EFFECT OF THE NUMBER OF OPTIONS ON THE QUALITY OF EST READING COMPREHENSION MULTIPLE-CHOICE EXAMS

Gilberto Berríos,
Carlina Rojas,
Noela Cartaya,
Yris Casart
Universidad Simón Bolívar -Sartenejas

Abstract

This study aims at identifying how the quality of English for Science and Technology (EST) reading comprehension multiple-choice tests is affected by reducing the number of options in each item from 4 to 3. The literature proposes that 3-option multiple-choice tests have a similar quality to that of 4-option ones, demanding less effort and being more efficient. Items from 2 exams administered in 2001 at Universidad Simón Bolívar were modified. The new forms were administered in 2002. From a sample of about 25% of the population, descriptive statistics were obtained: mean difficulty, mean discrimination, and reliability coefficient. Results confirm the practicality of the 3-option format without affecting the quality of the test. We recommend the adoption of this format for the reading program in question. A subsequent decision will need to address either increasing number of items or decreasing administration time.

Key words: language testing, multiple-choice tests, test quality, number of options.

EFECTO DEL NÚMERO DE OPCIONES EN LA CALIDAD DE LAS PRUEBAS DE SELECCIÓN SIMPLE DE COMPRENSIÓN DE LECTURA DE TEXTOS TÉCNICO-CIENTÍFICOS EN INGLÉS

Resumen

Este estudio pretende determinar cómo se afecta la calidad de las pruebas de comprensión de lectura de textos técnico-científicos en inglés al reducir el número de opciones de 4 a 3 en cada ítem de selección simple. La literatura del área propone que las pruebas de selección simple de 3 opciones tienen una calidad similar a las de 4 opciones, al tiempo que exigen menor esfuerzo y son más eficientes. Se modificaron dos exámenes con ítemes de 4 opciones, administrados en el 2001 en la Universidad Simón Bolívar (USB). Las nuevas versiones de 3 opciones se administraron en el 2002. Se obtuvieron estadísticas descriptivas ―dificultad media, discriminación media y coeficiente de confiabilidad― de los datos obtenidos en 2001 y 2002 de una muestra de alrededor del 25% de la población. Los resultados confirmaron la practicidad del formato de 3 opciones y mostraron cambios positivos en cuanto a la confiabilidad. Aspectos como validez, autenticidad, interactividad e impacto no se vieron afectados de manera adversa. Recomendamos la adopción de este formato para el programa de lectura de la USB. Como consecuencia de esa adopción, se desprende que, o bien se aumenta el número de ítemes, o se disminuye el tiempo de administración.

Palabras clave: evaluación de L2, pruebas de selección simple, calidad de las pruebas, número de opciones.

Recibido: 14/03/2005 **Aceptado:** 24/05/2005

Theory vs. praxis

Although several theoretical and empirical works have suggested that three is the optimum number of options for multiple-choice questions (Owen & Froman, 1987; Haladyna & Downing, 1993), a number of texts on second language evaluation, teaching practice as well as exams such as TOEFL (*Test of English as a Foreign Language*) still favor the use of four- or five-option items. It seems that the reason for this practice is the assumption that four- and five-option items reduce the probability that low ability students may answer right by guessing.

A look at various reference works in the language testing literature reveals that the issue of number of options is not given detailed coverage. Bachman & Palmer's (1996) book does not mention an optimal number of options under its 'selected response' items section. And, while some authors explicitly recognize that four-option multiple-choice questions are the most common format (Cohen, 1994; Bailey, 1998; Alderson, 2000), others do so implicitly by offering examples throughout their books that follow this practice (Read, 2000; Douglas, 2000; Davidson & Lynch, 2002). All of these authors would agree that "using four alternatives instead of three decreases the likelihood of getting the item right by chance alone" (Cohen, 1994, p. 232).

However, Haladyna (1994) states that "item writers are often frustrated in finding a useful fourth or fifth option because they do not exist" (p. 75). Likewise, Alderson, Clapham, and Wall (1995) recommend using four options, but they also add that "if it is impossible to think of a third attractive wrong answer, then it is sensible to have only three alternatives for some items" (p. 48).

Coverage of the number of options issue is more detailed in the empirical literature. Owen and Froman (1987) reviewed theoretical and empirical studies which pointed out that three-option items should be widely used. Some references go back to the early 20th century: for example Toops' work (as cited in Owen & Froman) dates back to 1921. Owen and Froman also mention Williams and Ebel's work published in 1957: they took a four-option, 150-item test and dropped the least discriminating distractors in order to produce two- and three-option exams. These exams ended up having similar or higher score reliability indexes than standard four-option exams. In other words, it seems that overall test quality as reflected by these indexes increased as a result of eliminating those wrong answers that were chosen either by no student at all or by students whose score was greater than or equal to the score of students choosing the right option.

Haladyna and Downing (1993) assert that multiple-choice items rarely have more than two efficient distractors (i.e., incorrect options), which suggests that "three options per item may be a natural limit for multiple-choice item writers in most circumstances" (p. 1008). These authors summarize the advantages of using only two distractors in the reduction of (a) item writing time, (b) exam length, (c) printing costs, and (d) administration time. In general, a number of researchers agree that three-option multiple-choice exams require less effort from test writers and test takers and also that these tests are more efficient (Owen & Froman, 1987; Trevisan, Sax & Michael, 1991; Haladyna, 1994), something which is also stated variously in a November 2001 discussion held in LTEST-L, an electronic discussion group dealing with language testing (Retrieved July 19, 2004, from: http://lists.psu.edu/cgi-bin/wa?A0=ltest-l). These advantages could be translated into an increase in the number of items that can be given in a certain time, which in turn increases score reliability. In fact, Zimmerman and Humphries (as cited in Owen &

Froman) found in 1953 that reliability was increased and administration time was reduced as the number of distractors decreased. All these authors propose that more studies should be done in order to support the effectiveness of the three-option format.

