

Replicating Expert-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 (UPV/EHU), (Spain)

Abstract

Learning Management Systems (LMS) with calibrated items are a very important part of the kernel of e-learning systems. They can use those items either to determine the learning progress of the students, to assign new students to their appropriate skill level, or to test student proficiency.

The aim of this paper is to identify tasks and decisions during an expert-based item calibration, in order to replicate or improve that process. Unfortunately, in most of the related bibliography, authors focus on the virtues of their systems, hardly mentioning how they did the calibration. Moreover, the calibration of items is usually done by experts; though their performance is rarely discussed. Accordingly, the involved tasks are not identified, which makes more difficult the replication or improvement of this process.

To identify underlying processes, a controlled experiment has been carried out with a bank of 252 textual items used in a Basque language LMS. It has been worthy to develop the experiment from scratch, to identify the underlying processes and their time dependencies precisely, as well as to foresee key decision makings. The calibration process can be divided into two stages: sample data gathering, and data analysis. Two groups of experts took part in the data gathering stage, with 74 and 42 participants respectively. However, due to the size of the experiment, this paper will focus on the second stage of the calibration: It describes how data filtering criteria were defined and applied. Then, two trait estimators were created: item difficulty and grammatical skill. Later, the calibration was computed and tested with the values of the Kappa coefficient. Finally, the developed processes were validated by the performed analysis on both subsamples in terms of confidence intervals and percentages.

As a result of this experiment, other developers may replicate the developed tasks or draw inspiration from it to establish off-line the most likely values among the consensus forecasts of traits.

1. Introduction

Hezinet [1] is a pioneering commercial Adaptive Hypermedia System for Basque language learning for adults. Currently, it is being used in Spain and other 14 countries. Its didactic module contains more than 10,000 calibrated items that are characterized by two main traits: their *difficulty level* and the *grammatical learning skill* they develop. Up to now Basque linguists and pedagogues have estimated both values, though their performance has not been documented. Moreover, there is another way of calibrating the items, which is based on the Item Response Theory [2].

During the item calibration, there are two main tasks to be accomplished: *Data Gathering*, and *Data Analysis and Calibration*. A controlled experiment has been carried out with a bank of 252 items with the aim of identifying their underlying processes. The development of the data gathering consisted on two field tests (FT1 and FT2). In each field test, paper questionnaires were administered to linguists and pedagogues (experts). Each questionnaire contained 42 items and experts were required to

provide three pieces of information per item: (1) the correct answer of the item and their estimates for (2) its difficulty and (3) its learning skill. In FT1, 74 experts took part and a total of 3119 items were administered, obtaining a sample of 10 judgements per item. In FT2, only 42 experts collaborated on a total of 1768 items, and with a rate of 7 judgements per item. Therefore, at the end of the former stage there was a sample with 4887 contributions (three values per item) from 116 experts. This paper will focus on the latter stage, since the former stage is fully described in [3].

The paper is organized as follows: section 2 describes how the sample was filtered to develop later a two-stage calibration. Section 3 presents the process followed to estimate the item difficulties. Section 4 does the same to obtain the items' grammatical skill estimations. The fifth section is dedicated to determining whether there was or not any difference between judgements of FT1 and FT2 experts. Finally, section 6 draws some conclusions and discusses future work.

2. Data filtering

Following some recommendations from [4, 5] and considering the targeted two-stage calibration, some criteria have been defined to filter unreliable items and to preserve the quality of expert contributions:

- **C. Ex2.** An expert contribution is considered valid if it only indicates one difficulty level.
- **C. It1.** An item is not discarded, if at least 50% of the experts answer it correctly.
- **C. It2.** An item is not discarded, if at least 75% of the difficulty judgments given by the experts are grouped in 4 consecutive difficulty levels (of a total of 12 discrete values).
- **C. Ex1.** A questionnaire completed by an expert is discarded, if more than 25% of the answers given are wrong.
- **C. Ex3:** Any contribution not having one and only one skill value selected will be removed.

