Aligning Commercially Available Placement Test Scores to an Existing Language Program[1]

María Georgina Clark Rivas[2], Universidad de Sonora

Jorge Alejandro Villa Carrillo[3], Universidad de Sonora

Abstract

The use of technology for testing is a current tendency in English language teaching. Several universities are developing computer assisted language tests and others are using commercially developed tests for their language programs. The researchers selected the Oxford Online Placement Test (OOPT) as a possible alternative to using a paper and pencil placement instrument for the Department of Foreign Languages at the University of Sonora. Our hypothesis was that scores of a commercially available placement test can be aligned to an existing language program using the weighted average criterion. For this purpose, data was collected from the OOPT which was administered to 555 students in the different levels of the English program. Great variability was observed in the scores; therefore a criterion of weighted average was used to calibrate the instrument. Data that were too distant from the target course were censored. In using this criterion, if the weighted average is zero or near zero the interval score was accepted. For each set of intervals scores proposed, a set of weighted average was obtained. The objective was to be fair in the classification considering what the system is producing. In addition, a bootstrap statistical process was used to capitalize the variability of the sample with the objective of having an estimate of the confidence limits of the weighted average. Two sets of intervals were developed by the research team and a third set was recommended by Oxford University Press (OUP).

Resumen

El uso de la tecnología en evaluación es una tendencia actual en la enseñanza del inglés. Varias universidades están desarrollando exámenes asistidos por computadoras y otras utilizan exámenes desarrollados comercialmente para sus programas de idiomas. Los investigadores seleccionaron el *Oxford Online Placement Test (*OOPT) como una posible alternativa al uso del examen de colocación de papel y lápiz para el Departamento de Lenguas Extranjeras de la Universidad de Sonora. Nuestra hipótesis es que los intervalos de puntaje que corresponden a cada nivel del programa pueden ser determinados aplicando el criterio de promedio ponderado a los puntajes producidos por un examen comercial. Para este propósito se recogieron datos del instrumento de colocación administrado a 555 estudiantes inscritos en los diferentes niveles del programa de inglés. Debido a la gran variabilidad observada en los puntajes, se utilizó el promedio ponderado para calibrar el instrumento. Datos demasiado distantes del curso meta fueron censurados. Al usar este criterio, si el promedio ponderado es cero o cercano a cero, entonces puede aceptarse como representativo. Para cada conjunto de intervalos propuesto, un conjunto de promedio ponderado fue obtenido. El objetivo fue ser justo en la clasificación considerando lo que el sistema produce. Además, se utilizó el proceso estadístico *bootstrap* para capitalizar la variabilidad de la muestra con el fin de tener una estimación de los límites de confianza del promedio ponderado. En la calibración del examen al programa se desarrollaron dos conjuntos de intervalos por el equipo investigador y un tercer conjunto fue recomendado por Oxford University Press (OUP).

Introduction

English language assessment has been enhanced by the use of technology. Nowadays, it is possible to evaluate students' language proficiency in a fast and efficient way having results in real time. Additionally, it is also possible to evaluate a more complete set of

[1] This is a refereed article.

[2] mclark@lenext.uson.mx

[3] jorge2155485@gmail.com

language components and features. Designing a test has also become a possibility for language institutions. Several universities, as well as testing organizations, have developed language tests for their own use, but they have also published tests that are available to language programs worldwide. Examples of published tests include the findings of an exploratory survey into 80 English as a Second Language (ESL) programs for matriculated students in the U.S. where the Educational Testing Service (ETS) found that 51 different examination instruments were used for placement purposes, 36 of which were commercially developed tests (Ling, Wolf, Cho & Wang, 2014).

However, using published tests may not be that simple. The user of a published test needs to consider the particular characteristics and the recommended score intervals that the test developer offers. These characteristics may not correspond to the context where test administrators intend to use a published test. Bernhardt, Rivera and Kamil (2008) have stated the ideal situation when they discuss that "a placement test must be aligned with the curriculum to the extent that a student will improve both in language proficiency and in the score on the placement test as a result of having taken the optimal sequence of courses" (p.358). Although the alignment can be achieved as the curriculum is being developed for a particular program in a context where a program has been implemented for a number of years, it may also be achieved by identifying the score intervals in a placement instrument that represents the outcome of a program.

The integration of Mexican public universities to the global community has led to changes in institutional policies on foreign language education. In 2004, the University of Sonora established a policy, General Guidelines for a Curricular Model of the University of Sonora, *(Lineamientos Generales para un Modelo Curricular de la Universidad de Sonora, 2003),* which requires matriculated students in any undergraduate program to attain the English language competence equivalent to an A2 level according to the Common European Framework of Reference. The institutional policy on foreign language education, offers several options for the accreditation of this requirement. However, the preferred option to fulfill this requirement has been to obtain the credit for having approved the level four course of the English program in the Foreign Language Department. This fact has resulted in an increase in student population and a demand for more efficient ways of placing students in their corresponding classes.

