Item analysis of a multiple-choice test of incoming EFL students at a public university in Guanajuato State

Análisis de ítems de una prueba de opción múltiple realizada a estudiantes de ILE de nuevo ingreso en una universidad pública del estado de Guanajuato

Kenneth Geoffrey Richter
Universidad de Guanajuato,
Campus Guanajuato, División de
Ciencias Sociales y Humanidades,
Departamento de Lenguas
ken.richter@gmail.com

Ricardo Alejandro
Medel Romero
Universidad de Guanajuato,
Dirección de Apoyo a la
Investigación y al Posgrado (DAIP)
ricardo.medel@ugto.mx

Recepción del artículo: 13 de agosto de 2019
Aceptación: 15 de octubre de 2019
doi: 10.22201/enallt.01852647p.2020.71.900

Abstract

The importance of multiple-choice (MC) exams in Mexico can hardly be overstated. Despite their ubiquity in Mexican EFL classrooms, little research on their quality or impact has been carried out. The current paper reports on an item analysis conducted on a multiple-choice placement exam administered to 50 incoming English language students at a public university in Guanajuato State. All the items contained in this test were subjected to four different types of statistical analyses: item facility (IF), distractor analysis (DA), item discrimination (ID), and response frequency distribution (RF). The results showed that the exam was critically deficient in each of these areas. While the findings of the current analysis cannot be generalized beyond the exam in question, the results highlight the extreme difficulty of creating MC exams and the indispensability of piloting tests and conducting item analyses before applying them.

Keywords: English as a foreign language; assessment; item difficulty; item discrimination

Resumen

La importancia de los exámenes de opción múltiple (MC) en México no debe subestimarse. A pesar de su ubicuidad en las aulas de enseñanza de ILE (inglés como lengua extranjera), se han realizado pocas investigaciones sobre su calidad o impacto. El presente artículo informa sobre un análisis de ítems realizado en un examen de ubicación de opción múltiple administrado a 50 estudiantes de inglés de nuevo ingreso en una escuela pública de nivel superior en el estado de Guanajuato. Todos los ítems de la prueba fueron sometidos a cuatro diferentes análisis estadísticos: índice de dificultad (IF), análisis de distractores (DA), índice de discriminación (ID) y distribución de frecuencia (RF). Los resultados mostraron que el examen fue sumamente deficiente en cada una de estas áreas. Si bien los resultados de este estudio no se pueden generalizar más allá del examen en cuestión, los resultados subrayan la dificultad que implica crear exámenes tipo MC y la necesidad de realizar pruebas piloto y análisis de ítems antes de aplicarlos.

Palabras clave: inglés como lengua extranjera; evaluación; índice de dificultad; índice de discriminación

doi: 10.22201/enallt.01852647p.2020.71.900

1. Introduction

The performance of any given English as a Foreign Language (EFL) student on a language test is affected by two variables: his or her knowledge of English and the assessment instrument employed (Yi'an, 1998). Since information obtained through testing may have significant repercussions for both teachers and students, it is crucial that the assessment instrument provides accurate information about the knowledge of English.

The importance of multiple-choice (MC) exams in Mexico can hardly be overstated. All of the large-scale, standardized assessments of English language ability used in Mexico, for instance, use the MC format. These include the *Exámenes Nacionales de Ingreso* (EXANI), the English First Standard English Test, the Test of English as a Foreign Language (TOEFL), and the International English language Testing System (IELTS) (British Council, 2015; Payant & Barrón, 2014).

Multiple-choice exams are virtually omnipresent in the Mexican EFL classroom. In a summary report on educational evaluation and assessment in Mexico, the Organization for Economic Co-operation and Development (OECD, 2012: 2) characterizes Mexico's reliance on multiple-choice tests as "excessive". The reasons for the popularity of MC tests are manifold. Here, four of the most important ones are considered: (1) The Mexican education system is marked by its pedagogical traditionalism. While the Secretaría de Educación Pública (SEP) exhorts teachers to adopt modern, formative assessment practices (SEP, 2017), traditional instructional methods (i.e., those approaches that are synthetic, transmissional, teacher-centered, forms-focused, and examinations-driven) are pervasive (Borjian, 2015; Despagne, 2010; Pamplón & Ramírez, 2013). As Despagne (2010: 62) notes "Despite educational reform, English teaching is known to be deficient […] teaching is based on repetition drills, rote learning and memorization". These traditional pedagogical approaches encourage and are encouraged by traditional assessment practices, such as the use of gap fills, matching exercises, and multiple-choice exams. (2) The way language instructors are contracted negatively impacts teaching. The majority of English teachers in Mexico are not hired as regular staff but rather work as part-time specialists, many of whom are contracted for only a few hours a week. Those teachers are not paid for such customary teaching responsibilities as preparing classes, designing materials, grading homework, and scoring exams (Mexicanos Primero, 2015: 65). These instructors, then, have little motivation to utilize assessment instruments that take time outside of regular classroom hours to create and correct. MC tests are quick to devise and easy to score, making them the ideal exam for overworked teachers. (3) Many classrooms in Mexico — particularly those in public schools — are overcrowded. Of the 28 OECD countries for which there is data, only four countries had larger classes in lower secondary education than Mexico: Korea, Chile, Japan, and Turkey (see OECD). Large class sizes mean that assessments must be easy to both administer and mark – two of the chief hallmarks of multiple-choice exams (Tucker, 2015). (4) Many English teaching positions are filled by instructors with poor language skills and little or no training in the methodology of teaching of foreign languages (Rodríguez, 2014). In a study carried out by Mexicanos Primero (2015: 93), more than half of the English teachers who agreed to be tested were found to have English levels below those that their students were expected to achieve. Many such teachers may show a strong preference for objective tests such as multiple-choice exams over subjective assessments that require language expertise to create and mark.

Multiple-choice tests have been criticized on a number of grounds. The three main areas of criticism are that MC tests are detrimental to learning, that they possess poor construct validity, and that they are technically challenging to construct. The first criticism relates to the fact that when educational institutions choose or are required to use MC tests, instructors may feel pressured to teach to the test; that is, the pedagogic focus narrows to practicing questions that are similar to those included in the exam. In such cases, rote learning may be encouraged (i.e., an emphasis on reproduction of knowledge) at the expense of analysis, synthesis, or evaluation (Hamp-Lyons, 2004). Students may be deprived of the opportunity to engage with new material in meaningful ways: to ask their own questions, to have discussions, to explore new ideas (Fairtest, 2007). Oller (1979: 233) has argued that problems with MC tests are so severe that they are "intrinsically inimical to the interests of instruction".

Multiple-choice tests have also been criticized because of their poor construct validity. Validity here refers to "the weight of evidence that supports the claim that the test measures what it is intended to measure" (Tucker, 2015: 4). MC tests are a measurement of discrete knowledge and, as such, can only reflect what a learner knows *about* language; to paraphrase Tucker, they cannot capture the *doing* of language.

