Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 173 (2015) 5 – 10

Procedia
Social and Behavioral Sciences

32nd International Conference of the Spanish Association of Applied Linguistics (AESLA): Language Industries and Social Change

Identifying learning patterns in the upper-intermediate level of English through large-scale testing

Irina Argüelles-Alvarez[a]*, Margarita Martinez-Nuñez[b]

[a]*Dep. of Linguistics Applied to Science and Technology, Universidad Politécnica de Madrid, Crta. de Valencia, km 7, Madrid 28031, Spain*
[b]*Dep. of Bussiness Organization, Management and Statistics, Universidad Politécnica de Madrid, Crta. de Valencia, km 7, Madrid 28031, Spain*

Abstract

In this paper we will summarize the rationale and validation process of a multiple choice test developed at the Universidad Politécnica de Madrid (UPM) to regulate the students' access to the subject "English for Professional and Academic Communication" for which a B2 proficiency level, in accordance with the Common European Framework of Reference for Languages (CEFRL), was established as a minimum level. Item difficulty and item discrimination are studied and analyzed from the large-scale application of the test to 924 students. The aim of the study is to reach preliminary conclusions about possible areas where sequential learning on the part of students could be studied.

Keywords: Large-scale testing; learning patterns; item response models; teaching and assessment processes

1. Introduction

After having extensively documented the rationale, layout, description and validation process of a multiple choice test developed at the Universidad Politécnica de Madrid (UPM) to regulate the students' access to the subject "English for Professional and Academic Communication" (Argüelles Álvarez et al., 2011; Argüelles Álvarez & Pablo-Lerchundi, 2012; Argüelles Álvarez, 2013), it is probably time now to move further in the analysis of the

* Corresponding author. Tel.: +34-91-336-5229
E-mail address: irina@etsist.upm.es

doi:10.1016/j.sbspro.2015.02.022

results obtained from the large-scale application of the test to 924 students across University. In Argüelles Álvarez (2013), test item difficulty was analyzed applying qualitative techniques to reach preliminary conclusions about possible areas where sequential learning on the part of students could be studied. This key idea that learners progress through an order when acquiring grammatical structures, is supported by convincing evidence (Ortega, 2011) and represents one major finding of Second Language Acquisition (SLA) empirical research.

In what follows, we will first revise the initial conclusions we reached in the past (Argüelles Álvarez, 2013) with respect to item difficulty, as these could eventually be seen as evidence of existing learning patterns. Then, we will present further item analysis framed in item response theory (IRT) (Bachman, 1990, pp. 202-208) that apply item response models in order to make predictions about individual's performance on specific items. This further quantitative study of items gives us additional clues about the discriminability index of items and therefore, how they relate with one another. At the same time, we also aim at determine future lines for change in the original proficiency test. Besides the theoretical study, we will try to establish a connection between test results and eventual patterns in the students' learning process.

2. Test result

2.1. Test reliability

Although test reliability was already studied at the pilot stage (Arguelles Álvarez & Pablo-Lerchundi, 2012) recent results obtained at the large-scale application of the test across university are summarized in Table 1 with a result of: α Cronbach = 0.918.

Table 1. Test reliability

Cronbach Alfa	N elements
.918	924

This reliability coefficient in the range of 0 and 1, estimates the extent to which test takers would have obtained similar results in comparable parallel tests (Morales, 2012). As for the test validity it was extensively studied, analyzed and justified in Argüelles Álvarez (2013).

2.2. Item difficulty

Although for multiple-choice tests, the average item difficulty index is set higher to compensate possible guessing strategies, standardised tests aim at a range of 30% to 70% spread of difficulty, averaging out at approximately 50% (Davies et al., 1999, pp. 95-96). The degree of difficulty of a test item, calculated on the basis of a group test performance, can eventually lead us to conclusions about the degree of difficulty of the trait under test and items that are too easy (with an index close to 100%) or too difficult (with an index close to 0%) do not usually contribute to a test's discriminability. The items in our test averaged 51.30% as described in Argüelles Álvarez (2013).

2.3. Item discriminability

Item discrimination is a crucial feature to consider in criterion-referenced testing as here, discrimination implies the test's capacity to distinguish between masters and non-masters on the trait that the test is aimed to measure. Several statistical techniques can be used to calculate item discrimination. According to Morales (2012), the formula 1 below, for example, aims to calculate item discrimination as follows:

$$DI = \frac{CAUR - CALR}{Ngs}$$

Fig.1: Formula

Where:

- DI=item discrimination index;
- CAUR=number of correct answers in the upper range;
- CALR=number of correct answers in the lower range;
- Ngs=group size

And state the discrimination intervals below:

- >0.61 Very High
- 0.41 to 0.60 High
- 0.31 to 0.40 Average
- < 0.30 Low

In calculating the discrimination index (DI), first each student's test are scored and ordered. Next, the 27% of the students at the top and the 27% at the bottom are separated for the analysis. According to Wiersma & Jurs (1990, pp. 145), "27% is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analysis".