A published computer adaptive test (CAT) was considered as an alternative to the paper and pencil placement test currently used by the Department of Foreign Languages of the University of Sonora. One of the reasons is that a CAT is context free and as a consequence has the advantage of using fewer items to estimate the examinee's competence of the language. In other words, Brown (1997) states that, "IRT (Item Response Theory) can provide item-free estimates of students' abilities" (p.44).

Having selected a CAT, the next step was to develop a method for identifying the score intervals of a commercially available computer-based placement test in alignment to the outcome of the General English Program offered by the Department of Foreign Languages. Three sets of intervals were proposed for course placement in the General Courses of English from the Department of Foreign Language (Table 1): two by the

research team based on the scores of student performance in the sample collected, and one recommended by Oxford University Press based on the textbook currently used.

	Course levels in the General English Program								
Set of intervals	1A	1B	2	3	4	5	6	7	Advanced
Lenext 1	0-5	6-10	11-18	19-28	29-37	38-45	46-54	55-64	65+
Lenext 2	0-3	4-7	8-16	17-25	26-35	36-45	46-55	56-65	66+
OUP	0-5	6-10	11-19	20-28	29-37	38-47	48-56	57-65	66+

Table 1. Intervals proposed for course placement.

Developments in Computer-Assisted Language Testing

Suvorov and Heglheimer (2014) explain that computer-assisted language testing (CALT) makes use of technology to evaluate the examinee's second language performance. Additionally, the authors provide a framework of current attributes that describe CALT tests. Five out of the nine attributes provide categories representing characteristics and concepts that have become part of ELT understanding of the field of computer-assisted testing.

1. The attribute of *directionality* describes how tests can be linear, adaptive, or semi-adaptive. Linear tests present the same items in the same order to the examinees. Computer adaptive tests estimate the examinee's competence by offering test items based on the previous response. Each time the test taker answers a question, an algorithm estimates the examinee's competence. Semi-adaptive tests are adaptive at the level of testlets (Hendrickson, 2007). which are groups of items related to a particular content area and are analyzed as a unit.
2. The second attribute, *delivery format,* refers to how computer-assisted language tests are administered. First, we have computer-based tests which make use of software applications that are installed in a computer. Second, web-based tests make use of an online format to deliver the test and evaluate the examinee.
3. The integration of multimedia, such as audio, videos, images, animation and graphics in computer-assisted tests, is pointed out as *media density,* the third attribute. It can be divided into single media in the case of having an audio-only listening test, or multimedia in the case of a listening test that also includes video.
4. The attribute of *target skill* indicates that a test can be developed to evaluate a single language skill (listening, speaking, reading or writing) or it can integrate skills in order to test them, for example, listening and speaking, or reading and writing.
5. The last attribute in CALT mentions *scoring mechanisms* to evaluate the examinee's performance. Tests can be scored by human raters and/or computers, by matching the exact answers or analyzing the test taker's response using natural language processing (NLP) technology. (Suvorov & Heglheimer, 2014, pp.2-4)

Another criterion to describe the developments of CALT is in terms of the contributions that technology has made in test design. Chapelle (2008) discusses three major contributions. The first is concerned with the development and use of computer-adaptive tests. Using Item Response Theory, computer adaptive tests allow the tailoring of items to the test-takers' response, which results in a more accurate and individualized test.

Additional advantages include ease of administration, immediate test results, tracking of students' performance, improved test security and the inclusion of multimedia (Chalhoub-Deville, 2001; Dunkel, 1999). Dunkel (1999) points out that "CAT was developed to eliminate the time-consuming and inefficient (and traditional) tests that present easy questions to high-ability persons and excessively difficult questions to low-ability testees" (p.79). In other words, a CAT narrows the distance between item difficulty and the test-takers' ability. Early developments of CAT instruments for placement purposes have been reported by Chalhoub-Deville (2001) for the French, German and Spanish programs at Brigham Young University. An additional example from the University of Cambridge Local Examination Syndicate is Business Language Testing Service (BULATS) aimed at the corporate sector which currently offers BULATS Online Reading and Listening Test, BULATS Online Speaking Test, and BULATS Online Writing Test (Suvorov & Hegelheimer, 2014). Experiences using CATs have prompted discussion and research on "the way language is measured, the need for independent items and their selection of an adaptive algorithm" (Chapelle, 2008, p.125).

The incorporation of multimedia in language testing is the second improvement discussed by Chapelle (2008). In test development for listening skills, it is now possible for the test developer to focus on specific micro skills to suit a particular need. In addition, the use of images aids in the contextualization of language and in some instances the test-taker can control the speed and even request the repetition of the listening text. These improvements represent a major advancement over paper and pencil tests. Testing listening in a CAT is considered to be more authentic due to the high correspondence between the characteristics of a language test task and target language use task. In addition to this, the testing conditions involve using headphones which leave out environmental noise and allow the student to focus on the task. Bachman (2000) discusses the potential that advances in multimedia and web technology have in the design and development of more authentic and interactive tests.