These criteria were used to build three data sieves that were consecutively applied. Several simulations were made to determine the best way of grouping them as well as their order of application [6].

The **first sieve** applies criterion C.Ex2 and prevents those expert contributions that either do not provide difficulty estimation, or provide more than one, from being taken into consideration. That kind of contributions would be worthless to estimate the difficulty. Once an expert is considered unreliable, every contribution given by him/her will be removed from the sample, discarding the whole questionnaire.

The **second sieve** applies first C.It1, which purges unreliable items (those with a success rate lower than 50%; the authors significantly reduced the requisite to 50% taking into account that almost 9% of the items had been labeled as "likely erroneous" during the pilot testing). Then, C.it2 discards items for which there is no consensus possibility, since their difficulty distribution is much dispersed. And, finally, C.Ex1 eliminates those expert contributions that have not been rigorously filled up. After having applied the three filters successively twice, the results became stable.

The **third sieve** uses criterion C.Ex3 to discard contributions that are unsuitable to estimate the item grammar skill.

As a consequence of the filtering stage, 60 items (17,8%) were removed from the bank, as well as 5 complete questionnaires (4,3%).

3. Difficulty estimation

Neither the mode nor the *arithmetic mean* statistics could be used to estimate the difficulty since the former does not make differences between items of the same level, showing the necessity of a

continuous estimator. And, although the latter computes continuous values, outliers could alter the estimation. Therefore, a specific statistic (the *M.dif* estimator) was fixed to estimate the difficulty. It is, inspired by *the Delphi method* [7] and the *algorithm of maximum consensus* [8], but without physically joining several experts. The goal of *M.dif* is to establish off-line the most probable value among the most agreed estimations of difficulty.

M.dif estimator was defined by means of the following two rules:

- **M.dif1.** The difficulty of an item is computed as the mean of the relative frequencies of the values grouped in the interval of 4 levels (a third of the scale) with greater density of values. Since the sample has already been filtered, the interval should contain at least 75% of the contributions.
- **M.dif2.** If there were more than one interval meeting condition M.dif1, then the interval will be extended to 5 consecutive difficulty levels, and the interval with smaller deviation will be chosen.

The difficulty for the remaining 192 items was computed by applying the *M.dif* statistic to the sample obtained from the second sieve. Figure 1 shows the values obtained. The horizontal axis represents each of the 192 items and the vertical axis shows the estimated difficulty varying from 1 to 12. The particular estimated values and their deviations can be checked in [6]. The results suggest that the distribution of the difficulties of the items was unbalanced. Concretely, half of the bank has an intermediate difficulty level (between 4 and 8), and practically the other half has a low-middle difficulty. Thus, the number of items with high difficulty is small. The value of Kappa index was 0.675, which means that the inter-rate reliability of experts in their difficulty estimations was good [9].

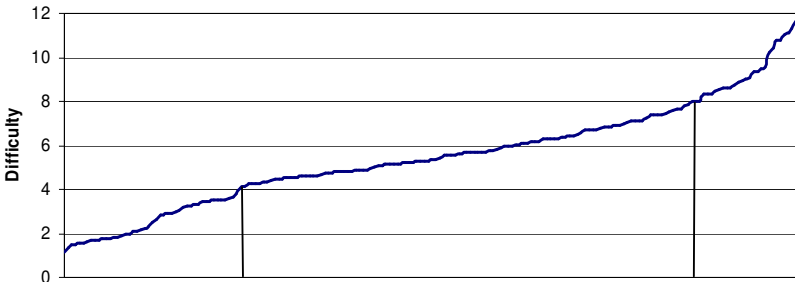


Fig. 1. Difficulty of items estimated by M.dif

4 Grammar skill estimation

The *majority element* and the *mode* statistics were firstly studied in order to establish the grammatical skill. Both options were ruled out since the former could not fix the skill of 10 items (6%) that showed nominal frequencies smaller than 51%. Something similar happened with the latter in the case of 5 items (3%) that had multiple modes.

