An Empirical Study About Calibration of Adaptive Hints in Web-Based Adaptive Testing Environments

Ricardo Conejo, Eduardo Guzmán, José-Luis Pérez-de-la-Cruz, and Eva Millán

Departamento de Lenguajes y Ciencias de la Computación,
Universidad de Málaga,
Bulevar Louis Pasteur, 35,
Málaga, 29071, Spain
{conejo, guzman, perez, eva}@lcc.uma.es

Abstract. In this paper we present a proposal for introducing hint adaptive selection in an adaptive web-based testing environment. To this end, a discussion of some aspects concerning the adaptive selection mechanism for hints is presented, which results in the statement of two axioms that such hints must fulfil. Then, an empirical study with real students is presented, whose goal is to evaluate a tentative bank of items with their associated hints to determine the usefulness of such hints for different knowledge levels and to calibrate both test items and hints.

1 Introduction

Testing is commonly used in many educational contexts with different purposes: grading, self-assessment, diagnostic assessment, etc. In order to improve the efficiency of the diagnosis process, adaptive testing systems select the next best question to be asked according to the relevant characteristics of the examinee. In this way, higher accuracy can be reached with a significant reduction in test length. In literature, there are different proposals for adaptive testing [1], [2]. One of the most commonly used is the *Item Response Theory* (IRT) [3], which has a well-founded theoretical background which assumes that the answer to a question depends on an unknown latent numerical trait. The latent trait θ represents a psychological factor that we want to measure and that is not directly observable. In educational environments, the latent trait corresponds to the knowledge of the subject being tested.

In any adaptive educational system, it is necessary to have accurate estimations of the student's knowledge level in order to take the most suitable instructional action. In this sense, *Computerized Adaptive Tests* (CATs) [4] based on IRT provide a powerful, efficient and reliable diagnosis tool. SIETTE [5], [6] is a web-based assessment environment that allows the construction and administration of conventional tests and CATs based on a discretization of IRT. One of the most relevant characteristics of SIETTE is that it is an open assessment tool, i.e, it can be easily integrated into any web-based learning system. In this

V. Wade, H. Ashman, and B. Smyth (Eds.): AH 2006, LNCS 4018, pp. 71–80, 2006.

way, SIETTE can be responsible for all the tasks concerning student modelling (basically creation and maintenance of the student model). This system can be accessed at *http://www.lcc.uma.es/SIETTE*.

One of the contributions to educational psychology in the XX century is Vigotskii's *Zone of Proximal Development* (ZPD) [7], defined as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with a more capable peer". A short operational definition useful for our purposes is given in [8]: the zone defined by the difference between a child's (in our case, person's) test performance under two conditions: with or without assistance.

Soon after the definition of the ZPD, attempts to apply this concept were made in the context of the test administration, under the two conditions described (with or without assistance), typically with the aim of classifying students in order to allocate them in the most appropriate educational program. But the main goal of the work presented here is different: to build a model that allows the integration of adaptive assistance in the adaptive testing procedure within the SIETTE system.

Hinting can be considered a general and effective tactic for human tutoring. In this sense, some researchers have put the emphasis on the mechanisms used by students to request hints when needed [9, 10]. On the other hand, many Intelligent Tutoring Systems also give hints to the student, like for example, ANDES [11], which calculates the score according to the correctness of the student's answer and the number of hints received; or AnimalWatch [12], which has different types of hints available (highly/low interactive and specific/symbolic hints) and adapts such hints to relevant features of the student such as the level of cognitive development and gender. It can be observed that human tutors maintain a rough assessment of the student's performance (the trait in our approach) in order to select a suitable hint [13].

Consequently, we will assume that assistance is represented by hints, h_1,..., h_n that provide different levels of support for each test question (commonly known as *items* in adaptive testing environments). By adaptive assistance we mean that the hint to be presented will be selected by the system depending on where the item is in the ZPD, in such a way that it provides the minimal amount of information and yet the student will still be able to correctly answer such an item.