The inclusion of natural language processing technologies in order to evaluate the test-takers' production of language is the last contribution discussed by Chapelle (2008). Brown (1992) envisaged the use of multimedia advances in language testing, such as the relationship between the test-taker and the computer using voice sensitive and hand-writing recognition devices. One example of NLP technology is *Criterion*, an automated system developed by Educational Testing Service which rates extended written responses. Chapelle and Douglas (2006) mention that "Criterion employs NLP techniques to parse textual input, assigning grammatical labels to items, and looking for markers of discourse structure and content vocabulary items" (p.36). A more recent example, developed by Pearson Education Inc. is the Varsant™ English Test, formerly known as *Phone Pass*. This test includes a speech recognition system which encloses an algorithm derived from a native speaker corpus of spoken English, not only from American or English varieties but also from native speakers from different parts of the world. The test compares the examinees' performance to a template of elicited language (Chapelle & Douglas, 2006; Suvorov & Hegelheimer, 2014).

Although existing research has addressed the use of commercial computer-based tests for placement purposes (Ling, et al., 2014), some researchers and institutions have developed CBTs and CATs for their own specific needs in a foreign language context. One example is the development of the New English Placement Test Online (NEPTON) used in a higher education institution in Cyprus (Papadima-Sophocleous, 2008; Suvorov & Hegelheimer, 2014) where the level of item difficulty was designed for each of their six English language competence levels. A further example is the Computerized Assessment System for English Communication (CASEC) developed by the Japan Institute for Educational Measurement to evaluate proficiency of English as a Foreign Language using an adaptive format (Nogami & Hayashi, 2010). None of the studies indicate methods nor procedures for aligning their tests. Although better authoring tools are available for test design, our research project aimed at aligning a commercially available computer-based test to meet the placement needs of our program.

An important issue to consider before test administration is the meaning of the scores in particular contexts. The user of a published test faces the risk of not knowing how to interpret the scores produced by the test. The situation may be that the recommended score intervals could not be aligned to the proficiency levels or course levels of a particular English program, and as a consequence, placement decisions could be misleading. Our hypothesis was that score intervals that corresponded to each level of our program could be determined for the use of the Oxford Online Placement Test, an online placement test published by Oxford University Press.

Method

The OOPT test was administered to 555 out of 5,879 students registered in the General Courses of English in 2013. At least two groups were randomly selected for each course level. The first and largest sample of 380 tests was taken at the beginning of the semester in January, and the second, 175 tests in May at the end of the semester. English Level 7 and the advanced courses were not considered. Figure 1 describes the sample.

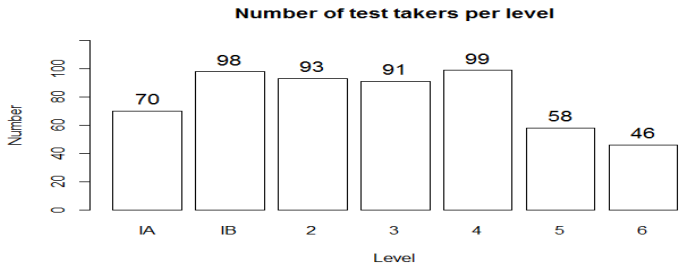

Figure 1. Number of students in the sample per level where 1A and 1B are introductory levels.

Criterion Selection

As can be observed in Tables 2, 3 and 4, the sample shows great variability at each target course which is defined as the course where students were expected to be placed. Some students scored above the target course, others scored at the target course and

some others scored below the target course. The expectation was for students to be placed in the target course represented in the table in bold numbers.

There were no level scores that could be used to represent English competence course level. With the data collected, this process of aligning the test was repeated twice which correspond to Lenext 1 and Lenext 2, and a third time with the intervals that Oxford University Press provided (see Tables 2, 3 and 4).

		Placement according to Lenext 1 scale								
		IA	IB	2	3	4	5	6	7	Adv
Last course accredited	IA	39	**16**	7	4					
	IB	25	12	**29**	18	8				
	2		13	13	**18**	17	10			
	3			2	15	**9**	16	7		
	4				9	11	**7**	10	16	
	5					5	5	**6**	4	37
	6						1	4	**5**	30

Table 2. Lenext 1 sample sizes for each combination of last course accredited and placed course.

		Placement according to Lenext 2 scale								
		IA	IB	2	3	4	5	6	7	Adv
Last course accredited	IA	29	**18**	13	5					
	IB	19	12	**25**	22	12				
	2		6	21	**13**	20	13			
	3			8	10	**14**	16	8		
	4				11	12	**8**	11	16	
	5					5	5	**6**	5	36
	6						2	6	**4**	29

Table 3. Lenext 2 sample sizes for each combination of last course accredited and placed course.

		Placement according to OUP scale								
		IA	IB	2	3	4	5	6	7	Adv
Last course accredited	IA	39	**16**	8	3					
	IB	25	12	**30**	17	8				
	2		13	13	**18**	12	11			
	3			4	13	**9**	18	7		
	4				9	11	**8**	10	16	
	5					5	6	**5**	5	36
	6						2	5	**4**	29

Table 4. OUP sample sizes for each combination of last course accredited and placed course.

