DIA: A Computerized Adaptive Testing Tool for Assessing Student Learning

Patrícia Nunes da Silva (Corresponding author)
Mathematical Analysis, Rio de Janeiro State University,
Rua São Francisco Xavier, 524
Zip code 20550-900
Rio de Janeiro, RJ, Brazil

Renata Cardoso Pires de Abreu
Colégio Santo Agostinho,
Rio de Janeiro, Brazil.

Carlos Frederico Fragoso de Barros e Vasconcellos
Computational Science, Rio de Janeiro State University,
Rio de Janeiro, Brazil.

Abstract

We present a computerized adaptive testing tool, termed DIA, for the assessment and provision of feedback to students from a formative evaluation perspective. We use Brazilian governmental guidelines for teaching mathematics (Brazil, 1998; Brazil, 1997) to construct a scale with goals increasingly ordered by the vertical development of mathematical knowledge. We construct a simulated item bank that meaningfully relates to our scale through the item response theory. We also analyze a feedback given by DIA.

Keywords: mathematical knowledge; formative evaluation; computerized adaptive testing; item response theory

1. Introduction

It is well known in the educational evaluation that assessment is essential to and inseparable from educational practice. In fact, we know that assessment is part of the learning process in any area of knowledge. However, its overriding presence can cause us to forget some fundamental questions: Why do we evaluate? What should we evaluate? How do we evaluate?

In formative assessment, judgments about the quality of student responses are used to shape and improve the student's competence (Sadler, 1989). The formative assessment gives the student the opportunity to become aware of his/her own difficulties, and possibly, to recognize and correct his/her own mistakes. For Hadji (2001), formative assessment takes a central role in learning actions. He explains that its main function is to contribute to the suitable regulation of such activities. Its task is to furnish useful information

for the regulation of the teaching-learning process. Perrenoud (1999) states that formative assessment helps students to learn and develop themselves, and thus, contributes to learning regulation.

An awareness of one's own knowledge is one aspect of metacognition. According to Brown (1978), as cited in Ribeiro (2003), the recognition of the difficulty in understanding a task or the awareness that something is not understood is a skill that seems to distinguish more and less capable students. Self-regulation, the ability to analyze and control the execution of a task and make necessary corrections is another important aspect of metacognition in the field of education. According to Perrenoud (1999), an assessment is more formative when it gives less importance to the classification and more to the regulation of learning.

Sadler (1989) points out that "feedback is a key element in formative assessment, and is usually defined in terms of information about how successfully something has been or is being done." The feedback of formative evaluation allows the student to assume the role of a regulator of his/her own learning process, thus conferring on him/her the ability to learn. According to Gipps (1998), it redefines the power relations in the evaluation, involving the learner as a partner who takes responsibility for his/her performance and monitors his/her own learning. The information provided by formative assessment is an additional ingredient for the teacher to reflect on the actual effects of his/her actions and, consequently, may regulate his/her pedagogical practice.

The merits of formative assessment impose great challenges. How do we distinguish and diagnose? How do we follow the learning process? Associating a methodology capable of accessing the skills and abilities of the student with a scale that clearly defines and ranks the descriptors associated with content permits feedback to each student.

The motivation of this research is to overcome low achievement in learning mathematics and to help students to take control of and to manage their own learning. We present a tool that supports formative assessment, provides information (feedback) pointing out the provisional ability level of the student in a specific subject and gives teachers the opportunity to revise their pedagogical actions; it also allows students to monitor the progress of their learning and performance, and to understand the difficulties and gaps in their individual learning processes. DIA tool uses the theory of response to the item to propose an adaptive computerized test that allows indicating to the student and teacher learning gaps. We present results obtained from simulated tests of its operation.

In Section 2, we present aspects of a computer adaptive test. We focus on describing the features of the item bank and the use of item response theory in selection criteria of the. In Section 3, we describe our tool. We indicate its underlying scale that will allow the interpretation of the results; how we use it to simulate an item bank and the provision of the feedback.