The work presented here further extends our investigation about introducing hints and feedback in adaptive testing, presented in [12]. Now, our main objectives are:

- Definition of a theoretical framework for adaptive hinting selection.
- Empirical study of the feasibility of the approach. To this end, an item bank has been developed for a course. Each item had a set of hints assigned. This bank has been tested in several experiments, all of them with real students. The final goal of these experiments was to validate the hints developed.

A further issue considered in paper is the calibration of the pairs of item-hints. Once this procedure has been accomplished, it will allow tests to be delivered in which both items and hints are adaptively selected according to the current estimation of the student's knowledge level.

2 Computerized Adaptive Testing and Item Response Theory

The CAT theory when combined with IRT allows a well-founded administration of adaptive tests. A *Computerized Adaptive Test* is a computer-based test where the decision to present a test item and the decision to finish the test are dynamically made depending on the examinee's performance in previous answers. If the CATs of two examinees are compared, each one of them will usually receive different sequences of items, and even different items. To properly administrate a CAT, each item i in the item bank is assigned an *Item Characteristic Curve* (ICC). An ICC is a function representing the probability of a correct answer to that item given the student's knowledge level θ, which is unknown but supposed to be constant during the whole test. The probability P_i of succeeding when answering a test item ($u_i = 1$) can be computed as $P_i = P(u_i = 1|\theta)$, and the probability Q_i of failing as $Q_i = P(u_i = 0|\theta) = 1 - P(u_i = 1|\theta)$. If the test is composed of n items, knowing their ICCs, and assuming local independence of items, a likelihood function L can be constructed as shown below:

$$L(u_1, u_2, ..., u_n|\theta) = \prod_{i=1}^{n} P^{u_i} Q^{1-u_i} \tag{1}$$

The maximum of this function gives an estimation of the most likely value of θ. A probability distribution of θ can be obtained by applying Bayes' rule n times. It is usually assumed that ICCs belong to a family of functions that depend on one, two or three parameters. These functions are constructed based on the normal or the logistic distribution functions. For example, in the three-parameter logistic model (3PL)[14] the ICC is described by:

$$CCI_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-1.7a_i(\theta - b_i)}} \tag{2}$$

where c_i is the *guessing factor*, b_i is the item *difficulty* and a_i is the *discrimination factor*. The *guessing factor* is the probability that a student with no knowledge at all answers the item correctly. The question *difficulty* represents the knowledge level in which the student has equal probability of answering or failing the item, in addition to the guessing factor. The *discrimination factor* is proportional to the slope of the curve.

Based upon the IRT and the CAT theory, our group has developed and implemented the SIETTE for adaptive testing construction and administration via the Web. In contrast with traditional IRT-based proposals, the knowledge level in SIETTE is a discrete variable that can take $n + 1$ values $v_0 < v_1 < ... < v_n$

(in the sections which follow, we will represent these values as $0, 1, ..., n$). In this way, in SIETTE ICCs are represented by vectors of $n + 1$ elements. Hence, computation of Bayes' rule is simply turned into a product of $n + 1$ values plus a normalization procedure. The main advantage of this discretization is that it improves the computational efficiency of the calculus, turning SIETTE into a scalable system. However, this entails a slight loss of accuracy in estimations.

3 Introducing Hints in an Adaptive Testing-Based Assessment Model

To introduce hints in the model, let us first define some terms:

- *Item.* We use this term to generically denote a question or exercise posed to a student. The solution of such task or question could be provided in different manners: by selecting one or more choices available within the item, or even allowing the examinee to write a brief text.
- A *test* is a sequence of items.
- *Hint.* A hint is an additional piece of information that is presented to the student after posing an item and before he/she answers it. Hints may provide an explanation of the stem, clues for rejecting one or more choices, indications on how to proceed, etc. Hints can be invoked in two different ways: a) *actively*, i.e., when the examinee asks for the hint by clicking a button; or b) *passively*, that is, when the hint is triggered as a consequence of his/her behavior while answering the item, indicating that the student has reached an impasse (for example, too much time waiting).