As can be noticed in the tables, data beyond two courses from the target course were removed and therefore the following five situations resulted from censoring the data:

(1) *The target course was* represented by the number zero,

(2) *one level above the target course* was represented by the number one,

(3) *two levels above the target course* was represented by the number two,

(4) one level below the target course was represented by the number minus one, and

(5) *two levels below the target course* was represented by the number minus two (Tables 2, 3 and 4).

As a consequence, the sample has been simplified to frequencies of these digits for each target course. In other words, instead of the scores, what is used is the number of courses the student is placed below, on, or above the target course. For each target course, what is counted is the number of students that fall in each situation. For instance, if looking at Lenext 2 (see Table 3 in *level two last course accredited*) six students were two courses below, twenty-one students were one course below, thirteen students were on the target course, twenty were one course above and thirteen were two courses above (Table 5).

Number of students	6	21	13	20	13
Placement situation (weight)	-2	-1	0	+1	+2

Table 5. Lenext 2 and level two last course accredited.

It is expected that the placement situation zero to be a dominant frequency, but as can be seen in Table 5, only thirteen students out of seventy-three were placed on the target course. Many combinations of frequency could have been obtained. Therefore, a simple average of the original scores was not appropriate, and the weighted average (***wa***), which uses the weights -2,-1, 0,+1, and +2 for each situation, was selected instead as a decision criterion. The ***wa*** is the sum of the number of students in each situation times the value of the situation divided by the total number of students involved.

$$\text{Weighted average} = \frac{\sum_{-2}^{+2}(number\ of\ students\ in\ the\ situation)(value\ of\ the\ situation)}{total\ students\ involved}$$

Consider the same case, *level two last course accredited in Lenext 2* (Table 3). The calculation was performed in the following way:

$$wa = \frac{6(-2) + 21(-1) + 13(0) + 20(1) + 13(2)}{73} = 0.17$$

This ***wa*** could be any number from -2 (the case in which all the students were two courses below) and +2 (the case in which all the students were placed two courses above). Let us suppose that most of the students, for example 47, were on the target course, which was a very good case, the ***wa*** would be very near to zero.

$$wa = \frac{7(-2) + 6(-1) + 47(0) + 10(1) + 3(2)}{73} = -0.05$$

Also, the ***wa*** would be near to zero if there was some degree of equilibrium. For instance, when approximately the same number of students was two courses below and two courses above, and approximately the same number of students was one course below and one course above.

$$wa = \frac{4(-2) + 10(-1) + 47(0) + 10(1) + 4(2)}{73} = 0$$

In a similar situation, there could be equilibrium in a negative situation such as the following; however, we did not have these extreme cases (Tables 2, 3, and 4):

$$wa = \frac{30(-2) + 6(-1) + 1(0) + 6(1) + 30(2)}{73} = 0$$

If the ***wa*** was zero or "*close*" to zero, the corresponding course interval was accepted. If the ***wa*** was "*distant*" from zero, then the corresponding course interval was not accepted. In that case there were two possibilities; one was when the ***wa*** was placed to the left of zero in which case the interval was labeled as *demanding*. The second possibility was when the ***wa*** was placed to the right of zero, in which case, the interval was labeled as *undemanding.* As a consequence of this, *a set of intervals where **wa** was nearer to zero was considered the best option.* The meaning of close to zero and distant to zero is explained in the next section.

Bootstrap Procedure

A bootstrap is often referred to as *computing-intensive statistics* which "makes use of extensive repeated calculations to explore a sampling distribution of a parameter estimator" (Venables & Ripley, 2002). The purpose of using the bootstrap method is to build a confidence interval (CI) for the weighted average in order to determine if the ***wa*** is closer to zero or distant from zero (p. 133). If the CI contains the number zero, it is said that the ***wa*** is zero. If the CI does not contain the number zero, it is said that the ***wa*** is not zero. A bootstrap procedure consists of resampling from the original dataset, in which the elements can be repeated in order to evaluate the variability of estimation. Hence the bootstrap is used to evaluate the *sd* of the weighted average to build the CI. In our case, what was needed was the distribution of the weighted average in order to obtain the percentiles 5 and 95. These percentiles were used for building a 90% confidence interval, which was used to make inferences about the ***wa***. With the purpose of accomplishing this objective, 1,000 samples were obtained for each course and their weighted average for each of them. A histogram was built and the percentiles 5 and 95 were calculated. They were considered as an approximation of the real CI for the true weighted average of the course. If the CI for a course contained the number zero, it meant that based on our sample, the ***wa*** was zero with a CI of 90%. If the CI did not contain the number zero and was placed to the left, it meant that the ***wa*** was less than zero with a CI of 90% and the interval was labeled as *demanding* because the ***wa*** was smaller than zero. Furthermore, if the CI did not contain the number zero and was placed to the right, it meant that the ***wa*** was greater than zero with a CI of 90% and it was labeled as *undemanding* because the ***wa*** was greater than zero. This process will be described in the next section.

Results

The objective of this project was to develop a method for identifying the score intervals of a published placement test in alignment with the outcome of a General English Program. The proposal involved two stages. The first stage was to analyze the sample in

order to identify possible sets of intervals for each course level. The second stage consisted of applying statistical tools, such as weighted average and bootstrapping with the purpose of selecting the best set of intervals for the course levels.