Finally, and most germane to the present paper, it has been argued that the development of multiple-choice tests is a formidable undertaking requiring a level of technical expertise that few teachers can be expected to possess (Oller, 1979). As Sajadi (2006: 200) notes, in educational contexts, instructors compose the largest population of test makers. It is critical, therefore, that instructors be well-trained in test construction. However, there is considerable evidence indicating that this is most likely not the case (Asim, Bassey, Akwegwu & Obi, 2005, as cited in Asim, Ekuri & Eni, 2013; Brooks & Normore, 2018; Davies, 2009; Despagne, 2010; González, 2017; INEE, 2015; Santibañez, Vernez & Razquin, 2005; SEP, 2010; Weigle, 2007). As Brooks and Normore (2018) remark, for many teachers, "their only exposure to the concepts and practices of educational assessment might have been a few sessions in their educational psychology classes or, perhaps, a unit in a methods class" (Chapter 6, "Assessment Literacy," para. 2).

Given the importance of multiple-choice tests in Mexican EFL classrooms, if they are to be taken seriously as evidence of student learning, it is critical that they are piloted and analyzed before being administered. Oller (1979: 245) convincingly argues that item analysis is an imperative requisite in the preparation of good multiple-choice tests and that every educator who uses this type of exam should be familiar with how to conduct such a review. Taking Oller's exhortation as its point of departure, the current paper reports on the psychometric soundness of a multiple-choice exam administered to 50 incoming students at a public university in Guanajuato State, in central Mexico.

2. Literature review

In general, vanishingly little attention has been paid to the use of language assessments at the levels of individual classrooms and schools in Mexico. Only a handful of articles report on research in this area (see, for instance, González, 2017; Schissel, Leung, López-Gopar & Davis, 2018). This can be measured against the wealth of studies on classroom language assessment in other countries (see, for example, Afflerbach, Armengol, Brooke, Carper, Cronin, Denman, Irwin, McGunnigle, Pardini & Kurtz, 1995; Alderson & Clapham, 1995; Milnes & Cheng, 2008; Cheng, Rogers & Hu, 2004; Cheng, Rogers & Wang, 2008; Cheng & Wang, 2007; Cheng & Warren, 2005; Colby-Kelly & Turner, 2007; Inbar-Lourie & Donitsa-Schmidt, 2009; Jia, Eslami & Burlbaw, 2006; Kasper, 2002; Ketabi & Ketabi, 2014; Lee, 2007; Li, Link, Ma, Yang & Hegelheimer, 2014; Llosa, 2008; Poehner & Lantolf, 2005; Puhl, 1997; Steadman, 1998; Topping, 1998; Waring, 2008).

Moreover, although multiple-choice tests are omnipresent in Mexican classrooms, as with language assessments in general, remarkably little research on their quality or impact has been carried out (García & Castañeda, 2006). Again, the paucity of studies in Mexico can be contrasted with the literature devoted to the use of multiple-choice tests in ESL assessment in other contexts. The research sites for these studies are located in countries such as the United States, Spain, Chile, Peru, Korea, Japan, Thailand, China, Nigeria, and Israel (Argüelles Álvarez, 2013; Asim *et al.*, 2013; Bensoussan, 1984; Currie & Chiramanee, 2010; Dobson, 1975;

Gyllstad, Vilkaitė & Schmitt, 2015; Hoshino, 2013; In'nami & Koizumi, 2009; Katalayi & Sivasubramaniam, 2013; Pike, 1979; Thanyapa & Currie, 2014; Zheng, Cheng & Klinger, 2007). The bulk of studies focusing on the MC format is characterized by relatively large sample sizes (more than 100 participants). A number of studies, however, focus on the level of classroom and educational institution and are marked by their small set of participants (Katalayi, 2018; Ko, 2010; Nevo, 1989; Rupp, Ferne & Choi, 2006; Teemant, 2010; Yi'an, 1998).

3. Method

The sample considered for the present research was a 50-item multiple-choice diagnostic English test administered by the university to 50 incoming students selected at random. At the time of the exam, the students' average age was 19 years. The exam was designed to measure three different language skills: listening comprehension, reading comprehension, and grammar. The subtests for these skills varied in terms of the number of items in each: the listening comprehension section consisted of 10 questions, the grammar section consisted of 20 questions, and the reading comprehension section consisted of 20 questions. Of the fifty items in the exam, ten items (the listening section) had three distractors; all the rest had four distractors. In total, 190 distractors were analyzed. All the items in this test were subjected to four different types of statistical analyses: item facility, distractor analysis, item discrimination, and response frequency distribution.

3.1. *Item facility*

According to Lado (1961: 342), "Item analysis is the study of validity, reliability, and difficulty of test items taken individually as if they were separate tests". Probably, the most important type of item analysis is item facility (also known as index of difficulty, facility value, or IF). The general purpose of IF is to highlight the degree of difficulty of each of the items in a test (Alderson, Clapham & Wall, 1995; Bailey, 1998; Brown, 2005; Fulcher & Davidson, 2007; Heaton, 1988; Hughes, 1989; Madsen, 1983).

IF is a number that ranges from 0.00 to 1.00. This number represents the percentage of people who answered a given item correctly. For the current study, the 50 test items in the university's diagnostic exam were statistically analyzed. To measure the IF of the items in this exam, it was necessary to sum up the total number of students who correctly answered the questions and divide that number by the total number of students who took the test, as shown in the next formula:

$$IF= \frac{\textit{Total no. of students who got the item right}}{\textit{Total no. of students taking the test}}$$

For each exam item, the number of students who answered the item correctly was divided by the total of 50 test-takers. This allowed detecting those items with a low facility index (close to 0.00, i.e., very difficult), those with a high facility index (close to 1.00, i.e., very easy), and those in the middle (around 0.50, which is considered ideal). As it is difficult to calculate the IF by hand, a formula in an Excel spreadsheet was utilized to calculate individual IF scores (see Table 1).

Table 1. Sample of responses to test items by 50 participants

Student	I1	I2	I3	I4	I5	I46	I47	I48	I49	I50	Total
St 1	1	1	1	1	1	0	0	1	0	1	32
St 2	1	1	1	1	1	0	0	0	0	1	31
St 3	1	1	1	1	1	0	1	1	0	1	31
St 4	1	1	1	1	1	1	1	0	0	0	30
St 5	1	1	1	1	1	0	0	0	0	0	30
St 46	0	1	1	1	1	0	1	0	1	0	15
St 47	0	0	1	1	0	0	0	0	0	1	15

(continued)

Table 1. Sample of responses to test items by 50 participants

Student	I 1	I 2	I 3	I 4	I 5	I 46	I 47	I 48	I 49	I 50	Total
St 48	1	1	1	1	0	1	0	0	0	0	15
St 49	0	0	1	0	1	0	0	0	0	1	14
St 50	1	1	1	1	1	0	1	0	0	0	14

Note: Some columns and rows have been hidden for a better fit of information.