Let us see a simple example. Consider the following test item:

```
What is the result of the expression: 1/8 + 1/4?
    a) 3/4  b)2/4   c)3/8   d)2/8
```

Possible hints may be:

```
Hint 1. 1/4 can be also represented as 2/8.
Hint 2. First, find equivalent fractions so they have
the same denominator.
Hint 3. Once fractions have the same denominator,
sum up numerators.
```

In the work presented here, a simplifying assumption is that *hints do not modify student's knowledge* (i.e., no student learning occurs either while testing or when receiving hints). This assumption is usual in adaptive testing (the trait θ remains constant during the test), and makes the model computationally tractable. In our case, this hypothesis means that hints do not cause a change in examinee's knowledge, but there is a change in the ICC shape. In this way, the hint brings the question from the ZPD to the student's knowledge level. In this sense, the combination of the item plus the hint can be considered as a new

item. This new (virtual) item can be treated and measured in the same way as the other items in the test: the new item is represented by a new ICC whose parameters can be estimated using the traditional techniques. However, both ICC's are not independent of each other. First of all, the use of a hint should make the question easier. This condition can be stated in mathematical terms by the following:

Axiom 1. Given an item q and a hint h, for all knowledge levels k, the following constraint must be fulfilled: $CCI_q(k) \leq CCI_{q+h}(k)$. ICC_q represents the original item characteristic curve and ICC_{q+h} represents the characteristic curve of the item with the hint.

If the examinee uses a combination of hints, the question should become even easier. Mathematically this condition can be written as follows:

Axiom 2. Given an item q, a set of hints H and a hint $h \notin H$, for all knowledge levels k, the following constraint must be fulfilled: $CCI_{q+H}(k) \leq CCI_{q+H+\{h\}}(k)$.

For a set of items and their corresponding hints, after the ICC parameters calibration[1] (of the real and virtual items), if the resulting ICCs do not satisfy the axioms above, it means that the piece of information given is a misleading element instead of a hint; therefore, it should be rejected. This simple approach provides us with a useful empirical method that allows validation of the proposed hints.

In adaptive environments, it makes sense to look for a criterion for adaptively selecting the best hint to be presented (from a set of available hints). Under the ZPD framework, if the student is not able to solve the item but this item is in his/her ZPD, the best hint to be presented would be the one that brings item I from the ZPD to the zone of the student's knowledge, and of course it will depend on how far on the ZPD the item is located. So, for example, if an item I has three associated hints h_1, h_2 and h_3 at different levels of detail, it means that each hint is suitable for a different part of the ZPD.

Therefore, the selection of h_i as the best hint to be presented would mean that the item I lies in ZPD_i for this particular student. A possibility for adaptive selection of hints is to use classical adaptive item selection mechanisms, e.g. given the knowledge estimation $\theta(k)$ for a student, and given two hints h_1, h_2, with $ICC_{q+\{h1\}}(k)$ and $ICC_{q+\{h2\}}(k)$, the best hint to use is the one that minimizes the expected variance of the posterior probability distribution. This mechanism is simple to implement and does not require substantial modifications in the adaptive testing procedure, because the test is only used for assessment and not as a learning tool. However, the use of adaptive hints in this context can provide positive stimuli and, as a consequence, increase student self-confidence.

[1] The calibration process consists of inferring ICCs from the student initial score. As a result, a first estimation of the student knowledge θ can be computed from these ICCs. This procedure will iterate until an equilibrium is reached. Therefore, students having asked for more hints (i.e., those students which presumably answered "easier" items), obtain a score which is very close to the one obtained when no hints are requested. ICCs take into account the lower or greater item "difficulty".

