

## Issues to be taken into account when calibrating items

**R. Arruabarrena Santos<sup>\*</sup>, J. López-Cuadrado**

Department of Computer Languages and Systems, University of the Basque Country (UPV-EHU), Apdo. 649, 20080 San Sebastián, Spain

The number of Learning Management Systems (LMS) is continuously growing, as well as these based on banks of calibrated items. To create the initial bank of knowledge of the application or to extend it in later versions, it is necessary to calibrate the items and for it, there are some choices. In this paper we show a checklist of issues to be taken into account when calibrating the items of a bank and we compare their costs using two different approaches: one considering estimations made by experts and another one considering estimation methods of the Item Response Theory (IRT). Moreover, both estimations are calculated for the traditional paper version and for a newer way using Information Technologies.

**Keywords:** calibration of items, estimation, costs, LMS

### 1. Educational systems based on calibrated item banks

The number of Learning Management Systems (LMS) based on banks of calibrated items is continuously growing. To create the initial bank of knowledge of the application or to extend it in later versions, it is necessary to calibrate the items. GHyM, our research group, has been the main designer of HEZINET, a Learning Management System (LMS) to learn Basque language off-line. BOGA is its new online version and it is used all over the world [1]. Some years ago, we obtained some sheets containing 252 textual items that could be used to increase the BOGA's bank, if they were calibrated by fixing their difficulty level in range of 1-12.

In 2003 we decided to improve those systems with the provided bank by including the capability of assessing the knowledge level of newcomer students [2]. We also determined that it was worthwhile to develop two parallel processes of calibration of items: one based on experts' estimations and the other based on Item Response Theory [3], and to conclude in which conditions would be better to use experts or a systematic procedure. Hence, we first distinguished that an off-line items' calibration of a LMS has two stages: the data gathering stage and the data analysis and calibration stage; then, we defined a methodology to compare not only the two elaborated calibrations but also their relative costs [4]. Both calibrations have recently finished, and this paper focus on their related costs, giving, at the same time, a list of checkpoints to be considered to calculate the costs of similar off-line calibrations. The comparison between the generated calibrations is deferred to another paper.

This paper is organized as follows: we start describing five checkpoints to be taken into account to calculate the cost of an off-line calibration of items. Then, we calculate the costs of our two calibrations in two different versions for each one. In the last section we compare the costs of the shown variants of calibration of items drawing some conclusions.

### 2. Checkpoints when calibrating items of a bank

Due to the fact that calibrations consume many resources, in [4] we presented a methodology that we have followed, during the data gathering and data analysis and calibration stages, in both experts based and statistical off-line items' calibrations. Basically, the plan consists in five tasks, once the project has been defined: (i) to set up active and passive participants, (ii) to establish the criteria to concrete the tools and materials to obtain information, (iii) to locate the passive participants, (iv) to conduct pilot tests if necessary and (v) to carry out the work. From these tasks and from those four instantiations of the meth-

---

<sup>\*</sup> rosa.arruabarrena@ehu.es, Teléfono: +34 943 01 50 43

odology, we have observed that there is a list of checkpoints to be taken into account when calibrating a bank of items to calculate and to control the expenses of the calibration. Next, we illustrate the identified checkpoints.

It is necessary (1) to carry out a detailed *plan and management* of the whole stage. This means to fix the calibration's parameters, criteria and tools to obtain the calibration's data for the data gathering stage; and also, for the data analysis and calibration stage, to concrete the procedures to filter the gathered data when necessary and to elaborate the calibration. It also includes specifying deliverables, scheduling tasks in time paying attention to requirements and real resources (human, costs and available time) and leaving time to face unforeseen events. Everything above mentioned will last during the whole life of the instantiation stage and punctual adjustments could happen. So, this checkpoint will not be punctual, requiring intermittent controls throughout the whole development of the stage. Moreover, the resultant costs will directly depend on the experience and rigour of the main managers and developers. In addition to this, these active participants will possibly need additional (2) *formation*, whose cost has to be considered. Finally, special care has to be taken with (3) *participants*; in particular, the different types and amounts of identified participants that have to be recruited, instructed and coordinated. When the number of participants is considerable, it may be helpful to register related information in an electronic support.

During (4) *implementation (or plan execution)* there are different points to check and to register. In the first stage, for instance, since the tool to gather information has been already fixed [5], paper and computerized questionnaires in our two calibrations, it is necessary to do the questionnaires and to conduct some pilot tests to validate them [6]. Afterwards, field tests have to be conducted until one obtains the desired amount of data. This means that all expenses from conduction have to be registered. However, in the second stage the calibration will be elaborated from the collected data which has been previously filtered. So, statistical software will be surely needed as well as time to interpret the statistical results.