To overcome this limitation, the *M.est* estimator was established in the following way:

- **M.est.1.** The grammar skill of an item is the mode of the estimations given by experts.
- **M.est.2.** When there are multiple modes, the least frequent grammar skill among the items fixed until that moment will be chosen.

Figure 2 shows the resulting distribution of the skill estimations computed by applying M.est to the sample obtained after the third sieve. See [6] for more details.

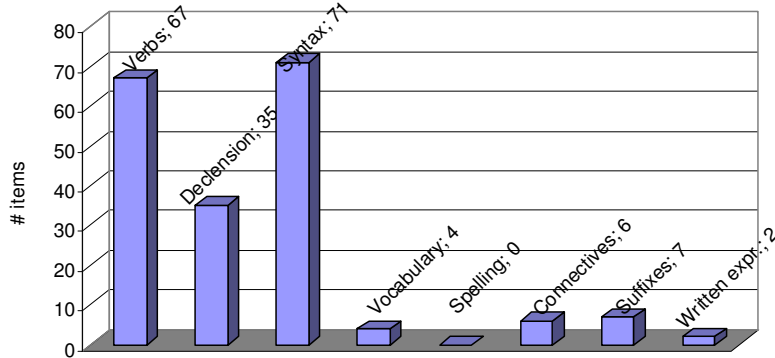


Fig. 2. Distribution of skills computed by M.est

One third of the items assesses *syntax*, another third is devoted to verbs, a sixth to declension and the remaining items to *vocabulary*, *connective*, *suffixes* or *written expression*. The agreement among the experts when judging items' grammar skill was denoted by a Kappa index value of 0.763, which can be considered very good [9]. Figure 3 shows the 192 skill estimations grouped into three singular sections according to the degrees of agreement. Concretely, there was complete agreement for 35% of the items (67/192); for 60% of the items (115/192) the considered skill turned out to be by majority agreement; and for the remaining 5% (10/192) it was estimated by minority agreement.

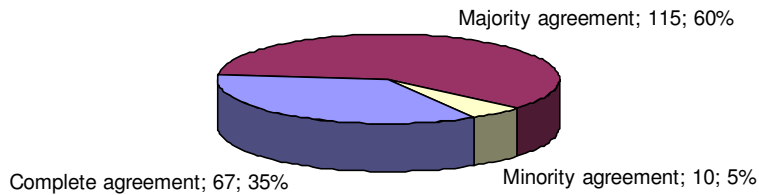


Fig. 3. Degrees in agreement among the experts in the skill estimations

5 Differential analysis

Although the trait estimation of the 192 items was computed using the judgements given by all the experts jointly, and independently of how or when the questionnaires were administrated, the authors consider that it is important to study if there was or not any difference between judgements of FT1 and FT2 experts.

The impact of the sieves over the whole sample and both subsamples in four particular moments are compared in figure 4. Specifically, the variability of the proportions of the volumes of the subsamples on the total sample was smaller than 1%.

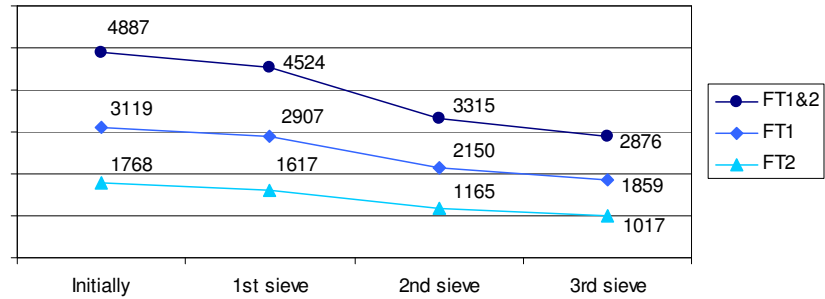


Fig. 4. Evolution of the number of judgements when filtering samples FT1, FT2 and both together

In order to analyze the consensus in the estimated difficulties, the item difficulties were estimated again by applying M.dif to the data sample gathered in FT1 and in FT2, as well as their 95% confidence intervals (CCII). Finally, a paired comparison was done to explore the overlapping between the calculated CCII. Concretely, the overlapping happened in 98% of the pairs of crossed intervals, reaching the 100% of the pairs of intervals when increasing the confidence interval to 99%. Since the range of the difficulty employed was [1-12], it can be concluded that there was no predominance by experts of either FT1 or FT2 in the values of the joint estimations.