4 Experiments with Real Students

An important first step towards the integration of this new adaptive hints approach into the SIETTE system is the calibration of the virtual items resulting from the combination of items and hints. Calibration of hints is a difficult goal, and, to this end, a methodology composed of several steps must be observed:

1. First of all, an item bank must be developed. Each item must also have several hints assigned.
2. Second, items must be administered to a student sample by means of a conventional (i.e. non-adaptive) test. After that, characteristic curves of real and virtual items must be calibrated.
3. Finally, once the ICCs have been inferred, adaptive administration of the test and the hints assigned with its items can be accomplished.

Regarding the first step, an item bank relating to a course of *Language Processors* has been developed. This course is taught at the Computer Science School in the University of Málaga (Spain). Each item has 2, 3 or 4 associated hints. Examples of such items are:

```
1. What is the output of the following LEX program with input abc?
    ab/c  { printf(''one'');   }
    c     { printf''two'');    }
    abc   { printf(''three''); }

    a) three b) one two  c) one  d) one two three

    Hint 1. yytext does not contain the characters on the right
    hand side of the lookahead operator ''/''.
    Hint 2. When the regular expression includes a lookahead
    operator, the length of the string matched corresponds to the
    part of the expression on the left of the operator.

2. Let T be the set of all ASCII characters from 0 to 127. The set
of all strings that can be formed with the symbols of T, can be
represented in LEX by the regular expression:

    a)(.|\n)*   b)[a-zA-Z0-9]*  c).*   d)[.]*

    Hint 1. The .(dot) operator represents any ASCII
    character, except the end of the line.
    Hint 2. The .(dot) operator does not have any special
    meaning when it is used inside the brackets [].
    Hint 3. The ASCII alphabet includes more than letters and
    digits, it also includes operators, punctuation symbols,
    parenthesis, and other special characters.
```

In the second step, three experiments have been carried out, with a total number of 263 individuals. Experiments included students taking the *Language Processors* course during 2003/04, 2004/05 and 2005/06. The sample size was 100, 80 and 83 students, respectively. All students were graded by means of a non-adaptive test-based exam composed of 20 items. These tests were administered using the SIETTE system. Students had a time limit of 45 minutes available to complete the test. The majority of students (87%), completed the test. 97% of them answered at least 18 items. All these data were used in the analysis described in this paper. For each test item, students were given the possibility of requesting a hint. It was a heterogeneous test where several types of items were combined (multiple-choice with just one correct choice, multiple-choice with more than one correct choice and fill-in-the-blank items). The same 20 items were posed to all students, but in a different order, in order to avoid cheating. Once a student requested a hint, it was randomly selected from the pool of hints assigned to the item.

The scoring method differed according to the experiment. In all of them, for each correctly solved, the student was awarded 1 point (in order to pass the exam students needed a minimum of ten points). However if a hint had been used, the correct solution only gave 0.5 points in the first experiment, 0,75 in the second, and 1 (i.e., no penalization) in the third one.

Table 1. Portion of students answering correctly

Item 1	*No hint*	*Hint 1*	*Hint 2*	**Item 2**	*No hint*	*Hint 1*	*Hint 2*	*Hint 3*
Correct	108	20	25	*Correct*	134	20	13	20
Total	198	29	34	*Total*	176	31	26	27
Percentage	54,5%	68,9%	73,5%	*Percentage*	76,1%	64,5%	50%	74%

Table 1 collects the results for items 1 and 2. It shows the number of students who correctly answered the corresponding item. The second row contains the portion of students, and the third row the corresponding percentage. For instance, the pair 108/198 in the first position of the first row, indicates that 198 individuals answered item 1 without asking for hints, and from this set, only 108 students gave the correct answer.

Fig. 1 shows the total percentage average of hints requested for each item and experiment. The total use of hints was 8, 7 and 53%, respectively. This suggests students only perceived qualitatively the penalty applied for hint requesting, since there are not significant differences between the first two experiments, in spite of the fact that the penalization in the second experiment was lower. The use of hints was much more frequent in the third experiment because, as explained before, they were not penalized. Still, there were students that decided not to ask for hints, probably because they knew (or thought they knew) the correct answer or because the time of the test was limited and they did not want to waste time reading the hint.