3.2. *Distractor analysis*

Once the IF scores for the 50 items were calculated, items were subjected to distractor analyses to ascertain the extent to which the distractors in each item functioned correctly (Bailey, 1998; Hughes, 1989; Madsen, 1983). The purpose of distractors in MC exams is to distract poorer test-takers, ensuring that those who possess the knowledge required are more likely to answer an item correctly than those who do not (Bailey, 1998). To find out if the distractors were working properly (i.e., to find out if the distractors were, in fact, distracting), a new MS Excel spreadsheet was constructed. First, the exam items were ordered in Column A (1 to 50). In Row 1, the different options were written out (A, B, C, and D). Subsequently, the options selected by each examinee were recorded individually in the spreadsheet. At the end of this process, it was possible to tally the number of responses each of the distractors received from the whole group of candidates (see Table 2).

Table 2. Distractor analysis sample

Item	A	B	C
1	8	40	1
2	5	43	1
3	1	2	47
4	6	39	5
5	27	7	16

(continued)

Table 2. Distractor analysis sample

Item	A	B	C
6	6	20	**24**
7	0	**38**	12
8	**33**	7	9
9	3	**39**	8
10	**23**	19	7

Note: Greyed cells mark the correct answers.

3.3. *Item discrimination*

The purpose of item facility is to differentiate the easy and difficult items of a test. Item discrimination (ID) analysis helps a test maker to visualize which items are working properly, i.e., discriminating between low-performing and high-performing test-takers. If, for instance, high-scorers and low-scorers both answer a given item correctly, that would indicate a problem with the item. If both high-scorers and low-scorers answer a given test item incorrectly, that would also indicate a problem with the item. Discrimination is important because exam items, as well as entire tests, should be able to partition students according to their differing levels of knowledge. Furthermore, when a test is able to discriminate, there is evidence of test reliability.

Since ID in testing refers to the degree to which test items distinguish between high-scorers and low-scorers, it is a basic principle that every test item needs to follow (Alderson *et al.*, 1995; Bailey, 1998; Brown, 2005; Fulcher & Davidson, 2007; Heaton, 1988; Hughes, 1989; Madsen, 1983). Similar to IF, the discrimination index ranges from +1.00 to -1.00. This range denotes the degree of discrimination that each item possesses: +1.00 signifying perfect discrimination, 0.00 signifying no discrimination at all, and -1.00 signifying perfectly wrong discrimination. Bailey (1998) argues that an acceptable value should be set at 0.25 or 0.35, Madsen (1983) holds that values of 0.15 or above are generally acceptable, Alderson *et al.* (1995) mention that values of 0.40 or above are preferred, and Heaton (1988) states that the acceptable values range from 0.30 and above. Following Heaton's (1988: 182) admonition that "[i]tems showing a discrimination index of below 0.30 are of doubtful use", the accepted ID threshold in the current research was set at 0.30.

Before computing ID, it was first necessary to arrange the scores of the tests in order to identify the high- and low-scorers (see Table 3). Then, Flanagan's method (see Hales, 1972) was utilized to approximate the correlation coefficient: the top 27.5% of the exams ($n = 14$) and the bottom 27.5% were analyzed. The ID value of each item was then computed by subtracting the IF of the low-scoring group from the IF of the high-scoring group for each item, as shown by the following formula:

ID = IF of high-scoring group – IF of low-scoring group

Table 3. Item discrimination sample

Student	I 1	I 2	I 3	I 4	I 5	I 46	I 47	I 48	I 49	I 50	Total
St 1	1	1	1	1	1	0	0	1	0	1	32
St 2	1	1	1	1	1	0	0	0	0	1	31
St 3	1	1	1	1	1	0	1	1	0	1	31
St 4	1	1	1	1	1	1	1	0	0	0	30
St 5	1	1	1	1	1	0	0	0	0	0	30
St 46	0	1	1	1	1	0	1	0	1	0	15
St 47	0	0	1	1	0	0	0	0	0	1	15
St 48	1	1	1	1	0	1	0	0	0	0	15
St 49	0	0	1	0	1	0	0	0	0	1	14
St 50	1	1	1	1	1	0	1	0	0	0	14

(continued)

Table 3. Item discrimination sample

Student	I1	I2	I3	I4	I5	I46	I47	I48	I49	I50	Total
IF upper	1.00	1.00	0.93	0.93	0.71	0.21	0.36	0.29	0.21	0.57	
IF lower	0.64	0.71	1.00	0.57	0.50	0.21	0.29	0.21	0.21	0.36	
ID	0.36	0.29	-0.07	0.36	0.21	0.00	0.07	0.07	0.00	0.21	

Note: Some columns and rows have been hidden for space.

Once an item discrimination index was created, it was necessary to scrutinize the results in order to discover why certain items had low discrimination scores.

3.4. *Response frequency distribution*

The response frequency distribution (RF) analysis was conducted to explain the findings of the ID procedure described above; that is, it was necessary to analyze in more detail why certain items received good or bad ID values. One of the goals of the RF analysis is to ascertain which distractors are more appealing to certain scorers (high or low). It is the aim of a RF analysis to compare the number of responses of each group on certain items to find the level of discrimination (Bailey, 1998). It also shows the frequency with which test items are selected by different groups.

A total of 28 tests and their responses were analyzed statistically. To carry out this analysis, once again, only the top-scoring and the bottom-scoring tests were recorded in Excel. First, the test items were ordered from 1 to 50. Subsequently, a space for low- and high-scorers was assigned to every item. Finally, columns were assigned to the different distractors (A, B, C, and D in the case of the grammar and reading sections, and A, B, and C in the case of the listening section). The number of times each option was selected by the low- and high-scorers was entered. In this way, a comparison could be made between the preferences of the two groups concerning the items, as shown in Table 4.

Table 4. Response frequency distribution sample

Item	High / Low scorers	A	B	C
1	High scorers	0	14	0
	Low scorers	5	9	0
2	High scorers	0	14	0
	Low scorers	3	10	1
3	High scorers	0	1	13
	Low scorers	0	0	14
4	High scorers	1	13	0
	Low scorers	2	9	2
5	High scorers	10	0	4
	Low scorers	7	3	4
6	High scorers	1	4	9
	Low scorers	2	6	6
7	High scorers	0	13	1
	Low scorers	0	10	4
8	High scorers	12	1	1
	Low scorers	9	2	4
9	High scorers	0	14	0
	Low scorers	1	8	4
10	High scorers	9	4	1
	Low scorers	3	6	4

Note: Greyed cells refer to the correct answer.