We have also noticed recently that some tests such as DELE (*Diploma de Español Lengua Extranjera*, the Cervantes Institute's Spanish proficiency test) and TestDaF (*Test Deutsch als Fremdsprache*, a German proficiency test) use multiple-choice items with three options in their reading comprehension sections.

The Modern Languages Department at Universidad Simón Bolívar- Sartenejas has been using four-option multiple-choice tests of reading comprehension since the early 80s. Most teachers involved in item-writing have found that writing a good third distractor is hard. Also, in analyzing test results, we have noticed that many times this fourth option was not appealing to test takers. In view of the above, we decided to explore the possibility of making and administering three-option departmental exams in order to corroborate what the literature says about the effectiveness of this format.

The study

We decided to answer the following research questions: How is the quality of our multiple-choice reading comprehension tests affected by reducing the number of options per item from four to three? If, as the literature seems to suggest, four-option item tests have the same quality as those with three-option items, why do we need a fourth option? Is it reasonable to advocate for the three-option format in our context?

Background

The First Year English Program of the Sartenejas campus of Universidad Simón Bolívar teaches reading strategies which allow students to understand science and technology texts in English. The Modern Languages Department offers this program to Engineering and Science students in their freshman year. It consists of three 48-hour courses which are offered quarterly (four hours per week during 12 weeks).

Within this framework, the departmental exams are achievement tests designed to measure the students' development of reading skills throughout the term. The first departmental exam is administered around the middle of the term and the second one towards the end. These exams are put together by the department's Exam Commission and then administered to the entire student population at the same time. Depending on the course, each exam has either 25 or 20 discrete items worth one point each. The first two, skill-oriented, courses have 25-item departmental exams. The last course is more topic-oriented and has 20-item departmental exams, leaving more evaluation weight to teacher quizzes that assess critical and comparative reading. Exams are put together from the Department's Item Bank. All teachers have access to the test about a couple of weeks before it is actually given.

Our departmental exams do not have testlets in the sense implied in the literature, i.e., a reading passage followed by several multiple-choice questions. Instead, the teachers who originated our exam system preferred to view these exams as sequences of *modules* (or *modular items*). A module is a short text followed by *one* multiple-choice question (Llinares de Alfonzo & Berríos Escalante, 1990, p. 43). The text ranges in length from a sentence to several short paragraphs, one-sentence "passages" typically used for testing understanding of complex noun phrases and other such low level objectives.

In 1980, an Evaluation Commission formed by Donna Archibald, Pauline de Marín, and Elías Foubert (...) wrote a *Working paper on testing* advocating for closed-ended (multiple-choice) departmental exams (Archibald, Marín & Foubert, 1980). In 1980-81, together with Cheryl Champeau and Susana Kertész, they produced a massive amount of five-option multiple-choice items which were later gradually tried out to become the standard four-option items that characterize most of our departmental exams. Their working folders attest to their use of the term "modules."

These teachers' outlook towards reading comprehension exam construction was innovative in that they tried to overcome some of the perceived drawbacks of traditional single-passage testlets, namely: (a) some questions may give away the answers to other questions; (b) undue advantage may be given to some students due to their background knowledge; (c) text characteristics limit the variety of objectives that can be tested; and (d) the order of questions is not very flexible (Llinares de Alfonzo & Berríos Escalante, 1990, p. 43).

Our *modular items* are shuffled in order to create four different versions of the test. Since this is a reading comprehension test, there is no correction for guessing: we expect that, in those few questions students are likely to find too difficult, a healthy measure of educated guessing can give the best of them a feel for what Goodman (1967) called the "psycholinguistic guessing game." In order to improve future item and exam quality, students' answers are routinely analyzed using the classical statistic software called Lertap 5 (Nelson, 2000). Administration time is ninety minutes.

Figure 1 shows a sample modular item with indication of its parts. Table 1 and Table 2 give an idea of the objectives and task types of our departmental exams by characterizing the tests used in this study. Regarding test structure, items are not presented in subtests; instead, all items appear in one sequence without any special rubric. Each item is worth one point, independently of the conceptual level of the task presented by its question. Reading ability has been operationalized as a syllabus-based construct where reading is seen as "a complex behavior which involves conscious and unconscious use of various strategies including problem-solving strategies, to build a model of the meaning which the writer is assumed to have intended" (Johnston, as cited in Mikulecky, 1989, p. 2). Our program is based on the Interactive Model, which combines both the bottom-up and the top-down processes alternatively or simultaneously incorporating the reader's background knowledge, language proficiency level, motivation and use of strategies among other factors, to the new information in the text (St. Louis, 2001).