Finally, not only the number of (5) *deliverables* but also the detail of them (including reports) rebound in the costs. However, it is necessary a minimum documentation reflecting the progression of the work, the elaborated products and their costs, and conclusions of the work.

In the following sections, attending to the described five checkpoints, we show the accumulated costs of our two calibration processes.

### 3. Costs of two processes of calibration of items

In [2] we fixed two parallel calibrations to improve BOGA: one using experts' opinions and the other using statistical methods based on IRT. Our aim was to determine the best calibration under certain conditions; consequently we planned each calibration and set up common parameters. In [4] we showed that an off-line calibration's process consists of a data gathering stage and a data analysis and calibration stage. The main features and results of the first stage for both calibrations are shown in [4,7-10]. The second stages are almost finished, being [11] a detailed report of the statistical calibration. The main developers of the experts' and of the statistical calibration are Rosa Arruabarrena Santos and Javier López-Cuadrado, respectively. Both are computer scientists and full lecturers at the University of the Basque Country.

Table 1 shows the related costs for checkpoints (1), (2) and (5) of each stage for both calibrations. The accumulated costs in each checkpoint are those explained in the former section. Obviously, as said before, these costs may vary depending on the number and capability of the participants (developers) as well as the rigour and detail of their work. The costs of the checkpoints (3) and (4) are explained in next sections.

#### 3.1 Empirical calibration with experts

According to [12], three or five experts are enough to revise didactic material. But, as the provided bank had 252 items, it was unviable that voluntary experts would calibrate the whole bank. So, items were divided into questionnaires which we estimated could be answered in 50 minutes by an expert. We con-

ducted two type of field test called FT1 and FT2 to obtain at least 7 assessments of the whole bank per each variant. In FT1, the questionnaires were delivered and fetched directly to-from experts' workplace, which are called euskaltegis<sup>1</sup>. In FT2, the questionnaires were sent and received by post. Due to the number of active participants (the main developer, her supervisor and 3 more punctual participants) and time restrictions, FT1 and FT2 were essentially conducted in sequence.

Mainly by phone, 42 experts from 12 euskaltegis took park in FT1, though we had agreed with 22% more experts. In average, we made 2,7 trips and 3,5 phone calls to euskaltegis. In FT2, a total of 80 experts from 40 euskaltegis agreed to help us during the field test, but at the end only 42 of them, from 20 euskaltegis, took part in that stage. In average, we made 2,1 phone calls. The time spent in those activities has also been accumulated in Table 1 in checkpoint (4), as well as time spent building two calibrations, one from data of each field test.

**Table 1** Cost estimations in hours for each stage of our off-line calibrations. The supraindex shows the number of whole bank's assessments.

|   | Experts' calibration |                  |                  | Statistical calibration |                   |                    |
|---|----------------------|------------------|------------------|-------------------------|-------------------|--------------------|
|   | FT1+2 <sup>17</sup>  | FT1 <sup>7</sup> | FT2 <sup>7</sup> | SP+NSP <sup>540</sup>   | SP <sup>500</sup> | NSP <sup>500</sup> |
| <b>Stage 1: Data gathering</b>                  |                      |                  |                  |                         |                   |                    |
| (1) Formation                                   | 128                  | 128              | 128              | 158                     | 158               | 158                |
| (2) Planning & management                       | 110                  | 105              | 95               | 98,5                    | 103,7             | 82,5               |
| (3) Participants: passives' time                | 112                  | 50               | 50               | 1333,3                  | 1038              | 1692               |
| (4) Implementation                              | 376                  | 268              | 248              | 571                     | 614               | 390,6              |
| (5) Deliverables                                | 174                  | 155              | 155              | 200                     | 200               | 200                |
| <b>Stage 2: Data analysis &amp; Calibration</b> |                      |                  |                  |                         |                   |                    |
| (1) Formation                                   | 140                  | 140              | 140              | 60                      | 60                | 60                 |
| (2) Planning & management                       | 35                   | 35               | 35               | 5                       | 5                 | 5                  |
| (3) Participants: passives' time                | -                    | -                | -                | -                       | -                 | -                  |
| (4) Implementation                              | 114,3                | 102              | 102              | 27                      | 27                | 27                 |
| (5) Deliverables                                | 100                  | 80               | 80               | 70                      | 70                | 70                 |
| Amounts of time costs                           | 1289                 | 1063             | 1033             | 2523                    | 2276              | 2285               |

This table gathers the total amounts of time costs for each calibration type and variant and for each checkpoint. The second column has accumulated the costs from the 17 full bank's assessments from which 10 were obtained through FT1 and the rest through FT2. Columns 3 and 4 only have the cost of 7 full bank's assessment corresponding to FT1 and FT2, respectively.

Apart from time costs, to build a calibration from experts' assessments requires phoning costs, paper copies of questionnaires, car availability and posting expenses (in FT1 and FT2, respectively)], and a computer with office software (similar to Word, Access, Excell, SPSS).

### 3.2 Analytical calibration based on IRT

Similarly, according to experts in psychometrical calibration [13], it is highly recommended to obtain at least 500 answers per item from a heterogeneous sample to elaborate a trustworthy calibration. Due to the length of the source bank, it was divided into 6 computerized questionnaires or subtests that could be completed in 20 minutes by a volunteer. A web application was designed and implemented to store the minimum needed of 3000 answers [14]. Once again, two type of field tests were conducted, which are called SP and NSP. In SP, test conductors went to learning centres and there they supervised laboratory sessions where students completed the tests. In NSP, anonymous participants answered the tests in non supervised sessions.

In SP, 2268 students' answers were validated and 52 rejected, that is, the 2,2% of supervised sessions [15] were rejected. Those sessions took place in 10 secondary schools and faculties and, in average, 10 phone calls per each centre were made. Moreover, in NSP 975 more sessions were validated, though a

<sup>1</sup> Euskaltegi is a certain type of homologate school to teach Basque

total of 1635 were finished, which means that the 40,4% of this type of sessions were rejected. To recruit volunteers, emails were sent to distribution lists and about 1400 more emails and 500 phone calls were made to confirm non supervised sessions. Five active participants took part in supervised sessions and 2 in NSP. The time spent by them, as well as time used for validating the sessions and elaborating the calibration has been added to Table 1 in checkpoint (4). The fourth column in Table 1 accumulates the costs of 540 full bank's assessments obtained through SP and NSP field tests. The last two columns have the estimations time costs for SP and NSP field tests to obtain 500 full bank's assessment.

Apart from time costs, in this case, there have been trips to the teaching centres in SP and phoning expenses, a personal computer, a web-server and software for both SP and NSP field tests (an office packet, SPSS, LISREL/PRELIS, XCALIBRE and Internet Information Server, respectively).

**Table 2** Total time costs considering or not the time spent by passive participants.

| Time spent by        | FT1+2 <sup>17</sup> | FT1 <sup>7</sup> | FT2 <sup>7</sup> | SP+NSP <sup>540</sup> | SP <sup>500</sup> | NSP <sup>500</sup> |
|----------------------|---------------------|------------------|------------------|-----------------------|-------------------|--------------------|
| Active participants  | 1177                | 1013             | 983              | 1190                  | 1238              | 993                |
| Passive participants | 112                 | 50               | 50               | 1333                  | 1038              | 1692               |
| In total             | 1289                | 1063             | 1033             | 2523                  | 2276              | 2685               |

Table 2 sums up the time spent in total all over the processes, only by active participants (developers) or by passive participants (all of them volunteers, having been experts and students).

#### 4. Conclusions and future work

The process of item calibration consumes many resources and there are different ways of doing it. Besides, we had a choice to improve a successful LMS to learn Basque by extending its bank, since we had been provided with a textual bank of 252 items. Having established a methodology to develop an off-line items' calibration process, we developed in parallel during 2 years two calibration processes: an empirical one using experts' opinions and an analytical one based on IRT. In relation to the mentioned working methodology, it has been helpful to set up a list of checkpoints to be considered to calculate and to control the costs of the calibration. This paper has presented this list, as well as the expenses of the developed calibrations for the checkpoints. Table 1 gathers time costs for each calibration type and variant and Table 2 summarizes time costs considering (or not) time spent by passive participants. From these two tables many conclusions can be drawn.

On the one hand, in the calibrations based on experts', the difference of costs between FT1 and FT2 are in checkpoint 2 and 4 in the data gathering stage, since in FT1 there is more control over the progression of the work, including trips and number of phone calls. Moreover, although the fact that in FT2 the rate of abandonment is closed to the 48% and more agreements have to be established (so more questionnaires have to be sent), FT2 variant continuous been cheaper to obtain the required 7 full banks' assessments. Another interesting alternative could be to implement FT1 and FT2 with computerized questionnaires, since costs associated to trips and copies would disappear.

On the other hand, when calibrating analytically, two factors have to be considered before drawing conclusions. Since a minimum of 500 answers per item are required in an IRT-based calibration, the number of volunteers to be recruited is large, as well as time spent by them answering computerized subtests. If we do not consider passives' time, the NSP variant is the cheapest. But this would mean that 5031 passive volunteers should take part in nonsupervised sessions by their own initiative in a fix period of time, since the rate of rejected answers is the 40,4% in NSP. If we consider passives' time, the SP is the cheapest. Nevertheless, the point is that appointments have to be made with schools' headmasters and active participants have to go to those schools to supervise subtest administrations. As a consequence, in SP almost all sessions are validated.