Before explaining the analysis, it should be mentioned that the *no answer* responses were not considered as incorrect answers in the distractor analysis and response frequency distribution, nor were they recorded in these statistics; consequently, the total number of responses for certain items in these techniques add up to less than 50, i.e., the total number of students answering the test.

4. Results and discussion

This section presents the results of the item analyses conducted on the 50 items in the test. It highlights the findings and offers recommendations concerning the improvement of the test.

4.1. *Item facility*

The IF range was set from 0.30 to 0.80. That is, items falling within this range were considered acceptable for inclusion in the test; items falling below 0.30 or above 0.80 were considered deficient and in need of modification or replacement.

As mentioned in the previous section, IF helps to detect those items that do not contribute to the reliability of a test because of their very high or very low levels of difficulty. A detailed description of the items that emerged as weak, average, or strong based on the results of the IF analysis is included below.

4.1.1. Listening subtest: items 1 to 10

According to the IF criteria, 30% of the items in the listening subtest fell outside the accepted range (0.30 to 0.80, see Table 5 below). Items 1, 2, and 3 all had an IF above 0.80 and would need to be improved or replaced in any subsequent version of the exam. Moreover, although items 4, 7, and 9 fell within the acceptable IF range, these were somewhat suspicious since they fell close to the *very easy* end (0.76) of the acceptable range.

Table 5. Item facility of the listening subtest (10 items)

Item	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
Test-takers	41	43	47	38	28	25	38	33	38	23
IF	0.82	0.86	0.94	0.76	0.56	0.50	0.76	0.66	0.76	0.46

Note: "Test-takers" refers to the total of test-takers answering the item correctly out of 50.

4.1.2. Grammar subtest: items 11 to 30

Thirty percent of the grammar items fell below the 0.30 criterion, i.e., they were too difficult for the level of the test (see Table 6). Items 12, 13, 22, 25, 28, and 29 were the most difficult items: most test-takers answered them incorrectly (IF values ranged from 0.12 to 0.24). In any future iteration of the exam, it would be advisable to rewrite or remove these items.

Table 6. Item facility of the grammar subtest (20 items)

Item	I 11	I 12	I 13	I 14	I 15	I 16	I 17	I 18	I 19	I 20	I 21	I 22	I 23	I 24	I 25	I 26	I 27	I 28	I 29	I 30
Test-takers	24	11	11	29	25	24	15	18	17	17	35	12	18	29	10	27	17	6	10	20
IF	0.48	0.22	0.22	0.58	0.50	0.48	0.30	0.36	0.34	0.34	0.70	0.24	0.36	0.58	0.20	0.54	0.34	0.12	0.20	0.40

Note: "Test-takers" refers to the total number of test-takers answering the item correctly out of 50.

Moreover, items 17, 19, 20, 23, and 27 should should be modified for falling directly on or very close to the limit, set at 0.30. These items should be reviewed carefully to ensure that they are not excessively difficult for the desired level of the exam (see the distractor analysis in the next section for more information).

4.1.3. Reading subtest: items 31 to 50

In this final section of the test, a total of ten items (50%) fell outside of the acceptable IF range: they were too difficult for the desired level of the test (Table 7).

Items 34, 35, 36, 38, 39, 40, 43, 46, 48, and 49 should be replaced or modified because they were all outside the acceptable range of difficulty. Moreover, items 31, 44, and 47 are suspicious, falling very close to the difficulty limit and deserve further investigation. Items 32, 37, 41, 45, and 50 can be considered the strongest items in this section, all of them falling near the ideal IF value (0.50). In summary, the reading section showed many problems in terms of item facility.

Table 7. Item facility of the reading subtest (20 items)

Item	I 31	I 32	I 33	I 34	I 35	I 36	I 37	I 38	I 39	I 40	I 41	I 42	I 43	I 44	I 45	I 46	I 47	I 48	I 49	I 50
Test-takers	16	29	18	12	12	13	20	4	12	13	20	38	11	16	27	10	16	14	11	26
IF	0.32	0.58	0.36	0.24	0.24	0.26	0.40	0.08	0.24	0.26	0.40	0.76	0.22	0.32	0.54	0.20	0.32	0.28	0.22	0.52

Note: "Test-takers" refers to the total of test-takers answering the item correctly out of 50.

4.2. *Distractor analysis*

One of the main causes of low or high IF scores is the poor quality of distractors. Madsen (1983: 184) states that "[n]o set percentage of responses has been agreed upon, but examiners usually feel uneasy about a distractor that isn't chosen by at least one or two examinees in a sample of 20 to 30 test papers". For this analysis, distractors that no or very few test-takers (i.e., five or fewer) chose are highlighted.

4.2.1. Listening subtest: items 1 to 10

In this first part of the test, a total of six items were found to have problems with one or two of their distractors. Distractors that few participants select contribute to the low reliability of a test (Hughes, 1989). Looking at Table 8, items 1, 2, 3, 4, 7, and 9 suffered from poor distractors.

Distractor C in items 1, 2, and 4 did not distract enough test-takers to be considered acceptable. Also, distractor A in items

Table 8. Distractor analysis: Listening subtest (10 items)

Item	A	B	C
1	8	**40**	1
2	5	**43**	1
3	1	2	**47**
4	6	**39**	5
5	**27**	7	16
6	6	20	**24**
7	0	**38**	12
8	**33**	7	9
9	3	**39**	8
10	**23**	19	7

Note: Greyed cells mark the correct answers.

3, 7, and 9 served no purpose, since no or very few participants chose it. Finally, distractor B in item 3 reflected a low level of distraction as it was chosen by only two students; this means that the distractor would need to be replaced in future versions of the test. In sum, the listening section can be considered to be rather weak, given that 60% of the items were marked by deficient distractors.

4.2.2. Grammar subtest: items 11 to 30

In this second section, 16 items were found to have problems with one, two, or even three of their distractors. Looking at Table 9, distractor A in item 23 can be considered to be of poor quality: only one examinee selected it. In items 14, 15, 20, and 30, distractor B was weak. In items 11, 12, 16, 18, 19, 21, and 29, only between two to five test-takers selected distractor C. Finally, in items 11, 15, 18, 21, 24, 25, 26, 28, and 29, distractor D was chosen by very few test-takers.

Table 9. Distractor analysis: Grammar subtest (20 items)

Item	A	B	C	D
11	18	**24**	3	5
12	9	23	5	**11**
13	11	15	**11**	12
14	**29**	5	8	6
15	19	2	**24**	4
16	15	**23**	4	6
17	**15**	19	8	6
18	**18**	23	3	5
19	12	17	4	**12**
20	6	4	**17**	23
21	6	**34**	4	4
22	**13**	12	9	15
23	1	8	21	**19**
24	**28**	9	9	2
25	29	8	**9**	3
26	10	**26**	8	5
27	13	11	6	**15**
28	24	14	**5**	2
29	**10**	31	2	2
30	**20**	5	15	4

Note: Greyed cells mark the correct answers.