Three sets of intervals were proposed in the first stage (Table 1). Two sets were proposed by the authors and a third set of intervals was recommended by Oxford University Press based on the textbook currently used in the English program. Since we have three sets of intervals, we also have three different sample sizes; for Lenext 1 it was 428 (Table 2), for Lenext 2 it was 440 (Table 3), for OUP it was 427 (Table 4). The differences in the number of students placed in each course using each set of intervals could make an important difference in a large university population.

Table 6 shows the results of applying the ***wa*** criterion to the proposed set of intervals. The Appendix describes the calculations of the ***wa*** using as an example the Lenext 2 case. This is part of applying statistical tools in order to select the best set of intervals for the current program.

We labeled a course interval as *demanding* when the CI was at the left of the zero, *adequate* when it contained the zero point, and *undemanding* when the CI was at the right of the zero. For the IA course in the Lenext 1 scale, the ***wa*** = -0.36, meaning an average setback of 0.36 for the student using the reference of the target course. In other words, the scores were low and the interval was labeled *demanding*. If this proposed interval was to be used, it would be likely that the student be placed in a course where his or her abilities are below the expectation. For the Level 2 course of the OUP scale, the ***wa*** =0.0 which meant neither a setback or an advance for the student, in other words the interval is *adequate*. This interval was aligned to what our system was producing. In the Lenext 2 scale the ***wa*** = 1.09 for the level 5 course, which meant an advance of more than one course for the student, in other words the interval was labeled *undemanding*. This interval was not aligned to what our system was producing.

Scales	Course levels in the General English Program							Weighted average	
	IA	IB	2	3	4	5	6	*wa* 1A-4	*wa* 1A-6
wa (Lenext 1)	-0.36	-0.30	-0.02	0.22	0.24	1.10	0.60	-0.09	0.13
Sd	0.11	0.14	0.15	0.16	0.20	0.18	0.12		
Cl	-0.54,-0.18	-0.53,-0.08	-0.28,0.22	-0.02,0.51	-0.11,0.56	0.80,1.4	0.37,0.8		
IC	D	D	A	A	A	U	U		
wa(Lenext 2)	-0.09	-0.04	0.17	0.10	0.15	1.09	0.46	0.05	0.22
Sd	0.11	0.13	0.15	0.17	0.20	0.18	0.14		
CI	-0.27,0.09	-0.26,0.17	-0.08,0.41	-0.16,0.37	-0.17,0.46	0.79,1.38	0.22,0.68		
IC	A	A	A	A	A	U	U		
wa (OUP)	-0.38	-0.31	0.00	0.21	0.24	1.07	0.50	-0.08	0.12
Sd	0.10	0.13	0.15	0.16	0.20	0.18	0.14		
CI	-0.54,-0.21	-0.52,-0.08	-0.26,0.23	-0.04,0.49	0.07,0.59	0.77,1.35	0.25,0.72		
IC	D	D	A	A	U	U	U		

Note. Wa 1A-4 = weighted average for levels 1A to 4; wa 1A-6 = weighted average for levels 1A to 6; Sd = standard deviation; CI = confidence interval; D = demanding; IC = interval classification; A = adequate; and U = undemanding. .

Table 6. Summary of results.

Level of Agreement

There was an exact match between the recommended course and the target course in the following percentages for each scale: 21.02% (Lenext 1), 20.0% (Lenext 2) and 21.07% (OUP) which means that the three proposed sets of intervals behaved very similarly. Approximately one out of five is a perfect match, what is an image of the great variability of the data. There are two types of disagreement. The first is when the course in which the student is placed is a more advanced course than the target course. This case may not represent a considerable problem if students have the option of choosing a more advanced course than the target course. The percentages for this case were 42.99% (Lenext1), 46.81% (Lenext2) and 42.15% (OUP). The remaining type of disagreement was when the course in which the student was placed was more elementary than the target course. This case can become a problem because the student might not be able to succeed in the target course. The percentages for this type of disagreement were 35.98% (Lenext1), 33.18% (Lenext2) and 36.76% (OUP). In this case, the best of all the sets of intervals was Lenext 2, because it had the smallest percentage of students placed in an unfavorable situation.

Weighted Average Analysis Results

An English language competence corresponding to the course level 4 is mandatory for undergraduate students of the University of Sonora. Taking into account this requirement, the question is which set of intervals is the best for the courses from IA to 4 using the ***wa*** criterion (Table 6, column 9). Lenext 1 had a ***wa = -0.09*** and OUP had a ***wa = -0.08*** which means setbacks; however, Lenext 2 had a ***wa = 0.05*** which was nearer to zero and means a small advantage using as reference the target course. Therefore, using the ***wa*** criterion, the best set of intervals for the courses from IA to 4, was Lenext 2. The same conclusion can be reached by referring to the labels *demanding, adequate* and *undemanding* in the interval classification (IC) in Table 6. Only the Lenext 2 scale was reported as *adequate* for each course.