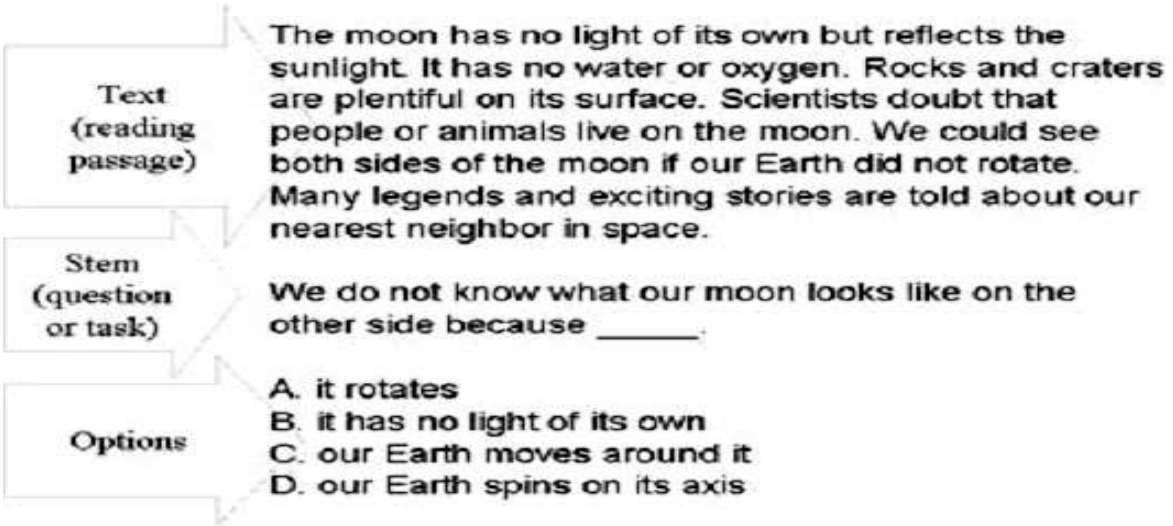

Figure 1. Sample modular item

Table 1. Test characterization by reading objective

	Number of items per objective			
Reading objective covered	1A01	1A02	3A01	3A02
Determine meaning from context	2	3	-	-
Determine referent	1	1	-	-
Identify text function	1	0	-	-
Recognize term being defined	1	1	-	-
Recognize what is being described	1	1	-	-
Recognize class relationships	1	1	-	-
Recognize paraphrase of specific information	2	2	-	-
Organize sentences into a coherent text	1	1	-	-
Recognize main idea or topic sentence	1	1	1	1
Select appropriate title	1	1	-	-
Recognize author's purpose	3	4	-	-
Infer from implicit information	1	1	5	5
Predict what follows from implicit information	1	1	3	3
Recognize cause and effect	-	-	1	1
Distinguish between fact and hypothesis	-	-	1	1
Infer from explicit information	8	7	9	9
Total items per exam	25	25	20	20

Table 2. Test characterization by item task type

		Number of items per task type			
Task type	**Task type description**	**1A01**	**1A02**	**3A01**	**3A02**
Implicit	Text ends in a blank: student chooses best ending	3	3	1	1
Implicit	Task partially expressed in stem: student infers task	5	5	14	14
Explicit	Task fully expressed in stem	17	17	5	5
Total items per exam		25	25	20	20

Objective

The aim of this study is to determine how the overall quality of multiple-choice exams is affected by the reduction from four to three options per item.

Method

The population was made up of about 600 first year students distributed in intact groups according to their assigned sections. All students took all items in each test. For each exam under study, a sample of roughly 25% of the population was analyzed. This sample was made up by all the students who took the base version of the exam (i.e., the version whose items are shuffled to originate the other three). Because of the way exam versions are distributed, the sample included students from all sections. This sample's initial proficiency as measured by our standardized entry exam, age range, and instruction was comparable to that of students not included in the sample.

The rationale for this sampling procedure comes from a previous study that "...did not find statistically significant differences between examinee score distributions as a function of exam version" (Berríos & Iribarren, 1989, p. 19; Spanish in original). The same study also found that no significant differences resulted from the order of presentation of items throughout versions. This is why a department policy has been in effect since the mid-90s to run item analysis statistics in a cost-effective way, i.e., only on the base version of the test, which is approximately 25% of the population.

In order to achieve our goal, we modified items from two departmental exams administered in 2001 using a procedure that is similar to Williams and Ebel's (as cited in Owen & Froman, 1987). The modified forms were administered in 2002. The original four-option exams were coded 1A01 and 3A01. The three-option forms were coded 1A02 and 3A02.[1]

Table 3 shows our modification procedure. To make the conversion from four to three options, the possible distractors to be dropped were identified on the basis of the Lertap 5 statistical analysis. The criteria to be taken into account were frequency, i.e., the number (n) of students choosing the option, and discrimination, i.e., the ability of an item to measure individual differences sensitively. Regarding frequency, the distractor that was chosen by the least number of subjects was eliminated. With regard to discrimination any distractor with a positive discrimination or with a negative discrimination too close to zero, was eliminated. Once the distractor was eliminated, the task was reviewed, the homogeneity of the options was checked, the text was re-read, and appropriate changes were made.

Table 3. Steps for eliminating the worst distractor when modifying four-option items.

1. Identify candidates for elimination on the basis of statistical information

1.1 Examine expected response discrimination — pb(r) must be positive.

- Mean must correspond to the most able group. See average (avg.) column.
- The option corresponding to the correct, i.e., expected, response cannot be eliminated.

1.2 Examine distractor discrimination —pb(r) must be negative.

- If it is high, for instance, -0.45, the distractor stays. If it is low, for example, -0.12, the distractor may have to be eliminated.
- Mean must correspond to least able groups.
- As long as pb(r) is closer to zero, the distractor attracts a group more like the high-ability one and thus its discrimination is worse. You can check this by looking again at the avg. column.
- A distractor with positive discrimination is attracting the high-ability group.

This distractor must either be rephrased to make it unappealing to high-ability students or, more probably, eliminated.