2. Computer Adaptive Test

As measures of knowledge obtained from averages involve various concepts, it becomes virtually impossible to identify the real gaps that prevent learning. It is necessary to use tools that can provide more detailed information on the level of knowledge of a student. In an adaptive computer test (CAT), each student is given a set of questions that are appropriate to his/her ability. To this end, the questions are

dynamically selected for each student, taking into account his/her individual performance during the test. Generally, the initial item is a random question of average difficulty. If it is correctly answered, the estimate of the student's ability is increased. Since the ability estimate has increased, it is assumed that he may also be able to answer a more difficult question. If the response is incorrect, the student's ability estimate is decreased, and an easier question is then presented (Lilley et al., 2004).

As Lilley et al. (2004) points out: "One of the principles of the CAT approach is that administering easy questions to a high-ability student is not efficient, as a correct response would provide low-value information about his or her ability. Likewise, an incorrect response from a less proficient student to a difficult question adds a little information about this individual's ability within the subject being tested. By selecting and administering questions that match the individual student's estimated level of ability, questions that present low-value information is avoided."

Heller et al. (2006) state that to enable a personalized learning process, one must attend to the needs and interests of each student. Moreover, it is necessary to guarantee that each student realizes his/her full potential. They point to individual ability and other characteristics or preferences as a starting point in the usual process. They also emphasize the role of theories such as item response theory[1] (IRT) in the realization of a personalized learning experience.

There are five basic ingredients for building a CAT: 1. item bank; 2. selection criteria of the items; 3. estimates of skills; 4. underlying scale that will allow the interpretation of the results; and 5. stopping criteria.

2.1 Item Bank

Costa (2009) states that a CAT requires the item bank to be composed of items with good pedagogical and psychometric aspects. For the psychometric evaluation of items, one finds strong statistical support in the IRT, which provides a quantitative analysis of certain item characteristics, such as difficulty and discrimination. When combined with IRT, the item selection is designed to permit a test to fit student skill levels. Thus, each student who participates in the evaluation may have a different test, depending on his/her competence. Usually, when using IRT, the item selected is the one that provides more information, given the ability of the individual.

To enjoy the advantages offered by IRT, the item pool from which the test items are selected must contain items of high quality for different levels of proficiency. This feature is markedly different from conventional tests that are built using items that better discriminate subjects with average skills. In addition to containing wide-ranging, high-quality items, the CAT item pool must meet the psychometric assumptions underlying the model and the demands of the calibration process and selection method (Flaugher, 1990).

Generally, at least four times the number of questions to be administered in a test sitting are required for verification. Furthermore, the questions should be evenly distributed across different ability levels (Lilley, 2007). For Ward (1981, as cited by Lilley (2007), contributing factors generally relate to the validity of objective tests in general; examples include "good syllabus coverage" and "precise questions" that can be

[1] IRT is a family of probabilistic models describing the characteristics of individuals that cannot be directly observed (latent variables) but can be inferred from the responses to test items. The IRT establishes a relationship between the latent variable or trait and an individual's response to a set of items on a test (Costa, 2009).

used to support the view that the CAT approach has content validity.

3. DIA: A Computerized Adaptive Testing Tool

In order to give feedback to the student of his/her learning gaps, we must hierarchize the knowledge to allow the identification of learning failures. To that end, we organize the content in terms of skills and abilities and associate them with a goal. These elements constitute the scale of our tool. Barr et al. (1975) indicate that the description of the curriculum in terms of skills and competencies and the selection of questions based on the capacity and shortcomings of each student allow the student to be simultaneously exposed to challenging and instructive situations. This fact makes assessment itself a part of the learning process.

Planning learning involves organization and sequencing of learning contents and the choice of learning evaluation methods. Nowadays, to improve our knowledge of mathematics learning, it is essential that we understand a key core practical concept that allows the teacher to reflect on the student's learning outcomes and can be used to guide a differential instruction: the concept of ability along with the concept of competence.

In Brazil, the National Curriculum Parameters (Parâmetros Curriculares Nacionais – PCN) are published by the Ministry of Education (MEC). For each level of education and each area of instruction, the curriculum guidelines and framework are defined in the PCN. The PCN establishes the basic competencies that will allow students to become subjects who are producers of knowledge and participants in the world of labor. Furthermore, they consider as an educational goal the students' personal development for the exercise of citizenship in a democratic context.