Another interesting question to consider is the **wa** for all the courses. In this case, the OUP scale had a ***wa* = 0.12**, the Lenext 1 scale had a ***wa* =** 0.13 and the Lenext 2 scale had a ***wa*** = 0.22 (Table 6, column 10).The ***wa*** in the OUP scale was closer to zero; therefore, when all the courses were considered, the OUP scale was the most appropriate option.

Discussion

This research shows that a set of score intervals for placement that correspond to each level of the General English Program at the University of Sonora can be developed using the weighted average criterion. The result provides an opportunity to use published placement tests as an alternative to paper and pencil tests currently used for placement purposes. The method used in this project allowed the alignment of a commercial placement test to calibrate a commercial test rather than to design and validate one since it is a major challenge that involves the expertise of a multidisciplinary team, especially for a CAT. This is important because by using a web-delivered exam, faculty

members gain valuable time that could be used for other tasks such as oral assessment (Bernhardt, Rivera & Kamil, 2008).

There are three important differences between the scales Lenext 2 and OUP. The first two differences are concerned with Levels IA and IB (Table 6). While the intervals 1A and 1B proposed by Lenext 2 were classified by the criterion as *adequate*, the intervals 1A and 1B in the scale OUP were classified as *demanding*. The third difference can be observed in the scale OUP for level 4 which classified the course level as undemanding; nevertheless, the scale Lenext 2 classified it as adequate. Lenext 2 provided an *adequate* classification for all the levels from IA to 4; therefore, it was selected as the most appropriate scale for the current context. However, if all the intervals are considered, the OUP is the most appropriate scale with a little advantage over Lenext1.

The three sets of intervals did not classify levels 5 and 6 appropriately (Table 6). Concerning level 5, 63.1% to 64.9% of students were placed in advanced level and 70.7% to 75% of level 6 students were placed in advance level as well. The variation depends on the set of intervals used. In both cases, the percentages were too high (Tables 2, 3 and 4). This was the reason why intervals in the three scales were labeled as *undemanding* for both levels. We hypothesized that there could be two populations mixed in the sample. One student population that corresponded to the natural development of the English competence in the courses of the program and another that already had a higher level of competence and was placed in those courses when entering the program.

There were several limitations in this research project. Student performance was under the influence of many factors and as a consequence it was not constant. Then the score observed was an approximation of the real performance (Pollitt, 2014). As variability reduces, the approximation is better. Another limitation was that at this point there were no samples of individual variation performance at different levels over the range of interest.

Conclusion

The objective of this project found that score intervals corresponding to each level of the English language program could be determined by applying weighted average criterion to the scores produced by a commercial test. The method presented in this research was developed by the authors with the aim of accomplishing the previous objective. Three sets of intervals that can be used for placement in the current English program in the Foreign Language Department were discussed in the results. The analysis considered two cases. The first corresponds to the courses from IA to 6 and the second includes the courses from 1A to 4. In the first case, the OUP scale is the most appropriate with a little advantage over Lenext1. The second case corresponds to our particular context. Therefore, it is concluded that the set of intervals of the scale Lenext 2 is the most appropriate.

In addition, General English Programs can be diverse in terms of curriculum, number of courses, language teachers, students, and context. Hence, a test calibration process should not depend on these variables. In other words, the proposed method must not

depend on a predetermined set of intervals or scales. The weighted average method proposed in this research, could be useful for programs that have not developed a placement instrument and for any combination of number of courses and scales.

There are more variables that must be considered in the model. Two of them are culture and administrative features. The increasing acceptance of computer-assisted instruments for testing English deserves deeper study. On the other hand, studies about computer process capacity are an important issue that must be addressed before using a computerized test for large and simultaneous administration.

References

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing, 17*(1), 1-42. doi:10.1177/026553220001700101

Bernhardt, B. B., 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), 356-365. doi: 10.1111/j.1944-9720.2004.tb02694.x

Brown, J. D. (1992). Technology and language education in the twenty-first century: media, message, and method. *Language Laboratory*, 29, 1-22.

Brown, J. D. (1997). Computers in language testing: present research and some future directions. *Language Learning and Technology, 1*(1), 44-59. Retrieved from http://llt.msu.edu/vol1num1/brown/default.html

Chalhoub-Deville, M. (2001). Language testing and technology: past and future. *Language Learning and Technology, 5*(2), 95-98.

Chapelle, C. A. & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press. Retrieved from http://llt.msu.edu/vol5num2/deville/default.html

Chapelle, C. A. (2008). Utilizing technology in language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education, 2nd Ed., Vol 7: Language Testing and Assessment,* (pp. 123-134). New York: Springer.