1.3 Identify the distractor that attracts the least number of subjects

- Review n or p.
- Sometimes one can eliminate the distractor that attracts the lowest number of students.

2. Review the item

2.1 Review the task

- Examine the explicit or implicit item question.
- If necessary the question may be edited for task clarity, focus, or conceptual reorientation.

2.2 Verify option homogeneity

- No option should be particularly different from the rest in length, lexis, syntax, format, etc.
- Edit options if to make them more homogeneous, that is, more similar to each other in the appropriate aspect mentioned.

2.3 Re-read the text

- Verify that the text is still valid for the options remaining after distractor elimination.
- Edit text only if absolutely necessary.

3. Record agreement

- After discussing with other Committee members, briefly describe in writing the kind of editing done to produce the new item: eliminated option, task changes, and changes in remaining options, text, or question.

4. Create and file a new printed version of the item in the Item Bank

With the aim of illustrating the process of elimination of the worst distractor, Table 4 presents statistics for two sample items. The number of students that chose any of the options is represented by the value 'n'. Also, the value 'p' represents the proportion of the sample that chose any of the options. The point-biserial correlation for each option is indicated under the column marked 'pb(r)' and the average grade of the group that chose any of the options is listed under the column marked as 'avg.' Column 'z' indicates the z-scores. The asterisk (*) indicates the right answer. The text of the items is also given to provide a practical reference to teachers willing to become test constructors.

Table 4. Statistics for two sample items from 1A01

Item 06[a]	option	n	p	pb(r)	avg.	z
	*A	95	0.56	0.23	16.28	0.31
	B	9	0.05	0.02	15.33	0.07
	C	41	0.24	-0.26	13.17	-0.46
	D	25	0.15	-0.18	13.32	-0.43

Item 08[b]	option	n	p	pb(r)	avg.	z
	A	44	0.26	-0.38	12.48	-0.64
	B	18	0.11	-0.08	14.06	-0.25
	C	9	0.05	-0.22	11.22	-0.95
	*D	98	0.58	0.40	16.79	0.43

[a] Item 06 was like this:

06. In its simplest terms, a solar heating/solar cooling system is any system which reduces consumption of conventional fuels by utilizing the sun's energy as a method of heating or cooling. The system can be either passive or active. People utilize passive solar energy when opening the curtains on the sunny side of the house during the cold months to let the sun add its warmth. There is a minor amount of control over this system during the cold months to let the sun add its warmth. There is a minor amount of control over this system during the summer when the curtains are closed to block out the sun's rays and the unwanted heat radiated within the house.

It is predictable that the sentence which continues this paragraph should _____.

A. define or describe an active solar energy system

B. discuss the variables concerned with fuel costs

C. suggest an alternative method to that of solar heating and cooling

D. draw a definitive conclusion

[b] Item 08 was like this:

08. Smallpox, one of the world's most dreaded plagues, is an acute infectious disease characterized by fever, and after two days, an eruption, which passes through the stages of papule, vesicle and pustule. These then dry up, leaving more or less distinct scars.

Which of the following CANNOT be concluded from this text?

A. One's skin is left with scars after the disease.

B. Papules, vesicles and pustules follow a high fever.

C. Two symptoms of smallpox are fever and skin eruption.

D. The scars that are left by smallpox are dangerous.

There are some considerations on item analysis that need to be examined in order to appraise these two sample items. First, it is important to be aware of the fact that, in a multiple-choice test, we need to know how language ability is reflected in option selection. High language

ability should correlate highly with choosing the correct option, whereas low language ability should correlate highly with choosing any of the distractors. This is what point-biserial correlation shows.

Concerning the data provided by the point-biserial correlation, it is pertinent to state here that, according to Haladyna (1994), the point-biserial index found in most item analysis computer programs is the product-moment correlation between distractor performance and total test score. It is an index which "considers the average performance of those selecting the distractor versus the average of those not selecting the distractor" (p.157). This coefficient helps us see how students who selected an option did on the whole test as a criterion measure. But, what is a high pb(r)? Nelson (2000, p. 100) says "anything over 0.30," and then cites Hills (p. 102): "...many experts look with disfavor in items with correlation discrimination values less than +.30. Teachers will often be not as good at writing items as experts are, however, and acceptable items may have discrimination values in teacher-made tests as low as +.15..." The value of pb(r) for the correct option is known as the item's discrimination index. On the other hand, the value of pb(r) for *distractors* is expected to also perform strongly but towards the negative end, and so a pb(r) of -0.30 is a better discrimination value for a distractor than -0.15. A positive pb(r), say +0.10, for a distractor means it is attracting high-ability students. This distractor distracts people from the wrong group and we should not want it in our test. A negative pb(r) of -0.05 for a distractor may mean it is attracting random guessers. It is probably the case that the distractor implausibility is only attracting "a few random guessers" (Haladyna, 1994, p. 157).

Also, it is relevant to review the meaning of z-scores. Z-scores are standardized scores, "the distance of a score from the mean, as measured by standard deviation units" (Ary, Jacobs, & Razavieh, p. 110). This scale is simpler to interpret than we might think: the mean score is set to zero and distance to it is indicated by positive z-scores for students above test average and by negative z-scores for students *below* test average. So, a positive z-score is logically expected from students who chose the correct option and a negative z-score is expected from those who chose any distractor. You can use the z-scores column to quickly get a grasp of what kind of student chooses an option. This is sometimes more informative than knowing the actual raw scores.