In his report on mathematics education in the United Kingdom, Cockcroft (1982) notices the role of the hierarchical nature of mathematics in the learning process and its relationship to the differences in achievement among students: "Mathematics is a hierarchical subject. This does not mean that there is an absolute order in which it is necessary to study the subject but that ability to proceed with new work is very often dependent on a sufficient understanding of one or more pieces of work which have gone before."

In Brazil, the national assessment agency is the Basic Education Assessment System (Sistema Nacional de Avaliação da Educação Básica – SAEB), and national assessments are undertaken by the National Institute of Educational Studies and Research (Instituto Nacional de Estudos e Pesquisas – INEP). In the late '90s, after the national curriculum standards were approved, INEP redesigned the reference matrix "based on descriptors involving two levels of specification: competences and abilities. The same assessment topics are used for all grades being assessed, but the priorities and levels of complexity are adjusted for the higher grades" (Ferrer, 2006, p. 65). In the SAEB 2001: Novas Perspectivas (SAEB 2001: New Perspectives; 2002) context, competence involves the ability to meet complex demands by thinking of multiple alternatives to solve given problems. Abilities are related to objective and practical aspects of "knowing how." The idea that students should develop certain cognitive competencies in the learning process and show some abilities based on these competencies guided the construction of the matrices. In this reference matrix, there are four core themes for mathematics: 1. space and shape; 2. magnitudes and measures; 3. numbers and operations/algebra and functions; and 4. data processing.

The development stage of our scale took into account PCN (BRASIL, 1998) and SAEB's reference matrix

(BRASIL, 2005). It was also necessary to rethink reference matrix descriptors and to propose new goals.

3.1 Our Scale

The development of our scale went through different stages. First, we selected the competencies to be analyzed in the area of geometry. Second, we searched the literature for studies of competencies and abilities in the Brazilian high school curriculum. In later stages, we observed aspects of the mathematical ability progression of the chosen area in the reference matrix of descriptors from SAEB (Brazil, 1997). The developed scale covered two themes of SAEB's array reference: space and shape, and magnitudes and measures. In Tables 1 and 2, we present two competencies (M4 and M5) and corresponding abilities associated with these themes.

Table 1. Space and shape and related abilities

M4- Use geometric knowledge to better understand and participate in the world in which they live	
H16	Recognize and understand phenomena presented in geometric language
H17	Recognize and understand geometric concepts in everyday life
H18	Apply geometric approaches in day-to-day problem solving
H19	Use geometric reasoning to pick from different problem-solving strategies proposed to solve a day-to-day problem
H20	Use geometric reasoning for evaluating the impact of proposed actions on day-to-day life

Competency M4 and corresponding abilities.

Table 2. Magnitudes and measures and related abilities

M5- Develop and extend magnitudes and measures notions and use them to understand the world in which they live and to solve day-to-day problems.	
H21	Recognize and interpret data using the standard system of measurement and its proper notation.
H22	Establish relationships between different systems of measurement and the modeling of natural and everyday life phenomena.
H23	Develop knowledge and understanding of different systems of measurement or units of measurement in day-to-day problem-solving.
H24	Demonstrate good reasoning in gathering and relating experimental measurements and estimates to understand their underlying phenomena.
H25	Use measures and estimates to recognize well-conceived actions in day-to-day

	life.

Competency M5 and corresponding abilities.

Thereafter, we designed goals that are related to those abilities. These goals resemble the descriptors that underlie the SAEB's mathematics assessments. They are hierarchically organized by grades, and levels of complexity are adjusted for the higher grades. Furthermore, they are basic requirements to develop the abilities they are related to. This structure is intended to enable a good interaction with the pool of items from which the test items were drawn in view of measuring the ability of interest. When designing the goals, we assumed they are intended to 9th grade. We have also considered related goals from 6th, 7th and 8th grades. Our scale has 30 goals from 9th grade, 35 goals from 6th, 18 from 7th and 22 from 8th grades (Anonymous, 2012).