4.2.3. Reading subtest: items 31 to 50

Table 10 shows that a total of 15 out of 20 items (i.e., 75%) on the reading section had problems with one or two of their distractors. Distractor A appeared to be very weak in items 36 and 37; in item 36, only four participants chose A; in item 37, only two participants chose A. Distractor B in items 31, 32, and 33 was not chosen by a sufficient number of test-takers to be considered sound. Similarly, distractor C in items 32, 36, 40, 41, and 42 was chosen by less than three students. In items 35, 38, 42, 43, 45, 46, and 50, distractor D likewise did not attract many test-takers.

Table 10. Distractor analysis: Reading subtest (20 items)

Item	A	B	C	D
31	9	4	**16**	19
32	15	3	2	**28**
33	**18**	2	7	21
34	10	7	19	**12**
35	**10**	27	9	2
36	4	**12**	2	31
37	2	17	8	**19**
38	10	33	**4**	2
39	**12**	12	11	13
40	21	**12**	3	12
41	21	6	2	**20**
42	8	**38**	0	3
43	**11**	22	11	4
44	19	10	**5**	14
45	**28**	7	9	5
46	**8**	13	26	1
47	20	6	6	**16**
48	13	18	**12**	7
49	24	7	8	**9**
50	11	**26**	6	5

Note: Greyed cells mark the correct answers.

4.3. *Item discrimination issues*

The following analysis considers those items that appeared to have good, average, or poor discrimination. As previously noted, the ID value in the current research was set at 0.30 and above.

4.3.1. Listening subtest: items 1 to 10

Once the ID value was set, the three subtests were subjected to a discrimination analysis. Table 11 shows that items 2, 3, 5, 6, and 7 have low levels of discrimination. These five items account for 50% of the items in the listening subtest.

Table 11. Item discrimination: Listening subtest (10 items)

Item	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
ID	0.36	0.29	-0.07	0.36	0.21	0.21	0.21	0.36	0.43	0.43

Further, item 3 obtained a negative discrimination value (-0.07), which indicates that the item is discriminating backward. That is to say, more poor-performing students correctly answered the question relative to high-performing students.

Item 2 is very close to the ID threshold with a value of 0.29, suggesting that with some modification, the item could probably be improved in future versions of the exam. Items 5, 6, and 7 have poor discrimination indices (0.21), so these items would be good candidates for elimination. On the other hand, although items 1, 4, and 8 (0.36 ID value) were above the acceptability threshold, they are somewhat close to limit and so revision is advised. Finally, items 9 and 10, with an ID of 0.43, appear to discriminate quite well.

4.3.2. Grammar subtest: Items 11 to 30

Table 12 shows that items 12, 16, 17, 23, 24, 25, 26, 27, 28, 29, and 30 (55% of the total items) had ID values below the acceptable threshold; the other nine items (11, 13, 14, 15, 18, 19, 20, 21, and 22) discriminated well.

TABLE 12. Item discrimination: Grammar subtest (20 items)

ITEM	I 11	I 12	I 13	I 14	I 15	I 16	I 17	I 18	I 19	I 20
ID	0.50	-0.07	0.43	0.36	0.57	0.29	0.29	0.36	0.36	0.71
ITEM	I 21	I 22	I 23	I 24	I 25	I 26	I 27	I 28	I 29	I 30
ID	0.43	0.50	0.00	0.21	-0.14	0.29	0.14	0.21	0.00	-0.29

Among the weak group of items, items 16, 17, and 26 had an ID value of 0.29, which suggests that these could be improved in future versions of the test. The ID value of the remaining items was very low. In fact, items 23 and 29 obtained discrimination indices of 0.00, meaning that the items did not discriminate at all (i.e., low- and high-scorers alike performed equally well on them). Even worse, more low-performing students correctly answered items 12, 25, and 30 than did high-performing students.

4.3.3. Reading subtest: items 31 to 50

The reading subtest was the weakest section of the exam, with 14 items (70% of the total number of items in this section) falling below the 0.30 ID threshold. Table 13 shows that only six items out of twenty were found to discriminate satisfactorily.

TABLE 13. Item discrimination: Reading subtest (20 items

ITEM	I 31	I 32	I 33	I 34	I 35	I 36	I 37	I 38	I 39	I 40
ID	-0.36	0.36	0.21	0.00	0.29	0.21	0.64	0.00	0.00	0.14
ITEM	I 41	I 42	I 43	I 44	I 45	I 46	I 47	I 48	I 49	I 50
ID	0.43	0.43	0.00	0.43	0.43	0.00	0.07	0.07	0.00	0.21

Items 31, 33, 34, 35, 36, 38, 39, 40, 43, 46, 47, 48, 49, and 50 all had ID scores below the acceptable range (values ranged from

-0.36 to +0.29). Items 34, 38, 39, 43, 46, and 49 shared an ID value of 0.00, which means that these items did nothing to distinguish between high- and low-performing students. Item 31 had a value of -0.36, i.e., only low-scoring students were able to answer this item correctly. This item should be removed in future versions of the exam. Item 37 was the strongest of the acceptable items, with an ID index of 0.64.

4.4. *Response frequency distribution issues*

As mentioned above, the RF distribution technique is utilized for a variety of purposes. In the present work, the RF distribution analysis was carried out to support the findings of the ID statistics. That is to say, it was carried out in an effort to explain why certain items received a low ID value.

4.4.1. Listening subtest: items 1 to 10

Looking at Table 14, it can be seen why items 2, 3, 5, 6, and 7 received a low score in the ID analysis: the correct answer was chosen by both low- and high-scorers.

Table 14. Response frequency distribution: Listening subtest (10 items)

Item	High / Low scorers	A	B	C
1	High scorers	0	14	0
	Low scorers	5	9	0
2	High scorers	0	14	0
	Low scorers	3	10	1
3	High scorers	0	1	13
	Low scorers	0	0	14
4	High scorers	1	13	0
	Low scorers	2	9	2

(continued)

Table 14. Response frequency distribution: Listening subtest (10 items)

Item	High / Low scorers	A	B	C
5	High scorers	10	0	4
	Low scorers	7	3	4
6	High scorers	1	4	9
	Low scorers	2	6	6
7	High scorers	0	13	1
	Low scorers	0	10	4
8	High scorers	12	1	1
	Low scorers	9	2	4
9	High scorers	0	14	0
	Low scorers	1	8	4
10	High scorers	9	4	1
	Low scorers	3	6	4

Note: Greyed cells refer to the correct answer.

There is little variation between the responses of the two groups, and that led to low ID indices. Items 5, 6, and 7 also did not discriminate well between the low- and high-performing students. The weakest item in this section was item 3. More low-performing students selected this item correctly (n = 14) than high-performing students (n = 13). This explains why the item's ID had a negative value.