With these considerations in mind, we can look back to Table 4. In item 06, option B is the worst distractor because it was chosen by the smallest number of students (n = 9), which represents a proportion of 5% of the sample (p = 0.05). Also, this option's discrimination index as shown by the point-biserial correlation —pb(r) = 0.02— is positive, something undesirable for a distractor. The average score of the students who chose this distractor was 15.33. The z-scores column shows us that this score is actually above average, which means that this option was chosen by students from the upper ability group.

In the case of item 08, option B is also the worst distractor, not because it was chosen by the smallest number of students (n = 18; p = 0.11), but because its discrimination index —pb(r) = -0.08—, although negative, is closer to zero than that of distractors A and C. In fact, a look at the last two columns confirms that students choosing option B in item 08 are more able than those choosing either option A or C.

To illustrate the change that resulted from our modification procedure, Table 5 shows the statistics for the same two sample items of Table 4 (this time converted to three options). The modified items are also given for the benefit of test constructors.

Table 5. Statistics for two sample items from 1A02 ─ converted

Item 06[a]	option	n	p	pb(r)	avg.	z
	*A	94	0.58	0.23	17.39	0.29
	B	41	0.25	-0.25	14.39	-0.43
	C	26	0.16	-0.13	14.92	-0.31
[b]	N	1	0.01	-0.14	9.00	-1.74
Item 08[c]	option	n	P	pb(r)	avg.	z
	A	44	0.27	-0.31	14.09	-0.51
	B	10	0.06	-0.18	13.30	-0.70
	*C	108	0.67	0.28	17.31	0.27

[a] The modified item 06 was like this:

06. In its simplest terms, a solar heating/solar cooling system is any system which reduces consumption of conventional fuels by utilizing the sun′s energy as a method of heating or cooling. The system can be either passive or active. People utilize passive solar energy when opening the curtains on the sunny side of the house during the cold months to let the sun add its warmth. There is a minor amount of control over this system during the cold months to let the sun add its warmth. There is a minor amount of control over this system during the summer when the curtains are closed to block out the sun′s rays and the unwanted heat radiated within the house.

Most probably, the next idea should _____.

A. define or describe an active solar energy system

B. suggest an alternative method to that of solar heating and cooling

C. draw a definitive conclusion about solar energy systems

[b]In item 06, N means ‘did not answer′

[c]The modified item 08 was like this:

08. Smallpox, one of the world′s most dreaded plagues, is an acute infectious disease characterized by fever, and after two days, an eruption, which passes through the stages of papule, vesicle and pustule. These then dry up, leaving more or less distinct scars.

Which of the following CANNOT be concluded from this text?

A. One′s skin is left with scars after the disease.

B. Two symptoms of smallpox are fever and skin eruption.

C. The scars that are left by smallpox are dangerous.

Haladyna (1994, p. 146) says that test score reliability depends on item discrimination: "The size of the discrimination index is informative about the relation of the item to the total domain of knowledge, as presented by the total test score." It follows then that improving distractor quality will improve reliability. This is why we examined option behavior by using a statistical program that allowed us to make decisions based on distractor quality.

Besides item statistics, Lertap 5 also provides total-test descriptive statistics. We were specifically interested in average item difficulty, average item discrimination, and score reliability, which were the same criteria used by Owen & Froman (1987). Additional descriptive statistics for the whole test were run in Excel.

In order to understand the average difficulty of the test, we should examine the difficulty index first which is represented by the value 'p' for an item's correct answer. It is the result of expressing the number of students who chose the correct option as a proportion (or percentage) of all students taking the test. Roughly speaking, easy items have high p-values, e.g., 0.85 (85%) and up, whereas low p-values indicate difficult items, e.g., 0.40 (40%) and lower. Mid-range difficulty indexes are around 0.50 to 0.65 (50% to 65%). Some authors prefer to call this an index of item facility to avoid potential confusions since the name 'facility' goes in the same direction of the index. *Therefore, the test's average diffi*culty is the average of all the items' difficulty indexes. Likewise, the test's average discrimination, is the average of all the items' point-biserial correlation (pb(r)-value), which was discussed above.

The index of reliability used by Lertap 5 is Cronbach's alpha coefficient. It is a measure of internal test consistency which has to do with how well test items relate to each other. A high alpha means items correlate highly. As is the case with correlation indexes, coefficient alpha's maximum value is 1, and it implies a perfect correlation. Any value of alpha between .60 and .85 is valid for this kind of tests (Nelson, 2000).

We also compared test administration time in a qualitative way by informal communication with all proctors as well as students.

Although our type of test does not seem to have an order of presentation effect (Berríos and Iribarren, 1989), we decided to keep items in 1A02 and 3A02 in exactly the same positions as in the original forms 1A01 and 3A01 to facilitate item-by-item comparisons later.

Results

Table 6 shows a comparison between the statistics for 1A01 (four options) and 1A02 (three options) and also between 3A01 (four options) and 3A02 (three options).

As seen in Table 6, there is an increase in the difficulty, discrimination, and reliability indexes associated with the three-option test (1A02) as compared to the four-option one (1A01). These increments make the three-option test easier and more discriminating than the four-option form. Scores from the three-option test are also more reliable. Since these are achievement tests, the increase in difficulty (or facility) index is desirable because it could mean that more students are accomplishing the objectives of the course. The three-option form still discriminates

similarly between the upper and lower ability students. The reliability of scores from the three-option form is marginally higher than that from the four-option one.