Depending on whether one considers or not time spent by passives in all the studied variants of calibration, expert calibrations are cheaper than IRT-based ones from time's cost point of view. When we do not consider passives' time, although FT2 and NSP are similar in spent time, the analytical calibrations have software expenses to compute the volume of data gathered and active participants need to have

good formation in statistics. So, in the considered conditions, FT2 is definitively the cheapest calibration variant among those studied.

In future work, we will consider new conditions, as other lengths of the source bank and repeating the process with new banks. Moreover, we would like to determine under which conditions it is not worthwhile the automation of the process. Apart from costs, comparisons between the generated results, i.e. between obtained item difficulties, are to be done.

To conclude, the lecturer should consider that the given resultant costs are estimations, since they may vary depending on the number and capability of the participants (developers) as well as the rigour and detail of their work. Another point is the period to conduct field test; it must always be fixed, certain moments are unproductive for volunteers and the increase in the number of active participants does not guarantee to get a proportional effect while gathering data or during work development.

**Acknowledgements** We would like to thank Anaje Armendáriz by her active participation in our projects and T.A. Pérez Fernández by his critical reviews of them.

## References

- [1] S. Sanz-Lumbier, J. Gutiérrez, T. A. Pérez, S. Sanz-Santamaría, J. A. Vadillo and M. Villamañe. Hezinet. The hypermedia system that makes the Basque language easy to learn. in IASTED: International Conference on Computers and Advanced Technology in Education. Cancún (Mexico): ACTA Press. (2002) pp. 344-349.
- [2] R. Arruabarrena, J. A. Vadillo and J. Gutiérrez. Are experts difficulty guessing and statistical results comparable? 2nd International Conference on Multimedia and Information & Communication Technologies in Education: m-ICTE2003. Badajoz. (2003) Vol. 1, pp. 542-545.
- [3] J. López-Cuadrado, T. A. Pérez, J. A. Vadillo and R. Arruabarrena. Integrating adaptive testing in an educational system. First International Conference on Educational Technology in Cultural Context: ETCC2002. Joensuu (Finland): University of Joensuu. (2002) Vol. 3, pp. 133-149.
- [4] R. Arruabarrena and T. A. Pérez. Pruebas de campo para calibrar un banco de ítems vía expertos. University of the Basque Country UPV/EHU/LSI/TR 08-2005. (2005) pp. 96
- [5] R. Arruabarrena, T. A. Pérez, J. Gutiérrez, J. López-Cuadrado and J. A. Vadillo. On evaluating adaptive systems for education. 2nd. International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems: AH2002. Málaga: Springer-Verlag. (2002) Vol. n. 2347, pp. 363-367.
- [6] M. Tessmer. Planning and conducting: formative evaluations. London: Kogan Page Limited. [1993], pp. 159.
- [7] R. Arruabarrena, S. Sanz-Santamaría and T. A. Pérez. Quality techniques help in LMS' improvement. Conf. Internacional de Tecnologías de la Información y Comunicación en la Educación: ICTE2005. Cáceres (Spain): FORMATEX (Badajoz). (2005) Vol. 1, pp. 542-545.
- [8] R. Arruabarrena. Filtrado de un banco de ítems. University of the Basque Country UPV/EHU/LSI/TR 02-2005. (2005) pp. 60.
- [9] J. López-Cuadrado and R. Arruabarrena. Diseño de anclaje de un banco de ítems. University of the Basque Country. UPV/EHU/LSI/TR 12-2005. (2005) pp. 109.
- [10] J. López-Cuadrado Administración de subtests de anclaje para calibrar un banco de ítems University of the Basque Country. UPV/EHU/LSI/TR 8-2006. (2006) pp. 96.
- [11] J. López-Cuadrado and A. J. Armendáriz. Obtención de estimaciones de los parámetros durante la calibración de un banco de ítems. University of the Basque Country. UPV/EHU/LSI/TR 13-2006. (2006) pp. 271.
- [12] B. Shneiderman. Designing the User Interface, 3rd edition. Reading-Massachusetts: Addison Wesley Longman, Inc. 639. [1998].
- [13] J. Olea, V. Ponsoda and G. Prieto. Tests informatizados: fundamentos y aplicaciones. Ediciones Pirámide. ed. Colección "Psicología". Madrid (Spain). [1999].
- [14] J. López-Cuadrado, A. J. Armendáriz and T. A. Pérez-Fernández. A supporting tool for the adaptive assessment of an e-learning system. Conf. Internacional de Tecnologías de la Información y Comunicación en la Educación: ICTE2005. Cáceres (Spain): FORMATEX (Badajoz). (2005) Vol. 1, pp. 295-299.
- [15] J. López-Cuadrado, A. J. Armendáriz and T. A. Pérez-Fernández. Should unsupervised web surveys be used to calibrate items? ePortfolio. Oxford (UK): European Institute for E-Learning (EIFEL). (2006) (cd).