Dunkel, P. A. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology, 2*(2), 77-93. Retrieved from http://llt.msu.edu/vol2num2/article4/index.html

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice. 26*(2), 44-52. doi:10.1111/j.1745-3992.2007.00093.x

Ling, G., Wolf, M. K., Cho, Y., & Wang, Y. (2014). *English-as-a-Second-Language Programs for Matriculated Students in the United States: An Exploratory Survey and Some Issues* (Research Report ETS RR-14-11). Educational Testing Service. doi: 10.1002/ets2.12048

Nogami, Y. & Hayashi, N. (2010). A Japanese Adaptive Test of English as a Foreign Language: Development and Operational Aspects. In W. J. van der Linden & C.A.W. Glas (Eds.) *Elements of adaptive testing, statistics for social and behavioral sciences* (pp.191-211). New York: Springer-Verlag. doi:10.1007/978-0-387-85461-8_10

Papadima-Sophocleous, S. (2008). A hybrid of a CBT- and a CAT-based new English placement test online (NEPTON). *CALICO Journal,* 25(2), 276-304. doi:10.1558/cj.v25i2.276-304

Pearson Education, Inc. (2013). Versant English Test [computer and telephone delivery]. Published instrument. Retrieved from https://www.versanttest.com/products/english.jsp

Pollitt, A. (2014). *The Oxford Online Placement Test: The Meaning of OOPT Scores*. Retrieved from https://www.oxfordenglishtesting.com/defaultmrarticle.aspx?id=3074

Suvorov, R. & Hegelheimer, V. (2014). Computer-assisted language testing. In A. J. Kunnan (Ed.). *The companion to language assessment, First Edition* (pp.1-20). New Jersey: John Wiley & Sons, Inc. doi: 10.1002/9781118411360.wbcla083

Universidad de Sonora. Lineamientos Generales para un Modelo Curricular de la Universidad de Sonora. (2003). GACETA Órgano Informativo de la University of Sonora. Hermosillo, Son: Draw Graphic.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S.* (4th Ed.) New York: Springer.

Appendix Working Case Lenext 2: Tables 3 and 6

This section includes the calculations of the ***wa*** for each level of the set of intervals in Lenext 2 *last course accredited*. Two histograms are presented, one showing the frequency of the number of students placed in each level, and another for the CI of the ***wa.*** The last part of the section describes the calculations of the ***wa*** that corresponds to the set of courses 1A to 4 and to the set of courses 1A to 6.

Last course accredited: 1A

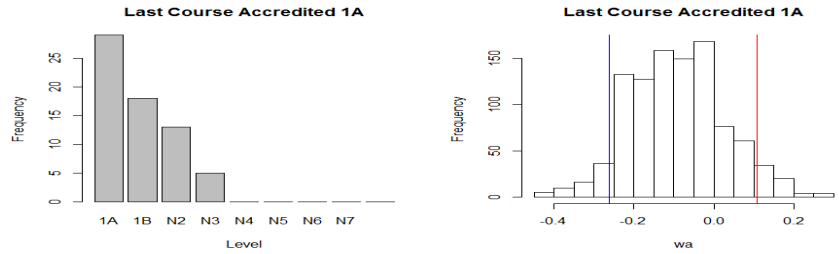

Figure A1. Number of students per level and histogram of **wa,** 1A.

$$wa = \frac{0(-2) + 29(-1) + 18(0) + 13(1) + 5(2)}{65} = -0.09$$

The graph shows a histogram with the weighted average of all 1000 samples. The 90% confidence interval is built with the percentiles 5 and 95, (-0.27, 0.09) which is marked with the vertical lines. Each time that a new 1,000 sample is generated, a new confidence interval will be obtained, but it will be approximately equal.

Last course accredited: 1B

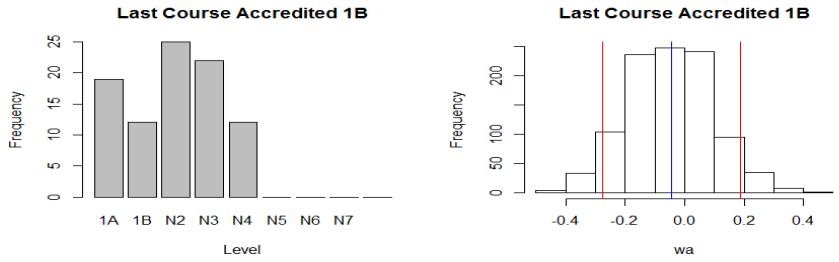

Figure A2. Number of students per level and histogram of **wa**, 1B.

$$wa = \frac{19(-2) + 12(-1) + 25(0) + 22(1) + 12(2)}{90} = -0.04$$

The graph shows a histogram with the weighted average of all 1000 samples. The 90% confidence interval is built with the percentiles 5 and 95, (-0.26, 0.17) which is marked with the vertical lines.

Last course accredited: 2

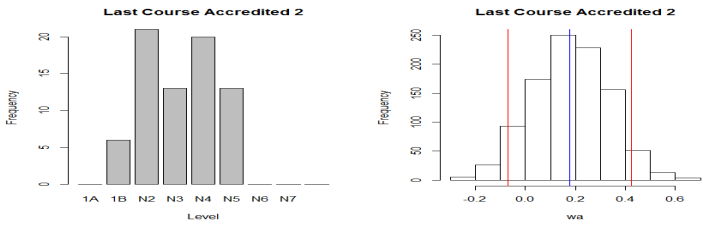

Figure A3. Number of students per level and histogram of **wa,** 2.

$$wa = \frac{6(-2) + 21(-1) + 13(0) + 20(1) + 13(2)}{73} = 0.17$$

The graph shows a histogram with the weighted average of all 1,000 samples. The 90% confidence interval is built with the percentiles 5 and 95, (-0.08, 0.41) which is marked with the vertical lines.