Table 6. 1A01 vs. 1A02 results and 3A01 vs. 3A02 results

	1A01	1A02	3A0	3A02
	4-option	3-option	4-option	3-option
number of examinees	662	676	570	506
number of examinees in sample (n)	170	162	135	131
average difficulty	0.60	0.65	0.57	0.63
average discrimination	0.24	0.27	0.29	0.30
reliability (coefficient alpha)	0.69	0.72	0.73	0.73
number of items	25	25	20	20
mean (or average)	15.05	16.19	11.36	12.60
median	15	16	11	13
mode	15	16	13	14
sample variance	16.45	17.18	14.68	14.52
standard deviation	4.06	4.15	3.83	3.81
minimum score	4	5	3	2
maximum score	24	24	19	20
range	20	19	16	18
standard error of measurement	2.26	2.17	2	1.96
kurtosis	-0.39	-0.61	-0.71	-0.62

skewness	-0.35	-0.34	-0.01	-0.17

Values in Table 6 for 3A01 (four options) and 3A02 (three options) indicate similar results in terms of the increase of difficulty and discrimination. Score reliability remained constant from 3A01 to 3A02. The central tendency and dispersion measures of each pair of tests (1A01-1A02 and 3A01-3A02) are similar, which tells us that score distributions can be considered equivalent. The standard error of measurement, a practical index of the accuracy of test scores, decreased for both modified forms (1A02 and 3A02).

Regarding administration times, all proctors reported that students took less time with the three-option form. A commonly reported fact was that fewer students stayed until test end was called. Students also reported taking less time to answer each question.

Discussion

To discuss our results, we would like to use Bachman and Palmer′s (1996) model on the qualities of test usefulness. These qualities are reliability, construct validity, authenticity, interactiveness, impact, and practicality.

The quality of our multiple-choice exams was in fact affected by reducing the number of options from four to three. For one thing, the three-option multiple-choice format lowered administration time. Also, the mean difficulty, mean discrimination, and reliability coefficients of the four- and three-option tests studied are, to say the least, equivalent. This seems to corroborate the claim made by the literature about the little practical significance of using items with more than three options, especially in view of the complexities implied by writing adequate third or fourth distractors.

From an information user′s point of view, the most efficient measure is the one that yields the same accuracy with the use of fewer resources. This study suggests that the three-option multiple-choice tests studied are more efficient than their four-option counterparts in terms of the accuracy/resource ratio. An added benefit comes from their lower standard error of measurement, which implies higher score accuracy in the three-option tests.

All this indicates that a positive change occurred with respect to practicality, i.e., how a test is developed and implemented with the resources available (Bachman & Palmer, 1996, pp. 37-37, 39-40). Clearly, our results advocate for three-option tests as easier to implement within our human, material, and time resource restrictions. From our experience in writing four-option tests, we can also extrapolate that designing and developing three-option ones from scratch will also be simpler. There is also a positive change with respect to reliability, i.e., how consistent scores are across tests, across test tasks, or across test task characteristics (Bachman & Palmer, 1996, pp. 19-21). The modification procedure actually minimized with some success variations in scores, particularly in the longer test, 1A02: we would like to think that at least some of the marginal improvement on this particular test′s reliability comes from our intervention of the test task characteristics known as task specification. In other words, sometimes the few changes introduced actually seem to have produced modular items whose question element (stem or otherwise) ended up being clearer and more to the point, thus pointing test-takers in the appropriate direction to approach the task.

From a methodological point of view, we strived to guarantee robustness by using samples from comparable populations not only in terms of initial proficiency and age range, but also in terms of size and general conditions of instruction. The exams were also comparable in that items converted from four to three options covered exactly the same topics since reading passages remained the same. The exams also had almost exactly the same objectives since most stems (the part of our modular items presenting the task or question) were unchanged.

Although this is not strong evidence of construct validity, it does suggest that, at the very least, validity was not adversely affected in the modified test forms. And, if we consider Bachman and Palmer's note that authenticity and interactiveness can be seen as related to construct validity (1990, pp. 42-43), we can safely assume that these qualities of our modified tests also stayed the same as in the original forms. Authenticity refers to the relevance of our test tasks to the target language use domain, in our case, understanding an EST reading text for a specific purpose. And that was an aspect that our modification procedure expected to improve when a task (the item's question) was reviewed for specificity and/or clarity. Students in 2002 had no access to the 2001 form of their exam to provide their opinion but, to judge from the teachers' informal oral reports, the modified forms of the test were clearer and more to the point. Regarding interactiveness, this quality of test usefulness denotes the contribution of examinees' individual characteristics in carrying out our test tasks. Again, task nature did not change dramatically from the original to the modified versions, which suggests that in any case students had to approach them in a similar manner as their 2001 predecessors.

Impact also did not seem to be adversely affected by our intervention of the original tests. Impact is the name given to the value- and goal-based consequences that a test has for both the individuals and the educational system involved. According to Bachman and Palmer (1996, pp. 31-33), impact on test-takers hinges on the test taking experience (which, in our case, was reported as fairer by students taking the three-option forms), but also on the feedback they received about their test performance and the decisions made about them on the basis of their score (these were not changed at all in the modified versions).

Bachman and Palmer (1996) describe authenticity in terms of how well the characteristics of a test task correspond to the characteristics of a target language use domain task. In other words, the authenticity of a test is the test's match with real-life tasks in the target language. Although we had not intended to do so, the review step of our modification procedure introduced minor changes in an item's stem and options which may have made students perceive test tasks as more authentic, i.e., more life-like in that they were more like actual questions a reader would try to answer when faced with EST texts.

So, on the whole, it seems reasonable to say that the quality of our three-option tests was, in the worst case, similar to that of our four-option tests and, in the best case, marginally improved with respect to our four-option tests. But there is still another issue.

A central aspect to discuss in a study of multiple-choice tests is guessing. We think that there is a place for guessing in multiple-choice tests of EST reading comprehension. If we are to follow Dycos's (1997) useful essay, high-ability students have a good chance at correctly guessing word meaning based on local context, i.e., the immediate intrasentential and sentential level. They will also have a good chance at using well all the intersentential and discourse level information available to them, as well as their world knowledge: this is what good L2 readers do in the absence of a dictionary or an English speaker. A multiple-choice question whose answer students are not sure about may trigger Goodman's (1967) psycholinguistic guessing game in its

appropriate, top-down, form. Do we want our items to elicit this kind of guessing? Not by design. But if an item is well written and the students' reading ability is high, good use of the guessing strategy will give them more context so that the initial difficulty of the item may become less as a consequence of a reading proficiency they brought with them to the exam. The result is that answering an item by blindly or systematically choosing an option is less likely in high-ability students. And even when one of these students gives up on an item or two and guesses their answer, we can be pretty much sure that there has been an honest reading attempt before.

The story is different when we talk about other types of students. Dycos says that low-ability students tend to use the guessing strategy but only because they have to expand their scant vocabulary. Mid-ability students use the guessing strategy less than either high- or low-ability ones. For both low- and mid-ability students, though, guessing as a reading strategy is needed but is not necessarily well mastered yet. This fact plus the massive amount of guessing needed by the weaker of these students will most likely conduce them to wrong hypotheses and therefore to mindful selection of wrong answers. If the strategy is used a little better, it might lead to choosing the correct answer without fully understanding the text. In the case of our tests, this guessing is not entirely bad because, again, there has been an honest reading attempt.

In the case of random or systematic guessers, there is not much any multiple-choice test constructor can do since all such tests contain an element of guessing, regardless the amount of options the items might have. If our test could be longer, we could in fact completely ignore the influence of guessing since "the probability of getting a higher-than-deserved score by guessing is very small as the test gets longer" (Haladyna, 1994, p. 152). However, we will hardly be able to take our modular tests to compete in length with the Williams and Ebel's 150-item test which is at the base of our study.

We think we can live with the present state of guessing in our departmental exams based on Hambleton and Swaminathan's study (cited in Haladyna, 1994, p. 152) which found that the guessing parameter's influence is small in relation to the influence of the discrimination parameter. Besides, if test reliability stayed the same (or goes up) in the modified test forms, item discrimination (which directly influences reliability) must be about the same (or better) than in the original test forms. The implication is that students in 2002 got about the same (or even fewer) chances of guessing randomly. We can be reasonably satisfied that this aspect of the test has not become a problem within the context of this study. So it seems that using four- or five-option items to reduce the probability that low ability students may answer right solely by guessing is too costly a test construction approach, especially since some guessing is expected and even encouraged in reading comprehension tests.

Conclusions

Our study has corroborated to a certain degree the effectiveness of the three-option format and, in so doing, has addressed several authors' concerns about this highly attractive format.

If we were to follow the indication given by the literature, the number of items in departmental exams could be increased. So, for example, instead of 25 items, each exam could have 30 items of the type described in this study. This would allow us to better sample the students' abilities, which would be perceived by both teachers and students as a more valid measure of their achievement. Score reliability from such three-option multiple-choice tests would probably be higher than that obtained from our traditional departmental exams. If, on the

other hand, the same number of items continue to be used, it follows that administration time should be reduced proportionately.

Regarding potential drawbacks, our study focused on global exam indexes such as average difficulty and average discrimination. Doing so may blur our perception of the effect that the conversion had at the item level. A desirable course of action would be to look more deeply into our item analysis outcome in order to better characterize the quality of the items that result from the conversion procedure. This would eventually lead us to gauge the quality of the exams in question as a measure of EST reading comprehension. Additionally, evaluating the conversion procedure would also allow us to estimate the feasibility of full Item Bank conversion as opposed to writing three-option items from scratch.

We are currently looking into other item analysis tools to examine the adequacy of our decisions as suggested by colleagues who read previous versions of this paper. For example, Haladyna (1994) cites two tools that we did not take into account in the design of this study: the frequency table, which is a tabular presentation of how options were chosen by ordinal score groups, and trace lines, which is nothing but a graphical representation of the frequency table. According to the literature, these tools allow a more meaningful and interpretable understanding of the behavior of item performance as a function of total test performance. Recently, we learned that Lertap′s latest version is equipped with both tools. Looking at our data using these tools may help corroborate the decisions we made regarding item modification.

Although not precisely a drawback, we must say that our modification procedure clearly went beyond the procedure we initially set out to apply, i.e., merely eliminating the worst distractor as in Williams and Ebel′s 1957 study. We went beyond their procedure when we reviewed each item′s text, stem, and remaining options for adjustments. Looking back on our decision, it does not seem pedagogically inappropriate, although methodologically it implies our study is not properly a replication.

Our study focused on the general quality of three-option versus four-option multiple-choice tests. It would be useful to carry out a detailed item-by-item comparison in order to improve the modification procedure we used and, beyond that, to identify what features make a three-option reading comprehension item effective.