The discrimination index is therefore the number of students in the masters (higher) group who answered the item correctly minus the number of students in the non-masters (lower) group who answered the item correctly, divided by the number of students in the largest group. Wood (1960) stated that when more students in the lower group than in the upper group select the right answer to an item, the item is showing negative discrimination or negative validity. Therefore, if we assume that the criterion itself has validity, the item is not only useless but is actually decreasing the validity of the test and consequently, should be discarded.

Classical analysis has traditionally calculated item discrimination by means of correlation techniques such as the usual "Pearson product-moment correlation coefficients". The latter, take advantage of the fact that individual item scores can be only 0 or 1 (Engelhart, 1965; Guilford & Fruchter, 1978). The mean item-total correlation coefficient may be estimated from the mean and standard deviation of the total scores, both expressed as fractions of N (number of items) (Burton, 2001).

3. Discussion

In order to reach preliminary conclusions from the results obtained, we are mainly concerned with the classification easy/difficult and discriminability indexes of the discrete functional-grammar items that make up the first part of the test. As largely described in previous research (Argüelles Álvarez et al., 2011; Argüelles Álvarez & Pablo-Lerchundi, 2012; Argüelles Álvarez, 2013), the second part of the test, presents the stimulus material in the form of text with cloze-type tasks and reading comprehension questions that are not discussed herein as the decision was made from the beginning to study the results in the reading section apart.

3.1. Classification easy-difficult

Table 2 below, summarizes the classification easy-difficult presented in Argüelles Álvarez (2013):

Table 2. Classification easy-difficult of the discrete items in the test

	Correct answers <30%	Correct answers >70%
Number of items under this category	21	22

Among the grammar, functions or notions addressed in the 22 items that were answered correctly on the part of the test takers in the range of >70%, and therefore, classified in Argüelles Álvarez (2013) as "easy", many address temporal and aspectual meanings (anaphoric time, duration or frequency). Furthermore, notions related to time and temporality such as grammatical tense and aspect have shown to be correctly interpreted by the selection of the correct option among the four possible given: [...] already*[...], [...] during*[...], [...] usually get up*[...], [...] have ever known*[...]. Others worth mentioning are grammatical knowledge and use of subordinating conjunctions as in "[...] unless* you press the bell" (item 10). Finally, with regard to modal verbs, those indicating "impossibility" are classified in this range of >70%, while "certainty" falls in the category from 30% to 70%. 76,53% students answered correctly to item number 32 whereas only 58,91% selected the correct option in the case of item number 33. See examples 1 and 2 below:

Example 1: You _____ go wrong if you follow the instructions. Impossible. Options: a) might, b) must, c) could, d) can't*
Example 2: He _____ have taken the money. Certain. Options: a) may, b) must*, c) could, d) can't

Among the grammar, functions or notions addressed in the 21 items that were classified as "difficult" (<30%), the following can be highlighted: grammatical form and meaning of few, a few, little, a little; adjacency pair to assess grammatical form in the context of the adverb rather used as "more readily or willingly" or grammatical form of the Saxon genitives.

3.2. Discrimination Index

According to DI formula, there are not values over 0.60 in our sample and therefore, our classification is summarized as shown in Table 3:

Table 3. Discrimination Index (DI) classification

Discrimination Index (DI)	Frequency	Average Index Value	Items Mean	Typical Deviation
> 41% (high)	10	47.15	55.00%	0.49
30-40% (average)	34	34.84	51.00%	0.47
<30% (low)	56	22.76	58.00%	0.44

According to the item-total correlation indexes calculated by means of SPSS, the range over 35% has been considered as high, between 30-35%, average and under 35% the range has been regarded as low as shows Table 4.

Table 4. Discrimination Coefficient item-total

Discrimination Coefficient	Frequency	Coefficient Mean	Mean	Typical Deviation
> 35% (high)	22	40.23	44.00%	0.46
30-35% (average)	31	32.3	56.00%	0.46
<30% (low)	47	24.58	61.00%	0.45

From the intersection of both classifications the items that are definitely influencing test discrimination are shown in Table 5 below:

Table 5: Item discriminability (own classification)

	N (Frequency)	Item number
G1 (High discriminability)	16	10, 15, 19, 22, 26, 30, 43, 49, 52, 55, 64, 68, 76, 82, 83, 100.
G2 (Low discriminability)	15	2, 17, 21, 23, 24, 40, 47, 59, 60, 61, 65, 69, 71, 74, 85.