Table 3: Selected goals in our scale and related abilities

Goals	
O39	Writing algebraic expressions to determine the area of a plane figure.
O66	Solving problems involving different area units of measurement and their relationship.
O72	Finding areas of similar figures.
O85	Solving problems involving areas of plane figures.

Some of the designed goals on our scale.

As Huang (1996) states, in DIA, the contents of mathematics curriculum for elementary schools are represented by a directed acyclic graph (see Figure 1 for a partial view), which reflects the dependencies between the contents.

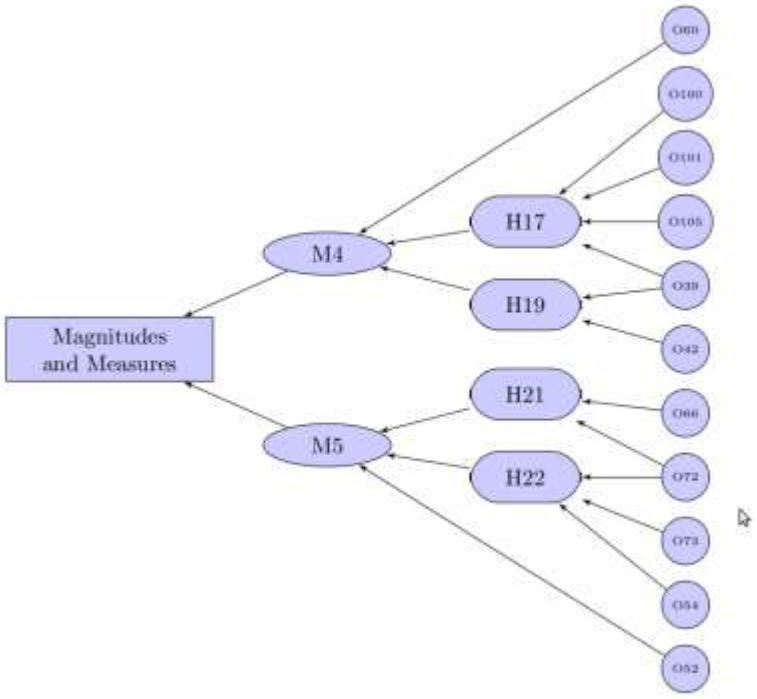

Figure 1. Contents of mathematics curriculum

Contents of mathematics curriculum represented by a directed acyclic graph.

The rectangle in the hierarchy represents one of the four themes of the reference matrix. Each ellipse represents a competency related to corresponding abilities. The designed goals describe and point out the basic and essential knowledge that can support the development of the related ability. For instance, to analyze the student's ability H22 – to establish relationships between different systems of measurement and the modeling of natural and everyday life phenomena, we indicate the goals O66 – of solving problems involving different area units of measurement and their relationship or even the goal O72 – of finding areas of similar figures.

Our scale is structured in four layers: Themes, Competencies, Abilities, and Goals. We emphasize that there is no solid hierarchy between concepts, nor between the goals. We do not exhaust the goals in the creation of this scale. We agree with Machado (1993) when he states " meaning is never definitively constructed. The bundle of relations that constitute it is continuously transformed, incorporating new relations or purifying others, which become less expressive."

3.2 Experiments

The experiments were carried out through simulations of respondents as well as their answers to the questions proposed by the algorithm. The probability of the student responding correctly to this item was calculated using the three-parameter logistic model. The probability of a correct response of an examinee at proficiency level θ to a dichotomous item is given by

$$P(U = 1|\theta_c) = \gamma + \frac{1-\gamma}{1+e^{-\alpha(\theta_c-\beta)}}.$$

Here U is the binary response of the examinee with latent ability level θ_c to an item with discrimination level α, difficulty level β, pseudoguessing level γ. U is coded as 1 for a correct response and 0 for an incorrect one. To simulate the response of a student of known ability θ_c to an item α, β and γ, a number $r, 0 < r < 1$ was generated randomly. If the probability $P(U = 1|\theta_c)$ was greater than or equal to r, it was assumed that the student would respond correctly to the item. Otherwise, the response was considered incorrect (Eggen and Straetmans, 2000, p. 14). For example, if there is a 0.70 correct response probability, $r = 0.75$ would produce an "incorrect" response. Responses were generated as each item is administered to the examinee in the simulation.