4.4.2. Grammar subtest: items 11 to 30

Table 15 shows that there was only a very small variance between the responses of low- and high-scoring students to to items 12, 16, 17, 23, 24, 25, 26, 27, 28, 29, and 30. This resulted in very low ID indices.

Table 15. Response frequency distribution: Grammar subtest (20 items)

Item	High / Low scorers	A	B	C	D
11	High scorers	3	10	1	0
	Low scorers	6	3	2	3
12	High scorers	2	9	2	1
	Low scorers	4	6	1	2
13	High scorers	3	2	7	2
	Low scorers	3	4	2	4
14	High scorers	10	1	2	1
	Low scorers	6	2	1	4
15	High scorers	1	1	12	0
	Low scorers	8	0	5	1
16	High scorers	4	8	0	1
	Low scorers	3	4	3	3
17	High scorers	7	6	1	0
	Low scorers	3	2	2	5
18	High scorers	8	4	1	1
	Low scorers	4	8	0	1
19	High scorers	2	4	0	8
	Low scorers	5	5	1	3
20	High scorers	0	0	11	3
	Low scorers	3	1	1	9
21	High scorers	0	14	0	0
	Low scorers	2	8	3	0
22	High scorers	8	1	3	2
	Low scorers	1	6	4	3
23	High scorers	0	1	8	4
	Low scorers	0	3	7	4
24	High scorers	9	1	3	1
	Low scorers	6	3	2	2

(continued)

TABLE 15. Response frequency distribution: Grammar subtest (20 items)

ITEM	HIGH / LOW SCORERS	A	B	C	D
25	High scorers	12	1	1	0
	Low scorers	5	4	2	3
26	High scorers	1	11	2	0
	Low scorers	3	7	3	1
27	High scorers	2	5	0	6
	Low scorers	5	2	1	4
28	High scorers	4	6	3	0
	Low scorers	8	3	1	0
29	High scorers	2	11	0	0
	Low scorers	2	8	2	0
30	High scorers	3	2	7	1
	Low scorers	6	0	3	2

Note: Greyed cells refer to the correct answer.

Items 16, 17, and 26 had ID values of 0.29 and thus can probably be improved in a subsequent version of the exams. Items 23 and 29 had an ID index of 0.00. Items 12, 25, and 30 obtained negative ID values, suggesting that more low-performing students answered these correctly than did high-performing students. Item 20 was the strongest in this section, with an ID of 0.71.

4.4.3. Reading subtest: items 31 to 50

In this final part of the test, a total of 14 items out of 20 had unacceptably low discrimination indices (Table 16).

The major finding in the reading subtest section is that items 34, 38, 39, 43, 46, and 49 did not differentiate at all between low- and high-performing students. Item 31 discriminated particularly poorly: more students in the low-performing group answered the item correctly than students in the high-performing group. Finally, item 35, with an ID value of 0.29, could probably be improved and used in a future exam. Among all the items which discriminated effectively, item 37, with an ID value of 0.64, was the strongest: 11 high-scorers got the item right, as compared to only two low-scorers.

Table 16. Response frequency distribution: Reading subtest (20 items)

Item	High / Low scorers	A	B	C	D
31	High scorers	1	1	2	9
	Low scorers	4	1	7	2
32	High scorers	2	0	0	12
	Low scorers	5	1	1	7
33	High scorers	6	0	1	7
	Low scorers	3	1	2	8
34	High scorers	3	2	4	4
	Low scorers	3	1	6	4
35	High scorers	5	7	2	0
	Low scorers	1	7	4	1
36	High scorers	0	4	0	10
	Low scorers	2	1	2	9
37	High scorers	0	3	0	11
	Low scorers	1	5	5	2
38	High scorers	4	9	1	0
	Low scorers	3	8	1	2
39	High scorers	4	5	0	5
	Low scorers	4	3	5	2
40	High scorers	1	5	1	7
	Low scorers	10	2	1	1
41	High scorers	4	1	0	9
	Low scorers	7	2	2	3

(continued)

Table 16. Response frequency distribution: Reading subtest (20 items)

Item	High / Low scorers	A	B	C	D
42	High scorers	0	14	0	0
	Low scorers	5	8	0	1
43	High scorers	3	8	3	0
	Low scorers	3	6	4	1
44	High scorers	1	4	3	6
	Low scorers	10	2	1	1
45	High scorers	11	0	3	0
	Low scorers	5	3	3	3
46	High scorers	3	2	9	0
	Low scorers	3	6	5	0
47	High scorers	6	0	3	5
	Low scorers	9	1	2	2
48	High scorers	5	3	4	2
	Low scorers	5	4	3	2
49	High scorers	9	1	2	3
	Low scorers	5	3	3	3
50	High scorers	2	8	1	3
	Low scorers	4	7	2	1

Note: Greyed cells refer to the correct answer.

4.4.4. Summary

Table 17 displays the number of poor test items per exam section. Of particular note is the poor quality of the distractors throughout. In the listening section, 60% of the distractors did not function well; a full 80% of the distractors were deficient in the grammar section; and in the reading section, 75% of the distractors failed to distract. Overall, the university's placement exam can be deemed a failure in terms of its item facility, distractors, and item discrimination.

Table 17. Problematic items per exam section

	IF	DA	ID	Average
Listening section	3 items (30%)	6 items (60%)	5 items (50%)	4.6 items or 46% (based on 10 items)
Grammar section	6 items (30%)	16 items (80%)	11 items (55%)	11 items or 55% (based on 20)
Reading section	10 items (50%)	15 items (75%)	14 items (70%)	13 items or 65% (based on 20)

5. Conclusion

There are several very good reasons why multiple-choice tests should not be used in Mexican EFL classrooms. Chief among these is that such tests contradict the expressed goals of Mexico's national language curriculum. The 2016 curricular proposal issued by Mexico's Ministry of Public Education called for language teachers to help their students acquire the skills, knowledge, attitudes, and values necessary to participate in oral and written social practices with native and non-native speakers of English; use language to organize thought and discourse; analyze and solve problems and access different cultural expressions; recognize the role of language in the construction of knowledge and cultural values; and develop an analytical and responsible attitude towards the problems that affect the world. Multiple-choice exams do not support any of these goals and, indeed, likely run counter to them.

This study has reported on a single, 50-item test taken by 50 students. While the findings of this analysis cannot be generalized beyond the exam in question, it is not unreasonable to assume that analyses of similar teacher-created tests would yield similar results. At the very least, our findings highlight the extreme difficulty of creating MC exams and the indispensability of piloting tests and conducting item analyses before applying them for collecting data and making educational decisions about language students. Given the intractable and possibly insuperable obstacles to doing so, language programs considering the use of MC exams should seriously question the rationale for employing them. It is hoped that the results of this small-scale study contributes to such reflection.