As is the case with item classification easy-difficult, the second part of the test (reading-comprehension) is not discussed herein. Therefore, our analysis is based on the discrete items from 1 to 67 with the aim to address complete results in further investigation.

It has been largely repeated that items that are too easy or too difficult should be removed because they do not contribute to test discriminability (Davies et al., 1999). From our data, item number 26 falls in the category of "high discriminability" although regarding its difficulty, the item was answered correctly on the part of the test takers in the range of <30% (difficult). Designed to test grammatical form and meaning (cohesive-ellipsis), item 26 seeks the correct function to express "in a similar manner or way" where students fail to identify so as the correct answer when the sentence provided as input is positive (Example 3). On the contrary, 62.40% answers are correct when the sentence given as input is negative (Example 4):

> Example 3: My father works at home and ____ does my mother. Options: a) so*, b) neither, c) either, d) same.
> Example 4: I haven't tried speed dating and ____ have my friends. Options: a) so, b) neither*, c) either, d) same.

On the opposite side, item number 10, already mentioned above, falls in the category of "high discriminability" although the item was answered correctly on the part of the test takers in the range of >70% (easy).

4. Conclusion

From our preliminary study, it can be firstly concluded that it can be actually demonstrable that learners progress through an order when acquiring grammatical structures. This order could be inferred from their answers to a multiple choice test where "implicit knowledge" (intuitive and rapidly processed) must be demonstrated.

Secondly, although according to Alderson & Wall (1993), there is little evidence for the claims made about the positive or negative impact of language testing, the effect of testing on instruction has had clear negative consequences in our context. For the last year, a negative backwash effect has been observed both within the institution and outside it, which has moved students to attend heavily grammar-based exam preparatory courses. This tendency must be reverted, which necessarily implies the need to design and validate new proficiency tests.

The item discrimination study presented herein represents a final stage in the development of a B2 proficiency test at the same time that it provides us with the necessary information to start the process again. The departing point this time will be adapting the items that have demonstrated to be clearly discriminating proficiency among test takers as part of a more comprehensive and adapted to the new context proficiency test.

References

Alderson, J. C. & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14 (2), 115-29.

Argüelles Álvarez, I., Pablo-Lerchundi, I., Herradón Díez, R. & Baños Expósito, J.M. (2011). Large-scale Testing of Proficiency in English: Back to Multiple Choice? *Proceedings of the BAAL Conference*, University of the West of England, 13-16.

Argüelles Álvarez, I. & Pablo-Lerchundi, I. (2012). …And back to multiple choice! Large-scale testing of proficiency in English: an experience. *ODISEA, Revista de Estudios Ingleses*, 13, 9-18.

Argüelles Álvarez, I. (2013). Large-scale assessment of language proficiency: Theoretical and pedagogical reflections on the use of multiple-choice tests. In L. Cerezo & M. Amengual (Eds.) Second Language Testing: Interfaces between Pedagogy and Assessment. *International Journal of English Studies*, 13 (2), 21-38.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Burton, R. F. (2001). Do item-discrimination indices really help us to improve our tests? *Assessment and Evaluation in Higher Education*, 20 (3), 213-220

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Doughty, C. & Williams, J. (Eds.) (1998). *Focus on form in classroom second language acquisition*. Cambridge: Cambridge University Press.

Engelhart, M. D. (1965). A comparison of several item discrimination indices. *Journal of Educational Measurement*, 2(1), 69–76.

Guilford, J. P. & Fruchter, B. (1978). *Fundamental Statistics in Psychology and Education*, 6th ed. New York: McGraw-Hill Book, Co.

Morales, P. (2012). Análisis de ítems en las pruebas objetivas. Facultad de Ciencias Humanas y Sociales. Universidad Pontificia Comillas. http://www.upcomillas.es/personal/peter/otrosdocumentos/AnalisisItemsPruebasObjetivas.pdf

Ortega, L. (2011). Sequences and processes in language learning. In M. H. Long & C. J. Doughty (Eds.) *The handbook of language teaching*. Oxford: Blackwell Publishing.

SPSS for Windows. (2012). Version 21.0.0. Chicago: SPSS Inc. (software on CD-ROM). Available in SPSS Inc. Website: http://www.spss.com/

Wiersma, W. & Jurs, S.G. (1990). *Educational measurement and testing* (2nd ed.). Boston, MA: Allyn and Bacon.

Wood, D.A. (1960). *Test construction: Development and interpretation of achievement tests*. Columbus, OH: Charles E. Merrill Books, Inc.