The study used a simulation of 23 students of 9th grade whose known ability θ_c were equally distributed in the interval $[-3,3]$. We assume that $\theta_c \in \{-2.75,2.5,\ldots,2.5,2.75\}$. Each virtual student answers a set of questions. If there is no error, it is presented to a new set. When there is an error, a process of refinement begins by offering questions associated with the objectives present in the questions incorrectly answered in the previous iteration. Within an assessment procedure, the CAT algorithm selects items based on the criterion of maximum information from a statistical perspective. This process repeats until an incorrect answer appears on an item with one or two objectives. At this point, a diagnosis is generated informing that the objective has not yet been fully developed or reached by the student. Each test comprises 25 iterations (Eggen and Straetmans, 2000).

3.2.1 Simulated Item Bank

Tests were performed using a simulated item bank. It was necessary to interweave the proposed goals to generate items that reflect "real" questions. Direito, Pereira, and Duarte (2010) state that the establishment of these relationships of dependence between learning objectives in a particular knowledge domain allows structuring the evaluation process and sequencing the teaching and learning processes. They also state that this approach is used in the development of several customized training proposals in technologically supported learning systems. There is always a possibility of creating new entanglements and new goals. In this way, we reiterate that there is no solid hierarchy between the concepts nor between the goals.

In our item bank, each question has at most 5 goals. To generate an item, we consider a goal O of the ninth grade and its k_O interlaced goals. We considered all combinations k to k of the interlaced goals with $k = 1, \ldots, \min(k_O, 4)$. Each of these combinations associated with goal O generates an item. In fact, each combination generated in the bank 13 questions. Each of these 13 questions has the same objectives but with difficulty parameter $\beta \in \{-3, -2.5, \ldots, 2.5, 3\}$. We set the discrimination parameter $\alpha = 1.2$. This value is the same as that proposed by Huang (1996) and corresponds to the mean value of the discrimination parameters of the item bank analyzed by Kingsbury and Weiss (1979). We assumed that all items are multiple-choice questions with five alternatives. As pointed out by Huang's (1996), since the random chance of success is one in five, we set the value of the pseudoguessing parameter $\gamma = 0.2$. To assign values to the parameters of difficulty, we made a simplifying hypothesis that all generated items occurred in the bank once with difficulty parameters $\beta \in \{-3, -2.5, \ldots, 2.5, 3\}$. That is, each of the 226 generated items occurs 13 times in the bank with different difficulties. In this way, we worked with an already calibrated item bank containing 2938 questions.

3.2.2 Analysis

We organize the simulation results in two graphs. In one of them, we present all tested goals. In the other, the result of the diagnosis. The graph proved to be very useful for presenting goals that are not yet achieved by the student. Goals of lower grades are also present since they are interlaced with ninth grade ones. Two schemes were created: Set of tested goals and diagnosis. In them, the goals for the 9th grade appear in orange. In pink, goals of the 8th grade, in green, 7th-grade goals and in blue color, 6th-grade goals.

We illustrate the feedback provided by DIA tool presenting the results of a student with known ability $\theta_c = -2.25$.

