBlanc, L. B., & Lally, C. G. (1997). Making the transition from secondary to postsecondary Spanish study: Achieving consistency in college placement for Florida's students. *Hispania, 80*(1), 124-135.

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

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Gainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(1), 12-31.

Lightbown, P. M., & Libben, G. (1984). The recognition and use of cognates by L2 learners. In R. Andersen (Ed.), *Second languages: A cross-linguistic perspective* (pp. 393-417). Rowley, MA: Newbury House Publishers.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. USA: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/passing_scores.pdf

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer & J.

Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35-53). Cambridge, UK: Cambridge University Press.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*, 141-154.

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society* (pp. 80-87). London: Centre for Information on Language Teaching and Research.

Mochida, A., & Harrington, M. (2006). The yes/no test as a measure of receptive vocabulary knowledge. *Language Testing, 23*(1), 73-98.

Morgan, D. L., & Michaelides, M. P. (2005). *Setting cut scores for college placement*. New York: The College Board. Retrieved from http://professionals.collegeboard.com/profdownload/accuplacer-setting-cut-scores.pdf

Moss, G. (1992). Cognate recognition: Its importance in the teaching of ESP reading courses to Spanish speakers. *English for Specific Purposes, 11*(2), 141-158.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston: Heinle & Heinle.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

Perpetual Technology Group (2008). *Welcome to CAPE (Computer Adaptive Placement Exam)*. Retrieved from https://www.aetip.com/Products/CAPE/CAPE2.cfm

Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.

Richards, J. C. (2008). *Moving beyond the plateau: From intermediate to advanced levels in language learning*. New York: Cambridge University Press. Retrieved from http://www.cambridge.org/other_files/downloads/esl/booklets/Richards-Beyond-Plateau.pdf

Ricker, K. L. (2006). Setting cut-scores: A critical review of the Angoff and modified Angoff methods. *The Alberta Journal of Educational Research, 52*(1), 53-64.

Sebastián, N., Martí, M. A., Carreiras, M. F., & Cuetos, F. (2000). *LEXESP: Léxico informatizado del español [LEXESP: A computerized lexicon of Spanish]*. Barcelona: Edicions Universitat de Barcelona.

Staehr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*(2), 139-152.

Teschner, R. V. (1990). Spanish speakers semi- and residually native: After the placement test is over. *Hispania, 73*(3), 816-822.

van Ee, J. H. (2007). "*You call that a word?" The effect of cognates on the Recognition-Based Vocabulary Test*. (Unpublished doctoral dissertation). Utrecht University, Utrecht, The Netherlands. Retrieved from http://igitur-archive.library.uu.nl/student-theses/2007-1009-200738/UUindex.html

Wesche, M., Paribakht, T. S., & Ready, D. (1996). A comparative study of four ESL placement instruments. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 199-209). Cambridge: Cambridge University Press.

Wherritt, I., & Cleary, T. A. (1990). A national survey of Spanish language testing for placement or outcome assessment at B.A.-granting institutions in the United States. *Foreign Language Annals, 23*(2), 157-165.

Wherritt, I., Cleary, T. A., & Druva-Roush, C. A. (1990). Development and analysis of a flexible Spanish language test for placement and outcome assessment. *Hispania, 73*(4), 1124-1129.

Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement*. USA: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf