Some practical issues in the design and operation of item bank calibrations. 
An experience in the context of Basque language assessment

Javier López-Cuadrado, Rosa Arruabarrena, Anaje Armendariz & Tomás A. Pérez

(pp. 104-115)

The Asian Conference on Language Learning
Official Conference Proceedings 2011

ISSN: 2186-4691
Proceedings URL: http://iafor.org/acll_proceedings.html

iafor

The International Academic Forum
www.iafor.org
Title:
Some practical issues in the design and operation of item bank calibrations. An experience in the context of Basque language assessment

Authors and affiliations:
Javier López-Cuadrado, Rosa Arruabarrena, Anaje Armendariz & Tomás A. Pérez
Department of Computer Languages and Systems
University of the Basque Country (UPV-EHU)

Contact information:
{ javilo, rosa.arruabarrena, anaje.armendariz, tomas.perez } @ehu.es

Topic of the submission:
Testing and evaluation
Title:
Some practical issues in the design and operation of item bank calibrations. An experience in the context of Basque language assessment

Authors and affiliations:
Javier López-Cuadrado, Rosa Arruabarrena, Anaje Armendariz & Tomás A. Pérez
Department of Computer Languages and Systems
University of the Basque Country (UPV-EHU)

Abstract

The developers of online language learning systems must be aware that every new learner has a different level of knowledge. Implementing Computer Adaptive Tests (CATs) is just one way to place incoming students at their appropriate level in the program. CATs emulate the intelligent behaviour of human evaluators because, depending on the previous responses given by the examinee, they dynamically select and administer the most appropriate items. In order for a CAT to be successful, these items have to be calibrated, which means that some parameters, such as their difficulty, must be previously estimated.

This paper presents our experience after having calibrated an item bank for the assessment of the Basque language. This item bank will feed the CAT generator that is intended to be the entry point to ELSA/Hezinet, an adaptive hypermedia system currently used for Basque language learning. The calibration process has been done in two complementary ways: guided by experts and by applying the psychometrical methodology.

We discuss some issues related to the calibration of the item bank that emerged from this work and that might be of interest to those who want to implement their own CAT generator. Some of the points treated here are the suitability of CATs for the assessment of a language, the assessment of the characteristics of the items (such as their quality or dimensionality), and main decisions to determine which calibration process best fits one's needs.

1. Introduction

Most educational systems are provided with a mechanism that assesses the progress of the students while acquiring knowledge. This is something critical to identify the success or the failure during the learning process. Actually, the developers of e-learning systems in general, and online language learning tools in particular, have to be aware of the fact that every new learner has their own level of knowledge. Thus, it is necessary to place the students at their corresponding stage, so they can progress properly as they interact with the e-learning system. Otherwise, students could get discouraged and lose interest.

Implementing Computer Adaptive Tests (CATs - Wainer, 2000) is just one way to place incoming students at their appropriate level in the program. Some e-learning systems and educative programs are based on this kind of adaptive evaluation, for instance, Cito’s tests
of Dutch as first and second language (Eggen, 2003); eCAT, which assesses the written English level of Spanish speakers (Olea et al., 2004); Geroline, a Web-based course for beginner German (Heift and Schulze, 2003); ELSA/Hezinet, an adaptive hypermedia system for the Basque language learning (López-Cuadrado, 2008); and the TOEFL computerized placement test (Wainer and Wang, 2000).

CATs emulate the intelligent behaviour of human evaluators. In fact, they select and administer the next item dynamically, depending on the previous responses given by the examinees (i.e. those that really provide useful information about student’s ability). One can find different theories that support the generation of CATs, as the Sequential Probability Ratio Test (Reckase, 1983), the combination of granularity hierarchies and Bayesian nets (Collins et al., 1996) or the Measurement Decision Theory (Rudner, 2002). However, the most used framework for computerized adaptive testing is the Item Response Theory (IRT - Lord, 1980), an item-oriented background that offers models that associate the ability of the examinee with the probability of a correct response. To choose the proper item from the bank, the CAT algorithm needs to know the values of some psychometric characteristics (called parameters) that feature the items. In terms of the IRT, this means that the item bank must be calibrated according to some model.

The most used IRT models are the 1-parameter logistic model (1PL - Rasch, 1960), which characterizes an assessment item by its difficulty, and the 3-parameter logistic model (3PL - Birnbaum, 1968), which also manages items’ discriminative power together with the guessing probability. Both models are one-dimensional, since they take just one trait into account, concretely the ability that the items measure. However, to ensure the item bank will work properly, it is necessary to explicitly verify this and other IRT-related properties that will be discussed later, in section 4.

This paper is organized as follows: next section introduces the most used ways to carry out an item bank calibration, that is, using statistical procedures and with the help of experts; section 3 presents our experience calibrating an item bank for the assessment of the Basque language, both under the guidance of a set of experts and by applying the psychometrical (statistical) methodology. Section 4 summarizes the lessons learned during these developments and discusses some issues related to the calibration of our item bank that might be of interest to those who want to implement their own CAT. Finally, section 5 draws the main conclusions.

2. The item bank calibration

The calibration of an item bank consists in determining the values for the item parameters, in terms of the chosen IRT model. The two most common ways to obtain these estimations are based on statistical methods and the judgement given by a set of experts. The former, which in this context will be called IRT calibration, allows obtaining not only the difficulty of every item in the scale used by the IRT, but also their discriminative power and guessing factor. The expert-based calibration, the latter one, is recommendable to estimate only the difficulty of the items, since the measurement of the other parameters can be complicated, but it can be applied in wider contexts rather than IRT-based computerized adaptive testing.
The first step of an IRT calibration consists in gathering the responses given to the items by a large group of examinees, so that item parameters can be statistically estimated. If the 1PL model is used, a minimum of 100 or 200 responses per item is required (Wright and Stone, 1979), whereas a sample of at least 500 individuals is recommended for the 3PL model (Bunderson et al., 1989). To perform such a dense task (many items, many individuals), and also because of security matters, it is recommended to distribute the evaluation items into several test forms (called subtests) and apply them separately. These subtests should share a number of items, which are called anchor items. Once the subtests have been administered to hundreds of individuals, parameter estimates can be computed statistically and then, thanks to the existence of the anchor items, equated in a common scale (Kolen and Brennan, 1995). When planning an anchor design, it is very important to select a set of anchor items that is representative of the whole item bank, and to distribute the items in a proper way to obtain subtests that share the same specifications about both content distribution and proportions.

Expert-based calibration is useful when authoring and calibration processes are done using separate roles. This kind of calibration consists in asking one or more experts in the field to give their personal, subjective, estimations for the item parameters. According to the literature (Arruabarrena et al., 2003), it is recommended to gather responses from 5 to 7 experts per item. The expert-based calibration is suitable for the 1PL model, which handles only the difficulty of the items and can be considered as a variation of the 3PL in which both the discriminative power and the guessing factor are constant and equal, respectively, to 1 and 0. Once the questionnaires have been created and distributed, the responses should be gathered and filtered. It is very important to reach a minimum of 5 answers per item. Only then, the difficulty for each item can be computed, for instance, by maximum consensus or by a bounded mean estimator.

3. Our experience

We have calibrated an item bank for the assessment of the Basque language by following two parallel calibration processes: Cal-Exp, an expert-based calibration intended to obtain difficulty estimates for the items (Arruabarrena et al., 2010), and Cal-IRT, an IRT-based calibration based on the 3PL model (López-Cuadrado et al., 2010).

Cal-IRT performed two types of subtest administration: first, supervised sessions, which were carried out in laboratories at schools, academies and universities, and second, non-supervised administrations, in which volunteers completed a subtest through the Internet by their own. During supervised sessions the administrator was responsible for assigning the examinees an identification code (IC), so every individual would be identified and attached to a session. The supervisor also had to control whether the examinees matched some conditions of administration and, therefore, invalidate those that had not met them. Thus, whenever a group of supervised test form administrations finished successfully, both the laboratory session and the administration of the set of subtests had to be validated. In contrast, the individuals that filled up a subtest not linked to a supervised session were required to use their e-mail address or telephone number as IC. Furthermore, this IC was
later used to contact every anonymous volunteer and determine whether the conditions had been satisfactory or not, that is to say, to decide if the test administration had to be validated or rejected. Subsequently, any incomplete application was rejected, as well as those test forms having an unknown, invalid or meaningless IC, and those answered by individuals having acknowledged that they had completed their subtest without paying much attention, by chance, just to try the application, with continuous interruptions or even with some extra help.

Cal-Exp was also carried out in two stages: initially, the questionnaires were delivered and fetched face-to-face, directly to-from experts' workplace, but later we decided to send and receive them by post, once we had previously contacted them by phone or e-mail. Experts were required to provide three pieces of information per item: the correct answer of the item and their subjective estimates for its difficulty and its learning skill. Asking for more than simply the difficulty of the items was useful to detect anomalous situations, i.e. items that were answered incorrectly by an expert and items assessing an unclear learning skill.

At the present time we are working in the development of a calibration supporting tool that is expected to guide educators and content developers during the decisions needed to be made in the progress of the calibration (Armendariz et al., 2009). Either it is an expert-based or an IRT-based process, the help tool will show the user how the most significant variables of the process are related and how changes in one of them could affect the others.

4. Lessons learned

This section summarizes the lessons learned after having finished the two calibration processes regarding the item bank for the assessment of the Basque language.

4.1. Lesson learned 1:

_The items to be calibrated should be of quality, up to date and homogeneously distributed along the difficulty scale._

This issue is critical in the particular case in which the resulting calibrated items are intended to feed a CAT, because these tests administer significantly fewer items than the ordinary paper-and-pencil trials. This means that much trust is placed in every single item. Moreover, the operation of the CAT depends on the answers given to any previously applied item, so the effects of including an erroneous or incorrectly functioning item in the bank could be critical. In addition, if the items are not homogeneously distributed along the difficulty scale, the CAT will not be able to deal with any possible response pattern, unless new items are calibrated and added to the bank.

The Basque language is continuously evolving since new regulations appear regularly trying to unify the dialects into a common language. Our initial item bank for the assessment of the Basque language was rather old, so we expected that some of the items would not be up to date due to the non-fulfilment of some recent grammatical rules. Thus, to ensure the correctness of the items, each of them was analysed by 2–4 reviewers, all of
which were either philologists or Basque linguists. After gathering their reports, a total of 265 mistakes were identified. Most of them were insignificant and their correction was automatic, essentially because they related to a typo and/or to the non-fulfilment of some punctuation or spelling rule. However, 76 severe errors affecting 56 items were found. They were, for instance, items that presented more than one possible correct answer, violations of new grammatical rules, ambiguities or lack of accuracy in the statement, even response options that, depending on the context, were all incorrect. Thanks to the contribution of the reviewers, most of these mistakes were rectified.

After having calibrated the item bank, we have found out that it will work better when assessing individuals with a low-middle level of knowledge of Basque language, since the distribution of the items was denser in these levels. As a result, we have planned to calibrate and include new items with high difficulty in order to obtain accurate ability estimates when assessing proficient examinees.

4.2. Lesson learned 2

Even then, 10–20% items will be removed for not passing the IRT model-fit analyses.

During a CAT-based assessment few items are administered, which is why the effect of erroneous items is especially critical. For this reason, it is necessary to remove from the bank any item that is inappropriate either from a didactic point of view or from the psychometrical perspective. The former can be found, for instance, thanks to the work carried out by the experts, whereas the latter can only be identified by means of reliability analyses of the responses gathered in the subtest administration stage. Some of these statistical studies are the identification of anomalous response patterns, classical reliability analyses such as the values of Cronbach’s alpha and the Spearman–Brown coefficient computed to obtain the item-subtest correlation, item differential functioning analyses, the computing of goodness-of-fit indicators, and the assessment of unidimensionality. However, in the case of item banks that assess knowledge on a given language, in most cases the verification of the restriction of unidimensionality, which is imposed by the IRT model, causes no problem.

In the case of the Cal-Exp calibration, we defined some criteria to filter unreliable items and to preserve the quality of expert contributions. For instance, "an item is not discarded, if at least 50% of the experts answer it correctly" and "a questionnaire completed by an expert is discarded if more than 25% of the given answers are wrong" were included in the set of conditions. These criteria were used to build three data sieves that were consecutively applied. As a consequence of the filtering stage, 60 items were labelled "to be removed from the bank", as well as 5 complete questionnaires.

Similarly, as a result of the reliability analyses conducted in Cal-IRT, 46 items were removed from the bank for either having an item-subtest correlation value close to zero or presenting some differential functioning.

The results of these analyses imply that more than the 17% of the initial set of items should be discarded, which is an amount halfway between the 10% removed during the calibration
of the eCat item bank (Olea et al., 1996) and the 25% invalidated during the calibration, in similar conditions, of a bank used to assess school contents (García et al., 1999).

4.3. Lesson learned 3

*Lessons learned 1 and 2 imply that the number of items to be calibrated should be large enough to (1) ensure a minimum of them pass the IRT model-fit analysis, and (2) minimize overexposure problems.*

The fact that even one fifth of your initial item bank could disappear while calibrating is an important piece of information to take into account, particularly if you want to build a reliable and efficient CAT. So, it is important to begin a calibration process having an initial set of items significantly larger than the foreseen one.

Furthermore, as time goes, and more and more examinees are assessed by the CAT, overexposure problems can arise. This means that psychometrically better items are administered very frequently while not so good items are hardly ever used. As a consequence, the majority of the CATs generated will share most of their items, whose confidentiality will be quickly compromised. There are several modifications of the CAT algorithm that minimize the effect of this problem (see, for instance, Eggen, 2001), but there is no doubt that having a larger item bank will also help.

4.4. Lesson learned 4

*Use an expert-based calibration for the model 1PL.*

The difficulty estimates obtained through a psychometrical IRT-based calibration is comparable to the results of an expert-based calibration, provided that a minimum of 7 experts are enquired (Arruabarrena, 2010). There are not statistically significant differences in the difficulty values obtained by the different calibration procedures for the same item bank.

Since similar results are obtained when only the difficulty of the items is needed, it is better to carry out an expert-based process. The reason is that using experts is cheaper and consumes less time and human resources than implementing an IRT-based calibration (Arruabarrena, 2010). Taking as the starting point the costs related to both Cal-Exp and Cal-IRT, we have extrapolated them for different item bank sizes, and concluded that the costs difference, in terms of time, human resources and money, is larger as the number of items to be calibrated increases.

4.5. Lesson learned 5

*When conducting an expert-based calibration, (1) be careful when selecting the set of experts, (2) contact them by (e-)mail, and (3) validate the results.*

Before attempting an expert-based calibration, it is very important not only to ensure that a set of experts will be available for that purpose, but also that the so-called experts are
actually expert in the field. Otherwise, unreliable estimates will surely be obtained. Notice
that an expert is someone that knows about the language that is assessed, but the other way
round is not always true. In Cal-Exp, questionnaires were administered to both linguists and
pedagogues, all them having experience in Basque language education.

Our experience says that, although the first contact should be face-to-face, once the
working conditions have been established, it is better to contact them by (e-)mail. It is also
recommendable, when possible, to contact an extra amount of experts to let them complete
less questionnaires, because overloading a small number of people with too much work will
imply a higher abandonment rate, especially if the experts are volunteers and therefore they
are not paid for their service.

Finally, the results must be filtered and validated by means of some known estimators or
indexes. In the case of the Cal-Exp calibration, we used Cohen's Kappa index to ensure that
the inter-rate reliability of experts in their difficulty estimations was good.

4.6. Lesson learned 6

*Use an IRT-calibration for the model 3PL.*

If the items to be calibrated are true/false or multiple-choice questions, there is no need of
experts to estimate their guessing parameters in a reasonably accurate way. Actually,
assuming that all options are equally plausible, the actual values will be close to \(1/n\) (where
\(n\) is the number of response choices of the item): that is, 0.5 for a true/false question, 0.33
for a multiple-choice item with 3 options, and so on. However, for other types of
assessment items, such as open-ended questions or checklists, it becomes very difficult to
manually obtain estimates for their guessing parameters.

The same thing happens with the discriminative parameter, which is related to how well the
item is able to differentiate examinees of slightly different ability. This value can hardly be
estimated, unless some statistical procedure is applied. For this reason, if one wants to
calibrate an item bank following the model 3PL, the method to be used is the one based on
the IRT.

4.7. Lesson learned 7

*When conducting an IRT-based calibration, (1) organize supervised laboratory sessions,
(2) administer subtests to similar groups, and (3) use specific software for statistical tasks.*

During the process Cal-IRT, the subtests were administered in two complementary ways.
First, supervised sessions were performed in schools, academies and universities, and
second, in non-supervised subtest administrations volunteers were allowed to complete a
test without supervision.