6. References

Afflerbach, Peter; Parker, Emelie Lowrey; Armengol, Regla; Brooke, Leigh Baxley; Carper, Kelly Redmond; Cronin, Sharon M.; Denman, Anne Cooper; Irwin, Patricia; McGunnigle, Jennifer; Pardini, Tess, & Kurtz, Nancy P. (1995). Reading assessment: Teachers' choices in classroom assessment. *The Reading Teacher*, *48*(7), 622–624.

Alderson, J. Charles, & Clapham, Caroline (1995). Assessing student performance in the ESL classroom. *TESOL Quarterly*, *29*(1), 184–187.

Alderson, J. Charles; Clapham, Caroline, & Wall, Dianne (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Argüelles Álvarez, Irina (2013). Large-scale assessment of language proficiency: Theoretical and pedagogical reflections on the use of multiple-choice tests. *International Journal of English Studies*, *13*(2), 21–38. https://doi.org/10.6018/ijes.13.2.185861

Asim, Alice E.; Ekuri, Emmanuel E., & Eni, Eni I. (2013). A diagnostic study of pre-service teachers' competency in multiple-choice item development. *Research in Education*, *89*(1), 13–22. https://doi.org/10.7227/RIE.89.1.2

Bailey, Kathleen M. (1998). *Learning about language assessment. Dilemmas, decisions, and directions*. New York: Heinle & Heinle.

Bensoussan, Marsha (1984). A comparison of cloze and multiple-choice reading comprehension tests of English as a Foreign Language. *Language Testing*, *1*(1), 101–104. https://doi.org/10.1177/026553228400100109

Borjian, Ali (2015). Learning English in Mexico: Perspectives from Mexican teachers of English. *The CATESOL Journal*, *27*(1), 163–173.

Brooks, Jeffrey S., & Normore, Anthony H. (2018). *Foundations of educational leadership: Developing excellent and equitable schools*. New York: Routledge.

Brown, James Dean (2005). *Testing in language programs. A comprehensive guide to English language assessment*. New York: McGraw-Hill.

Milnes, Terry, & Cheng, Liying (2008). Teachers' assessment of ESL students in mainstream classes: Challenges, strategies, and decision-making. *TESL Canada Journal*, *25*(2), 49–65.

Cheng, Liying; Rogers, Todd, & Hu, Huiqin (2004). ESL/EFL instructors' classroom assessment practices: Purposes, methods, and procedures. *Language Testing*, *21*(3), 360–389. https://doi.org/10.1191/0265532204lt288oa

Cheng, Liying; Rogers, Todd, & Wang, Xiaoying (2008). Assessment purposes and procedures in ESL/EFL classrooms. *Assessment & Evaluation in Higher Education*, *33*(1), 9–32. https://doi.org/10.1080/02602930601122555

Cheng, Liying, & Wang, Xiaoying (2007). Grading, feedback, and reporting in ESL/EFL classrooms. *Language Assessment Quarterly*, *4*(1), 85–107. https://doi.org/10.1080/15434300701348409

Cheng, Winnie, & Warren, Martin (2005). Peer assessment of language proficiency. *Language Testing*, *22*(1), 93–121. https://doi.org/10.1191/0265532205lt298oa

Colby-Kelly, Christian, & Turner, Carolyn E. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *The Canadian Modern Language Review*, *64*(1), 9–37. https://doi.org/10.3138/cmlr.64.1.009

Currie, Michael, & Chiramanee, Thanyapa (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing*, *27*(4), 471–491. https://doi.org/10.1177/0265532209356790

Fulcher, Glenn, & Davidson, Fred (2007). *Language testing and assessment. An advanced resource book*. New York: Routledge.

Davies, Paul (2009). Strategic management of ELT in public educational systems: Trying to reduce failure, increase success. *The Electronic Journal for English as a Second Language*, *13*(3), 1–22.

Despagne, Colette (2010). The difficulties of learning English: Perceptions and attitudes in Mexico. *Comparative and International Education Society of Canada*, *39*(2), 55–74.

Dobson, Barbara K. (1975). Student-generated distractors in ESL tests. In Ruth Crymes & William E. Norris (Eds.), *On TESOL 74* (pp. 181–188). Washington: TESOL.

Fairtest: The National Center for Fair and Open Testing (2007). *Multiple choice tests*. http://www.fairtest.org/facts/mctfcat.html

García, Raquel, & Castañeda, Sandra (2006). Validación de constructo en la comprensión de lectura en inglés como lengua extranjera. *Razón y Palabra*, *51*(11). http://www.razonypalabra.org.mx/anteriores/n51/garciacastaneda.html

González, Elsa Fernanda (2017). The challenge of EFL writing assessment in Mexican higher education. In Patricia Grounds & Caroline Moore (Eds.), *Higher education English language teaching and research in Mexico* (pp. 73–100). Mexico: British Council Mexico.

Gyllstad, Henrik; Vilkaitė, Laura, & Schmitt, Norbert (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics*, *166*(2), 278–306. https://doi.org/10.1075/itl.166.2.04gyl

Hamp-Lyons, Liz (2004). The impact of testing practices on teaching: Ideologies and alternatives. In: Jim Cummins & Chris Davison (Eds.) *International handbook of English language teaching* (pp. 487–504). Boston: Springer. https://doi.org/10.1007/978-0-387-46301-8_35

Hales, Loyde W. (1972). Method of obtaining the index of discrimination for item selection and selected test characteristics: A comparative study. *Educational and Psychological Measurement*, *32*(4), 929–937. https://doi.org/10.1177/001316447203200407

Heaton, John Brian (1988). *Writing English language tests*. New York: Longman.

Hoshino, Yuko (2013). Relationship between types of distractor and difficulty of multiple-choice vocabulary tests in sentential context. *Language Testing in Asia*, *3*(16). https://doi.org/10.1186/2229-0443-3-16

Hughes, Arthur (1989). *Testing for language teachers* (2nd. ed.). Cambridge: Cambridge University Press.

Inbar-Lourie, Ofra, & Donitsa-Schmidt, Smadar (2009). Exploring classroom assessment practices: The case of teachers of English as a foreign language. *Assessment in Education: Principles, Policy & Practice*, *16*(2), 185–204. https://doi.org/10.1080/09695940903075958

Instituto Nacional para la Evaluación de la Educación (INEE) (2015). *Los docentes en México. Informe 2015*. http://www.senado.gob.mx/comisiones/educacion/docs/docs_INEE/Docentes_Mexico_Informe2015.pdf

In'nami, Yo, & Koizumi, Rie (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244. https://doi.org/10.1177/0265532208101006

Jia, Yueming; Eslami, Zohreh R., & Burlbaw, Lynn M. (2006). ESL teachers' perceptions and factors influencing their use of classroom-based reading assessment. *Bilingual Research Journal*, *30*(2), 407–430. https://doi.org/10.1080/15235882.2006.10162883

Kasper, Loretta F. (2002). Technology as a tool for literacy in the age of information: Implications for the ESL classroom. *Teaching English in the Two Year College*, *30*(2), 129–144.