Last course accredited: 3

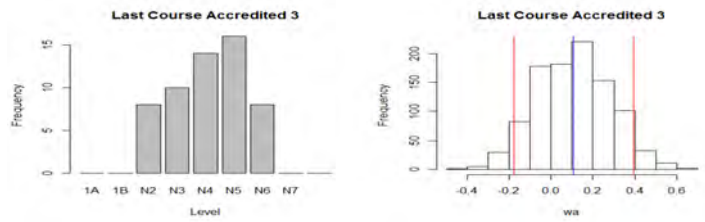

Figure A4. Number of students per level and histogram of **wa,** 3.

$$wa = \frac{8(-2) + 10(-1) + 14(0) + 16(1) + 8(2)}{56} = 0.10$$

The graph shows a histogram with the weighted average of all 1,000 samples. The 90% confidence interval is built with the percentiles 5 and 95, (-0.16, 0.37) which is marked with the vertical lines.

Last course accredited: 4

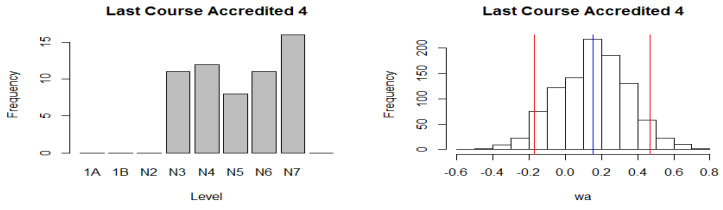

Figure A5. Number of students per level and histogram of **wa**, 4.

$$wa = \frac{11(-2) + 12(-1) + 8(0) + 11(1) + 16(2)}{58} = 0.15$$

The graph shows a histogram with the weighted average of all 1,000 samples. The 90% confidence interval is built with the percentiles 5 and 95, (-0.17, 0.46) which is marked with the vertical lines.

Last course accredited: 5

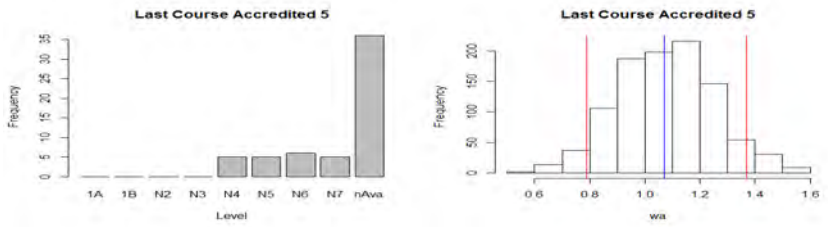

Figure A6. Number of students per level and histogram of **wa**, 5.

$$wa = \frac{5(-2) + 5(-1) + 6(0) + 5(1) + 36(2)}{57} = 1.09$$

The graph shows a histogram with the weighted average of all 1,000 samples. The 90% confidence interval is built with the percentiles 5 and 95, (0.79, 1.38) which is marked with the vertical lines. Note that this CI does not contain the number zero.

Last course accredited: 6

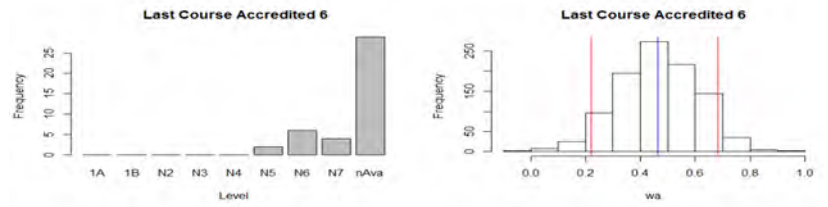

Figure A7. Number of students per level and histogram of **wa,** 6.

$$wa = \frac{2(-2) + 6(-1) + 4(0) + 29(1) + 0(2)}{41} = 0.46$$

The graph shows a histogram with the weighted average of all 1,000 samples. The 90% confidence interval is built with the percentiles 5 and 95, (0.22, 0.68) which is marked with the vertical lines. Note that this CI does not contain the number zero.

Weighted average for levels from 1A to 4 and for 1A to 6

Calculations of the ***wa*** that correspond to the set of courses 1A to 4 and to the set of courses 1A to 6 are presented in this section. The first set corresponds to level 4 which is mandatory for degree attainment for any student enrolled in a program of study at the University of Sonora. The second set corresponds to the course levels of the General English Program in the Foreign Language Department.

Last course accredited	1A	1B	2	3	4	5	6
Number of students	65	90	73	56	58	57	41

Table A1. Lenext 2 and number of students for last course accredited.

$$wa(1A-4) = \frac{65(-0.09) + 90(-0.04) + 73(0.17) + 56(0.10) + 58(0.15)}{342} = 0.05$$

$$wa(1A-6) = \frac{65(-0.09) + 90(-0.04) + 73(0.17) + 56(0.10) + 58(0.15) + 57(1.08) + 41(0.46)}{440} = 0.22$$