Replicating Experts' Based Item Calibrations

Rosa Arruabarrena, Javier López-Cuadrado, Tomás A. Pérez, Anaje Armendariz, José Á. Vadillo
E-mail: rosa.arruabarrena@ehu.es
University of The Basque Country (Spain)

Abstract

The aim of the developed work has been to identify subprocesses and decisions that are embedded in experts' based item calibrations, in order to replicate or improve them.

The number of Learning Management Systems (LMS) is steadily increasing, as well as those based in calibrated items. In particular, the last ones can also be used to determine the learning progress of students, to assign new students to their appropriate groups and to test proficiency. It is known that calibration processes can be divided into two stages, since firstly data has to be gathered and then the sample has to be analyzed and calibrated. Unfortunately, in most of the related bibliography, authors focus on the virtues of those systems, hardly mentioning anything about how calibrated items have been generated. Moreover, usually, the calibration of items is done by experts; though their performance is rarely mentioned. Accordingly, the involved subprocesses are not identified and they can neither be replicated nor improved.

To identify underlying processes, a controlled experiment was carried out with a bank with 252 textual items used to learn Basque language. Two groups of experts took part into the data gathering stage, with 74 and 42 participants each one [1]. However, due to the size of the experiment, this paper only discusses the second stage of the calibration. Thus, to perform the data analysis and calibration stage, firstly data filtering criteria were defined and applied, removing items from the bank. Then, two estimators were settled down to estimate the values of the parameters item difficulty and grammatical skill. Later, from the filtered sample a calibration was computed with the values of the kappa coefficient 0.675 and 0.763 for the previously mentioned parameters. Finally, the developed processes were validated by the performed analysis on both subsamples in terms of confidence intervals and percentages.

It has been worthy to develop the experiment from the scratch, to identify the underlying processes and their time dependencies precisely as well as to foresee key decisions makings. As a result of our controlled experiment, developers who seek to establish off-line the most likely values among the consensus forecasts of parameters may replicate the developed calibration or draw inspiration from it.