With regarding to the real usefulness of hints, which we define as the percentage of items successfully answered after requesting a hint, this was around 50% (more specifically, 38%, 47% and 56% in each of the three experiments).

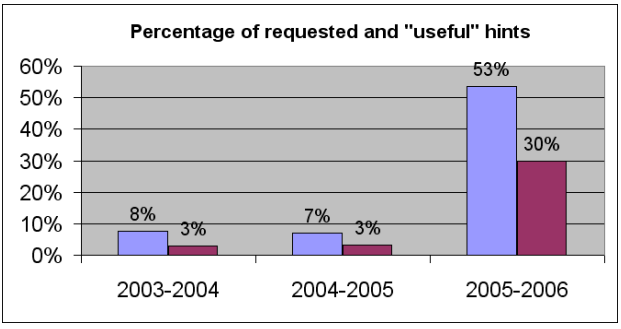

Fig. 1. Results about the use of hints in the three experiments

4.1 Item Calibration Under the 3PL Model Analysis

Much more interesting is a whole analysis of the ICC achieved using the IRT. For this purpose, we will assume all ICCs follow the 3PL model, formerly expressed in equation 2.

In this new experiment, 81 ICCs have been calibrated: 20 curves corresponding to the real items, and in addition, the curves of the 61 virtual items obtained from the combination of each pair *item+hint*. To this end, we have used one of the most popular item calibration tools in IRT, i.e., MULTILOG [15]. Results

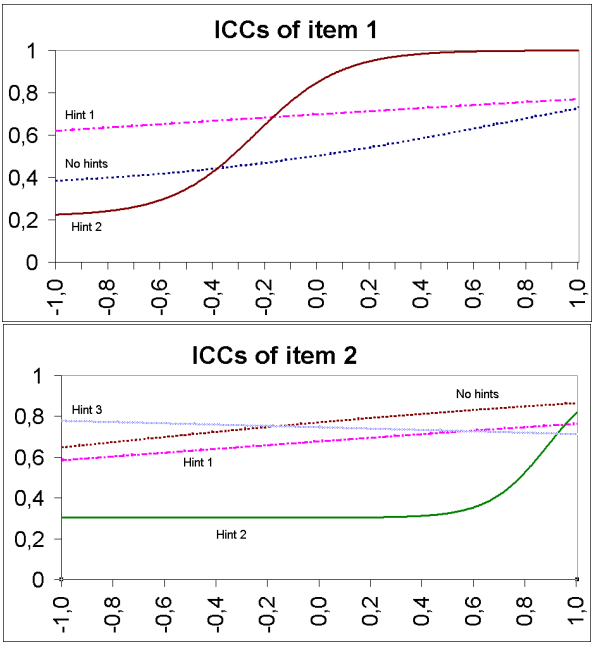

Fig. 2. ICCs of items 1 and 2

show that student knowledge level distribution (θ) is normal with a mean of $0,018$ and standard deviation of $0,337$. That is, $99,7\%$ of the sample exhibit θ values located at the interval $[-1,1]$. Fig. 2 depicts the results for items 1 and 2 in this interval. The horizontal axis displays the knowledge level (θ), and the vertical axis the probability of correctly answering the item. The calibrated ICCs of the items without hints are represented by a dotted line.

As can be seen, after this analysis, hints of item 1 are useful only from a certain knowledge threshold. This threshold is located at the intersection of both curves. Hint 1 is useful for the majority of students but, in contrast, hint 2 is only useful for those whose knowledge level satisfies the following constraint: $\theta > -0,35$, i.e., for the 84% of the sample individuals. Regarding item 2, observe that hint 1 does not provide any improvement in the success percentage, and its curve is very close to the one corresponding to the original item. Consequently, we can infer that it does not contain relevant information to help students solve the item. Likewise, hint 2 is counterproductive for the majority of students. Finally, hint 3 is as a misleading element (note that the slope of the ICC is negative) so this hint should be discarded.