Katalayi, Godefroid B. (2018). Elimination of distractors: A construct-irrelevant strategy? An investigation of examinees' response decision processes in an EFL multiple-choice reading test. *Theory and Practice in Language Studies*, *8*(7), 749–758. http://dx.doi.org/10.17507/tpls.0807.05

Katalayi, Godefroid B., & Sivasubramaniam, Sivakumar (2013). Careful reading versus expeditious reading: Investigating the construct validity of a multiple-choice reading test. *Theory and Practice in Language Studies*, *3*(6), 877–884. http://doi.org/10.4304/tpls.3.6.877-884

Ketabi, Somaye, & Ketabi, Saeed (2014). Classroom and formative assessment in second/foreign language teaching and learning. *Theory and Practice in Language Studies*, *4*(2), 435–440. https://doi.org/10.4304/tpls.4.2.435-440

Ko, Myong Hee (2010). A comparison of reading comprehension tests: Multiple-choice vs. Open-ended. *English Teaching*, *65*(1), 137–159.

Lado, Robert (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

Lee, Icy (2007). Assessment for learning: Integrating assessment, teaching, and learning in the ESL/EFL writing classroom. *The Canadian Modern Language Review*, *64*(1), 199–213. https://doi.org/10.3138/cmlr.64.1.199

Li, Zhi; Link, Stephanie; Ma, Hong; Yang, Hyejin, & Hegelheimer, Volker (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, *44*, 66–78. https://doi.org/10.1016/j.system.2014.02.007

Llosa, Lorena (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice*, *27*(3), 32–42. https://doi.org/10.1111/j.1745-3992.2008.00126.x

Madsen, Harold S. (1983). *Techniques in testing*. New York: Oxford University Press.

Mexicanos Primero (2015). *Sorry. El aprendizaje del inglés en México*. Mexico: Mexicanos Primero.

Nevo, Nava (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing*, *6*(2), 199–215. https://doi.org/10.1177/026553228900600206

Oller, John William (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Organisation for Economic Co-operation and Development (OECD). Stat: Education and training. https://stats.oecd.org/Index.aspx?DataSetCode=EDU_CLASS#

Organization for Economic Co-operation and Development (OECD) (2012). *Reviews of evaluation and assessment in education: Mexico*. http://www.oecd.org/education/school/Educational%20Evaluation%20and%20Assessment%20in%20Mexico%20-%20Strengths,%20Challenges%20and%20Policy%20Pointers.pdf

Pamplón Irigoyen, Elva Nora, & Ramírez Romero, José Luis (2013). The implementation of the PNIEB's language teaching methodology in schools in Sonora. *MEXTESOL Journal*, *37*(3), 1–14.

Payant, Caroline, & Barrón Serrano, Francisco Javier (2014). Assessing English in Mexico and Central America. In Antony John J. Kunnan (Ed.), *The companion to language assessment* (1st. ed.) (chapter 98). New Jersey: John Wiley & Sons. https://doi.org/10.1002/9781118411360.wbcla146

Pike, Lewis W. (1979). An evaluation of alternative item formats for testing English as a foreign language. *Educational Testing Service (ETS) Research Report Series*, *1*, 1–99. https://doi.org/10.1002/j.2333-8504.1979.tb01174.x

Poehner, Matthew E., & Lantolf, James P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research*, *9*(3), 233–265. https://doi.org/10.1191/1362168805lr166oa

Puhl, Carol A. (1997). Develop, not judge. Continuous assessment in the ESL classroom. *Forum*, *35*(2), 2–9.

Rodríguez Ramírez, Carlos (2014). Developing competencies under the national English program for basic education in Mexico: Is it possible? *MEXTESOL Journal*, *38*(2), 2–10.

Rupp, André A.; Ferne, Tracy, & Choi, Hyeran (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language testing*, *23*(4), 441–474. https://doi.org/10.1191/0265532206lt337oa

Sajadi, Fattaneh (2006). The effect of defective options on multiple-choice items and test characteristics. *Pazhuhesh-e Zabanha-ye Khareji: Special Issue, English*, *27*, 199–214.

Santibañez, Lucrecia; Vernez, Georges, & Razquin, Paula (2005). *Education in Mexico: Challenges and opportunities.* Santa Monica: RAND Education.

Schissel, Jaime L.; Leung, Constant; López-Gopar, Mario, & Davis, James R. (2018). Multilingual learners in language assessment: Assessment design for linguistically diverse communities. *Language and Education*, *32*(2), 167–182.

Secretaría de Educación Pública (SEP) (2010). *Programa nacional de inglés en educación básica. Segunda lengua: inglés. Programas de estudio 2010*. Mexico: Secretaría de Educación Pública.

Secretaría de Educación Pública (SEP) (2016). *Propuesta curricular para la educación obligatoria 2016*. Mexico: Secretaría de Educación Pública. https://www.gob.mx/cms/uploads/docs/Propuesta-Curricular-baja.pdf

Secretaría de Educación Pública (SEP) (2017). *Aprendizajes clave para la educación integral: plan y programas de estudio para la educación básica*. Mexico: Secretaría de Educación Pública. https://www.tamaulipas.gob.mx/educacion/wp-content/uploads/sites/3/2017/07/aprendizajes_clave_para_la_educacion_integral.pdf

Steadman, Mimi (1998). Using classroom assessment to change both teaching and learning. *New Directions for Teaching and Learning*, (75), 23–35.

Teemant, Annela (2010). ESL student perspectives on university classroom testing practices. *Journal of the Scholarship of Teaching and Learning*, *10*(3), 89–105.

Thanyapa, Inadaphat, & Currie, Michael (2014). The number of options in multiple-choice items in language tests: Does it make any difference? Evidence from Thailand. *Language Testing in Asia*, *4*(8). https://doi.org/10.1186/s40468-014-0008-7

Topping, Keith (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*(3), 249–276. https://doi.org/10.3102/00346543068003249

Tucker, Charlene G. (2015). *Psychometric considerations for performance assessment with implications for policy and practice.* New York: Educational Testing Service.

Waring, Hansun Zhang (2008). Using explicit positive assessment in the language classroom: IRF, feedback, and learning opportunities. *The Modern Language Journal*, *92*(4), 577–594. https://doi.org/10.1111/j.1540-4781.2008.00788.x

Weigle, Sara Cushing (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, *16*(3), 194–209. https://doi.org/10.1016/j.jslw.2007.07.004

Yi'an, Wu (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, *15*(1), 21–44. https://doi.org/10.1177/026553229801500102

Zheng, Ying; Cheng, Liying, & Klinger, Don A. (2007). Do test formats in reading comprehension affect second-language students' test performance differently? *TESL Canada Journal*, *25*(1), 65–80. https://doi.org/10.18806/tesl.v25i1.108