References

Alderson, J.C. (2000). Assessing reading. Cambridge, UK: Cambridge University Press. [Links]

Alderson, J.C., Clapham, C., & Wall, D. (1995). Language test construction and evaluation. Cambridge University Press. [Links]

Archibald, D., Marin, P., & Foubert, E. (1980). Working paper on testing. Unpublished manuscript, Departamento de Idiomas, Universidad Simón Bolívar, Caracas – Venezuela. [Links]

Ary, D., Jacobs, L.C., & Razavieh, A. (1979). Introduction to research in education. (2nd ed.). New York: Holt, Rinehart and Winston. [Links]

Bachman, L.F., & Palmer, A.S. (1996). Language testing in practice. Oxford, UK: Oxford University Press. [Links]

Bailey, K.M. (1998). Learning about language assessment: dilemmas, decisions, and directions. Pacific Grove, CA: Heinle & Heinle. [Links]

Berríos, G., & Iribarren, I.C. (1989). El efecto del orden de presentación en el diseño de exámenes. Perfiles: Revista de Educación, 19-20, 5-21. [Links]

Cohen, A.D. (1994). Assessing language ability in the classroom (2nd ed.). Boston, MA: Heinle & Heinle. [Links]

Davidson, F., & Lynch, B.K. (2002). Testcraft: a teacher's guide to writing and using language test specifications. New Haven, CT: Yale University Press. [Links]

Douglas, D. (2000). Assessing languages for specific purposes. Cambridge, UK: Cambridge University Press. [Links]

Dycos, D. (1997). Guessing word meaning from context: should we encourage it? Literacy across cultures [A publication of the Foreign Language Literacy N-SIG of JALT], 1/2. Retrieved June 3, 2005, from http://www2.aasa.ac.jp/~dcdycus/ LAC97/guessing.htm [Links]

Goodman, K. (1967). Reading: a psycholinguistic guessing game. Journal of the Reading Specialist, 6 (1), 126-135. [Links]

Haladyna, T.M., & Downing, S.M. (1993). How many options is enough for a multiple-choice test item? Educational and Psychological Measurement, 53, 999-1010. [Links]

Haladyna, T.M. (1994). Developing and validating multiple-choice test items. Hillsdale, NJ: Lawrence Erlbaum Associates. [Links]

Language Testing List (LTEST-L). (2001). [On-line discussion group]. Retrieved November 19, 2003, from http://lists.psu.edu/cgi-bin/wa?A0=ltest-l. [Links]

Llinares de Alfonzo, G., & Berríos Escalante, G. (1990). Writing MCIs for reading tests in science and technology. English Teaching Forum, 28(4), 43-45. [Links]

Mikulecky, B. S. (1989). A short course in teaching reading skills. Reading, MA: Addison-Wesley Publishing Company. [Links]

Nelson, L.R. (2005). Lertap 5.4.6. [Computer program for classical item analysis; available for trial from http://www.lertap.com/]. Perth, Australia: Curtin University of Technology. [Links]

Nelson, L.R. (2000). Item analysis for tests and surveys using Lertap 5. [Statistics package manual]. Perth, Australia: Faculty of Education, Curtin University of Technology. [Links]

Owen, S.V., & Froman, R.D. (1987). What's wrong with three-option multiple-choice items? Educational and Psychological Measurement, 47, 513-522. [Links]

Read, J. (2000). Assessing vocabulary. Cambridge, UK: Cambridge University Press. [Links]

St. Louis, R. (2001). What is reading? Unpublished manuscript, Departamento de Idiomas, Universidad Simón Bolívar, Caracas – Venezuela. [Links]

Trevisan, M.S., Sax, G., & Michael, W.B. (1991). The effects of the number of options per item and student ability on test validity and reliability. Educational and Psychological Measurement, 51, 829-837.

LOS AUTORES

Gilberto Berríos

M.A. in Applied Linguistics (Concordia University, Montreal, 1994). Trabaja en el Departamento de Idiomas la Universidad Simón Bolívar desde 1981. Jefe del Departamento de Idiomas y co-investigador del proyecto de investigación de "Rediseño del Banco de Ítemes para el Departamento de Idiomas de la USB". **Línea de Investigación:** Evaluación del aprendizaje de lenguas extranjeras gberrios@usb.ve

Carlina Rojas

M.A. in Hispanic Studies (University College London, 1999). Trabaja en el Departamento de Idiomas la Universidad Simón Bolívar desde el 2000. Jefe de la Comisión de Exámenes del Departamento de Idiomas y es co-investigadora del proyecto de investigación de "Rediseño del Banco de Ítemes para el Departamento de Idiomas de la USB". **Línea de Investigación:** Evaluación del aprendizaje de lenguas extranjeras y la literatura latinoamericana carojas@usb.ve

Noela Cartaya

Magíster en Lingüística Aplicada de la Universidad Simón Bolívar (2005). Trabaja en el Departamento de Idiomas la Universidad Simón Bolívar desde el 2001. **Línea de Investigación:** Comprensión de lectura y la adquisición de vocabulario en L2. ncartaya@usb.ve

Yris Casart

Magíster en Lingüística Aplicada en la Universidad Simón Bolívar (2004). Trabaja en el Departamento de Idiomas la Universidad Simón Bolívar desde el 2002. Miembro de la Comisión de Exámenes del Departamento de Idiomas y responsable del proyecto de investigación de "Rediseño del Banco de Ítemes para el Departamento de Idiomas de la USB". **Línea de Investigación:** Evaluación del aprendizaje de lenguas extranjeras y la adquisición de lengua materna. ycasart@usb.ve

Datos de la Edición Original Impresa

Berríos, G. Rojas, C. Cartaya, N y Casart Y (2005, Junio). Effect of the number of options on the quality of est reading comprehension multiple-choice exams. *Paradigma*, Vol. XXVI, Nº 1, Junio de 2005 / 89-116.