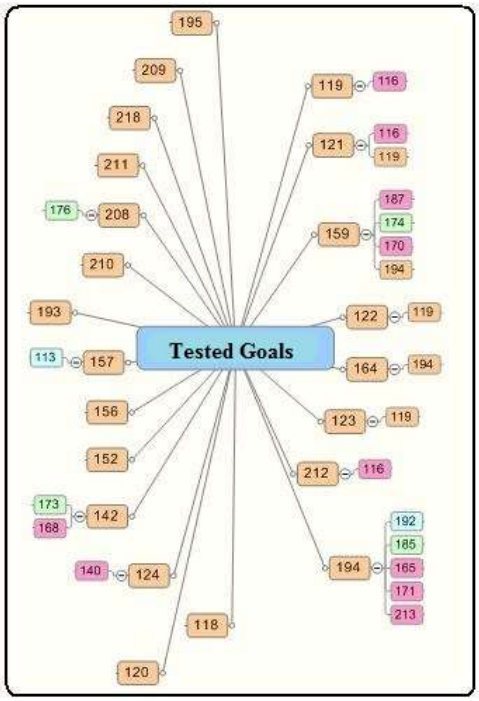

Figure 2. Tested Goals

Tested goals of a student with ability $\theta_c = -2.25$.

Figure 2 shows that during 25 iterations, a student with ability $\theta_c = -2.25$ answered questions that related to 24 goals of 9th grade and 16 interlaced goals: 10 of 8th grade, 4 of 7th grade, and 2 of 6th grade. Incorrected answers given during the test enable a diagnosis shown in Figure 3.

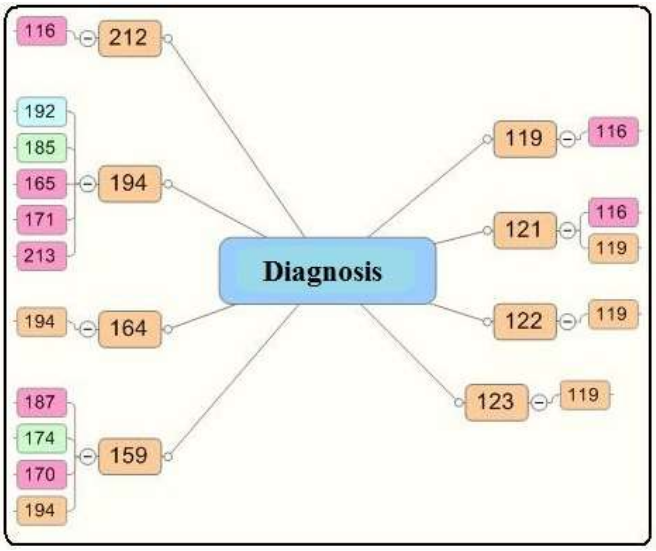

Figure 3. Diagnosis

Diagnosis of a student with ability $\theta_c = -2.25$.

In Figure 3, we see that 8 goals of 9th grade and 9 interlaced goals (6 of 8th grade, 2 of 7th grade, and 1 of 6th grade) were not achieved by the student. Instead of a long list of unachieved goals, the graph allows to connect them. It gives a path for solving learning gaps identified from previous grades or misunderstood

concepts.

4. Conclusion

Prevalent measures of knowledge in the educational system are based on averages, which highlight only those students who are well above or well below average. Several studies have indicated the positive effects of formative assessment. Among them, the surveys of Raabe (2005) and Pimentel and Omar (2006) use computer resources in decision making about the mediation regulation of learning, and a computer-assisted formative evaluation process. For Al-A'Ali (2007), the question is not whether the assessment should incorporate the use of technology but to do so responsibly, so as to preserve the validity, usefulness, and credibility of the results.

Comparing all tested goals and all those diagnosed in DIA tool, we concluded that the proposed work met expectations. The proposed scale and interlacing enabled dialogue between the scale and the item bank in order to allow the construction of a diagnosis. Through the diagnosis, the student has the possibility to observe which objectives were not yet reached, either by learning gaps identified from previous grades or by incomprehension of concepts tested.

One of the objectives of the DIA tool is to make teachers aware of the importance of goals that describe the content to be developed and feedback from appraisals of learning. The use of a tool such as DIA represents an alternative within the range of instruments and techniques available to the teacher and/or student to the evaluation of learning in a formative perspective.

5. Acknowledgement

The research is partially financed by

6. References

Anonymous, 2012.

M. Al-A'Ali, "Implementation of an Improved Adaptive Testing Theory," Educational Technology & Society, International Forum of Educational Technology & Society, USA, 2007, pp. 80-94.

A. Barr, M. Beart and R.C. Atkinson, The computer as a tutorial laboratory: The Stanford BIP project. Technical Report 260, Stanford Univ., CA. IMSSS, 1975.