5 Conclusions and Future Work

This paper has presented some ideas about hint adaptive selection in an adaptive testing environment, based upon IRT constructs. Hints are considered not as knowledge modifiers, but as modifiers of the ICC of an item. Some formal axioms that every model of hints must satisfy have been stated and informally justified.

We have also described the three different experiments with real students which we carried out between 2003 and 2006. In those experiments three student samples took a test composed of the same items. Once an item was posed, students were given the possibility of asking for a hint. Depending on the experiment, the use of a hint was penalized.

Finally, we have performed the ICC calibration based on well-founded IRT-based techniques. Calibration was done for each item, and also, for each virtual item, resulting from each pair *item+hint*. The input data used for calibration was the performance of real students who took the test in the three former experiments. Thanks to the ICCs inferred from calibration, we have got not only a set of calibrated ICCs, but a mechanism to discern between useful and useless hints, and to remove those hints that confuse the students.

Plans for immediate future work involve the adaptive administration of both items and hints and, accordingly, the evaluation of the benefits of our proposal.

References

1. Rudner, L.M.: An Examination of Decision-Theory Adaptive Testing Procedures. In: Annual meeting of the American Educational Research Association. (2002)
2. Chua Abdullah, S.: Student Modelling by Adaptive Testing - A Knowledge-based Approach. PhD thesis, University of Kent, Canterbury (2003)

3. Hambleton, R.K., Swaminathan, J., Rogers, H.J.: Fundamentals of Item Response Theory. Sage publications (1991)
4. Wainer, H.: Computerized Adaptive Testing: A Primer. Lawrence Erlbaum, Hillsdale, NJ (1990)
5. Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez de la Cruz, J.L., Ríos, A.: SIETTE: a Web-based tool for adaptive testing. Journal of Artificial Intelligence in Education **14** (2004) 29–61
6. Guzmán, E., Conejo, R.: A brief introduction to the new architecture of SIETTE. In Bra, P.D., Nejdl, W., eds.: Proceedings of the IIIth International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems(AH 2004). Lecture Notes in Computer Science. Number 3137. New York: Springer Verlag (2004) 405–408
7. Vygotskii, L.: Mind in Society: The Development of Higher Psychological Processes. Cambridge, MA: Harvard University Press (1978)
8. Wells, G.: Dialogic Inquiry: Towards a Socio-Cultural Practice and Theory of Education. New York: Cambridge University Press (1999)
9. Luckin, R., Hammerton, L.: Getting to know me: Helping learners understand their own learning needs through meta-cognitive scaffolding. In: Proceedings of the 6th World Conference of Intelligent Tutoring Systems. ITS'02. Springer-Verlag (2002) 759–771
10. Aleven, V., McLaren, B., Roll, I., Koedinger, K.: Towards tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. In: Proceedings of the 7th World Conference of Intelligent Tutoring Systems. ITS'04. Springer-Verlag (2000) 227–239
11. Gertner, A.S., Conati, C., VanLehn, K.: Procedural Help in Andes: Generating Hints Using a Bayesian Network Student Model. In: Proceedings of the 15th National Conference on Artificial Intelligence. Madison, Wisconsin (1998)
12. Arroyo, I., Beck, J.E., Woolf, B.P., Beal, C.R., Schultz, K.: Macroadapting Animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism. In: Proceedings of the 5th World Conference of Intelligent Tutoring Systems. ITS'00. Springer-Verlag (2000) 604–614
13. Hume, G.D., Michael, J., Rovick, A., Evens, M.W.: Hinting as a tactic in one-on-one tutoring. Journal of Learning Sciences **5**(1) (1996) 23–47
14. Birnbaum, A.: Some Latent Trait Models and Their Use in Inferring an Examinee's Mental Ability. In: Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley (1968)
15. Thissen, D.: MULTILOG: Multiple, categorical item analysis and test scoring using item response theory (version 5.1) (1988)