To decide which way is the best to be applied in future IRT-based calibrations, not only
costs in terms of time and resources should be taken into account, but also the implications
of the validation criteria and the abandonment rate. Actually, in the case of Cal-IRT, 40%
of the individuals that started a non-supervised administration either did not finish it or failed to pass the validation process. In contrast, this rate was smaller than 3.5% when there was someone supervising the fulfillment of the subtest. This is why we consider that it is better, when possible, to arrange supervised laboratory sessions rather than allowing non-supervised subtest administrations.

The administration stage is probably the hardest one during the item bank calibration. It requires time and effort to apply so many items to many individuals, so it is necessary to dedicate much time to the design and administration of the subtests. As we have pointed out in Section 2, it is desirable all the subtests to be as similar as possible, and also the characteristics of the subtests and the item bank to be comparable (i.e. each one should cover all the levels of difficulty and knowledge in the same proportion). Moreover, each subtest should be applied to homogeneous groups with heterogeneous levels of knowledge.

Finally, one can find specialized software that is helpful to automatically carry out some of the statistical studies and tasks related to an IRT-based calibration. For instance, in the case of Cal-IRT, PRELIS (Jöreskog and Sörbom, 1993b) and LISREL (Jöreskog and Sörbom, 1993a) were used in the study of unidimensionality, as well as the program XCALIBRE (ASC, 1997), which computed the item parameter estimations and analysed the data-to-model fit.

5. Conclusions

The systems that use Computerized Adaptive Tests aim to measure the level of knowledge of the students in a specific domain by emulating the intelligent behaviour of human evaluators. In other words, these tests dynamically select the most appropriate item depending on the examinee's performance during the test. This means that if they give a correct answer then the next question will be slightly more difficult, whereas incorrect answers will be followed by easier items.

Nowadays CATs are used to assess almost any educational area (López-Cuadrado, 2008), from musical aptitudes to the level of English as foreign language. This kind of tests is suitable for language learning, since they give accurate results by administering few items. Actually, many of the existing CATs are related to the measurement of knowledge on certain language, like Cito’s tests of Dutch as first and second language, eCAT, which assesses the written English level of Spanish speakers, Geroline, a Web-based course for beginner German, ELSA/Hezinet, an adaptive hypermedia system for the Basque language learning, and the TOEFL computerized placement test.

The IRT provides very powerful techniques to carry out the evaluation, particularly when using CATs, but it imposes significant constraints. The most important one has to do with the fact that, to identify the proper item from the item bank during a CAT administration, the implemented algorithm needs some parameters to be available, i.e. difficulty if the 1PL model is used, and difficulty, discrimination and guessing factor in the case of the 3PL model. The process in which these values are estimated is called calibration, and it can be done in two different ways: establishing the values of the parameters basing on judgements
from multiple experts, or using statistical procedures to obtain them from a sample of item administrations.

This paper has presented some lessons learned from the experience of having calibrated by both ways an item bank for the assessment of knowledge on Basque language, which we expect to be useful to those attempting to calibrate their own item banks.

The items to be calibrated should be of quality, up to date and homogeneously distributed along the difficulty scale. Even then, 10~20% items will be discarded for not passing the IRT model-fit analyses. This implies that the number of items to be calibrated should be large enough not only to ensure a minimum of them pass the IRT model-fit analysis, but also to minimize overexposure problems.

An expert-based calibration is recommended if the 1PL model is used to feature the items. In this case, it is important to carefully select the set of experts. We recommend any communication to be made by post or e-mail, and to methodically validate the answers given by the experts, discarding them when necessary.

An IRT-based calibration is the only choice for the 3PL model, mainly due to the difficulty in judging the discriminative power of the items. When conducting such a calibration, it is better to organize supervised laboratory sessions than to allow non-supervised subtest administrations. However, depending on the availability of human resources, this will not always be possible to achieve. If subtests need to be defined, it is important to arrange them carefully and administer them to similar groups of people. Finally, specific software such as PRELIS/LISREL or XCALIBRE is available to carry out statistical tasks that would be very hard to do otherwise, especially for those who do not have skills in psychometrics.

References