Brazil. Ministério da Educação e do Desporto. SE Fundamental. PCN: Ciências da Natureza, Matemática e Suas Tecnologias. Ensino Médio. Brasília, DF: MEC/SEF, 1998.

Brazil. MEC/INEP, Matrizes Curriculares de Referência para o SAEB. Brasília, DF, 1997.

A.L. Brown, "Knowing when, where, and how to remember: A problem of metacognition," in R. Glaser (Org.), Advances in instructional psychology, Hillsdale, N.J.: Erlbaum, 1978, pp.77-165.

W.H. Cockcroft, Mathematics Counts. London: HMSO, 1982.

D.R. Costa, Métodos Estatísticos em Testes Adaptativos Informatizados. Dissertação. (Mestrado em Estatística; Dep. Métodos Estatísticos, IM, UFRJ), Rio de Janeiro, 2009.

I. Direito, A.M.S. Pereira and A.M.O. Duarte, "A representação do conhecimento e competências: contributos da psicologia cognitiva para sistemas de aprendizagem apoiados por computador", Actas do VII Simpósio Nacional de Investigação em Psicologia, Universidade do Minho, Portugal, 2010, pp. 2552-2560.

T. Eggen and G. Straetmans, "Computerized adaptive testing for classifying examinees into three categories", Educational and Psychological Measurement, 2000, p. 713-734.

G. Ferrer, "Educational assessment systems", in Latin America: Current practice and future Challenges. Washington, DC: Partnership for Educational, 2006.

R. Flaugher, "Item Pool", in Computerized adaptive testing: A primer. Hillsdale, NJ:Erlbaum, 1990.

C. Gipps, "Avaliação de alunos e aprendizagem para uma sociedade em mudança", in Proc. Anais do Seminário Internacional de Avaliação educacional. Brasília: INEP, 1998.

C. Hadji, "Compreender que avaliar não é medir, mas confrontar um processo de negociação", in Avaliação desmistificada. (Translated by Patrícia C. Ramos). Porto Alegre: Artes Médicas, 2001, pp. 27-49.

J. Heller, C. Steiner, C. Hockmeyerr and D. Albert, "Competence-Based Knowledge Structures for Personalised Learning", Int. J. on E-Learning, 2006, pp. 75-88.

S. X. Huang, "On Content-Balanced Adaptive Testing. In Computer Aided Learning and Instruction", Science and Engineering, 1996, pp. 60-68.

G. Kingsbury and D. Weiss, An Adaptive Testing Strategy for Mastery Decisions, Defense Technical Information Center, 1979.

M. Lilley, The Development and Application of Computer-Adaptive Testing in a Higher Education Environment. Unpublished PhD thesis, School of Comp. Sc., Univ. of Hertfordshire, Hertfordshire, 2007.

M. Lilley, T. Barker and C. Britton, "The development and evaluation of a software prototype for computer-adaptive testing", Comput. Educ., 2004, pp.109-123.

N. Machado, "Interdisciplinaridade e matemática", Pro-Posições, 1993.

P. Perrenoud, Construir competências desde a escola. (Translated by Bruno Charles Magne). Porto Alegre: Artmed, 1999.

E.P. Pimentel and N. Omar, "Métricas para o Mapeamento do Conhecimento do Aprendiz em Ambientes Computacionais de Aprendizagem", in Anais XVII Simpósio Brasileiro de Informática na Educação, Brasília, 2006, pp. 247-256.

A.L.A. Raabe, Uma proposta de arquitetura de Sistema Tutor Inteligente baseada na Teoria das Experiências de Aprendizagem Mediadas. Tese (Doutorado). Programa de Pós-graduação em Informática na Educação, UFRGS, Porto Alegre, 2005.

C. Ribeiro, "Metacognição um apoio ao processo de aprendizagem", Psicol. Reflex. Crit., Porto Alegre, 2003.

C. Ward, Preparing and Using Objective Questions, Handbooks for Further Education, Nelson Thornes Ltd, 1981.