The influence of each subsample over the total was also studied for skill estimation in percentage terms, and similar results were obtained [6].

6. Conclusions and future work

An off-line item calibration based on judgements from multiple experts has been developed in the context of Basque language learning. The considered traits have been the item difficulty and the learning skill. For that purpose, it has been necessary to create and apply some specific estimators, which seek to decide the most likely values among the consensus forecasts off-line. The degrees of agreement among experts have been respectively good and very good for both traits. Nevertheless, the study has shown that the initial item bank is biased and that it does not cover the whole range of neither difficulty levels nor learning skills.

The input sample was the result of joining FT1 and FT2 subsamples. Comparing the subsamples and the joint sample, the authors have been able to conclude the following:

1. the sieves have removed nearly the same proportions of contributions in the three studied samples,
2. there has not been predominance in estimations by experts of any subsamples on the values of the joint estimations,
3. hence, a minimum of 7 judgments per item is enough to develop a similar calibration. Notice that this number should result from the filtering phase, where some judgements could be invalidated. Thus it would be recommendable to gather a greater number of judgements, depending on the abandon-rate,
4. the inter-rate reliability has to be verified.

Based on the empirical developments, a proposal of business process has been elaborated to develop item calibrations using expert judgments [10]. The proposal will allow to undertake future improvements and it will set the basis to create an expert system that will help during the calibration tasks.



References

- [1] Sanz-Lumbier, S., et al. Hezinet. *The Hypermedia System That Makes The Basque Language Easy To Learn*. in *IASTED: 5th International Conference on Computers and Advanced Technology in Education*. 2002. Cancún (Mexico): ACTA Press. pp. 344-349.
- [2] Hambleton, R.K., H. Swaminathan, and H.J. Rogers, *Fundamentals of Item Response Theory. Measurement methods for the social sciences*. Vol. 2. 1991, Newbury Park, CA: Sage.
- [3] Arruabarrena, R. and J. López-Cuadrado. *Issues to be taken into account when calibrating items*. in *Current Developments in Technology-Assisted Education*. 2006. Sevilla (España): Formatex Research Center-Badajoz. pp. 906-910.
- [4] Renom, J. and E. Doval, *Tests adaptativos informatizados: estructura y desarrollo*, in *Tests informatizados: fundamentos y aplicaciones*, J. Olea, V. Ponsoda, and G. Prieto, Editors. 1999, Ediciones Pirámide: Madrid (España). pp. 127-162.
- [5] López-Cuadrado, J., *Evaluación mediante test adaptativos informatizados en el contexto de un sistema adaptativo para el aprendizaje de la lengua vasca*, in *Lenguajes y Sistemas Informáticos*. 2008, Univ. País Vasco/ Euskal Herriko Unibertsitates: San Sebastián.
- [6] Arruabarrena, R. and A.J. Armendariz, *Estimación de los parámetros de los ítems de un sistema de e-learning vía expertos*. 2008, University of the Basque Country: San Sebastián.
- [7] Landeta, J., *Current Validity of the Delphi method in social sciences*. *Technological Forecasting and Social Change*, 2006. 73: pp. 467-482.
- [8] Chen, Y.-L. and L.-C. Cheng, *Mining maximum consensus sequences from group ranking data*. *European Journal of Operational Research*, 2009. 198(1): pp. 241-251.
- [9] Fleiss, J.L., *Statistical Methods for Rates and Proportions*, 2nd edition. 1981, New York: John Wiley.
- [10] Arruabarrena, R. and T.A. Pérez, *Calibración de ítems con expertos: procesos BPM, ejecución, análisis y mejora. Una investigación empírica*. 2010, University of the Basque Country: San Sebastián.