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.

Peer-review under responsibility of Universidad Pablo de Olavide.

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

Peer-review under responsibility of Universidad Pablo de Olavide.
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