Graduate Management Admission Council®

Adequacy of an Item Pool Measuring English Language Proficiency for Implementing CAT

Camila Akemi Karino

Denise Reis Costa

Cespe/University of Brasilia

Jacob Arie Laros

University of Brasilia

Presented at the CAT Research and Applications Around the World Poster Session, June 2, 2009

Abstract

The possibility of applying a different set of items according to the ability level of each respondent has stimulated, among other factors, an increasing use of computerized adaptive testing (CAT). The present study aimed to verify whether the psychometric quality of an item pool was sufficient to implement a procedure of CAT. The item pool was a database of the proficiency exam in English language used since 2004 by the University of Brasilia. The psychometric quality of the items was verified using both classical test theory and item response theory (IRT). The complete item pool consisted of 384 items which were applied in nine different test forms of 50 items each. Other than the first test form which consisted only of unique items, all other test forms consisted of 42 or 41 unique items and 8 or 9 common items. On average, each test form was administered to 330 students. The total number of respondents was 2,969. Each test was analyzed individually, and in a second stage the forms were calibrated jointly. In the individual analyses, 80 poorly discriminating items were eliminated. In the joint analysis, another 58 items with an *a* parameter less than .50 were eliminated. After the elimination of these items, the joint IRT analysis revealed a mean discrimination parameter of $a = .77$ (SD = .20). The remaining 312 items showed a range of *b* between −3.56 and 3.23. Nonetheless, the majority (75%) of the items were easy, with *b* parameter less than .10. The median value of the *c* parameter was .11 (SD =.04). The item pool was considered to be of sufficient psychometric quality to serve for the implementation of a CAT procedure, even though 36% of the items needed to be eliminated in order to satisfy pre-established psychometric criteria. The item pool will permit initial studies and should improve through repeated applications of the English exam using a CAT procedure.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Karino, C. A., Costa, D. R., & Laros, J. A. (2009). Adequacy of an item pool measuring proficiency in English language to implement a CAT procedure. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

**Camila Akemi Karino, Cespe/University of Brasilia,
QNC 08 HOME 29, Brasilia, Brazil, Email: camilaakarino@gmail.com**

Adequacy of an Item Pool Measuring English Language Proficiency for Implementing CAT

Computerized adaptive testing (CAT) is an assessment procedure which attempts to estimate an individual's proficiency interactively at the moment of test application. Based on the answers given by the individual, the most appropriate items are selected to effectively measure his/her proficiency or ability. In this way, each person answers a specific test that corresponds to his/her ability level.

The notion of a CAT is to emulate what a wise examiner would do (Wainer, 2000). That is, select for an examinee only those items that best measure his/her estimated ability level based on the answers that are being given by the examinee. This way, if the examinee responds correctly to an easy item, he/she receives a more difficult item; if the examinee incorrectly answers an item, an item a little easier than the previous one is presented. This process continues until a previously established precision of measurement has been reached. As opposed to traditional paper-and-pencil tests, each examinee can receive tests of different lengths. It is estimated that an assessment which uses a CAT procedure reduces the test length by about 50%, in addition to being a more precise measurement (Wainer, 2000).

Another benefit of CAT is that it can be less tiring, because the examinee does not need to answer items that inform little or nothing about their ability. Sands and Waters (1997) pointed out that presenting easy items to high ability examinees can be wearisome, because they are not challenging to these examinees. On the other hand, presenting many difficult items to low ability examinees can be frustrating. In both cases, the individuals might begin to answer items at random without much interest, which can contribute to errors in the estimation procedure.

Among other advantages of CAT are: flexibility to deliver test batteries in different places and times, the possibility of using multimedia resources in the construction of items, higher strictness in the control of application rules, simplification of the test scoring process and test security control (Sands & Waters, 1997). Once it is not known which items will compose the test, fraud becomes much more difficult.

Research with CAT began in the 1970s with Lord (1971) and Owen (1975). Currently, CAT has been used in many assessments: the Graduate Record Examination (GRE), Test of English as a Foreign Language (TOEFL) and also the Armed Services Vocational Aptitude Test Battery (ASBAV). In Brazil, this study is one of the first initiatives of implementing CAT.

Some theoretical studies have already been carried out in Brazil, for example, the dissertations of Oliveira (2002) and Gonçalves (2004) from the Institute of Ciências Matemáticas e de Computação, ICMC/USP. These studies focused on the development of CAT systems, but not a single implementation was made in a real assessment. The dissertation that most advanced the field of implementation of this technology in Brazil was developed by Costa (2009). The dissertation of Costa was the basis of the present study.

Certainly, the incipient development of CAT procedures in Brazil up to this moment is due to many difficulties. Three major difficulties can be mentioned in relation to the implementing of CAT: theoretical, operational and practical. CAT demands knowledge and domain of statistical

theories of very high complexity, one of which is item response theory (IRT), beyond the computing and construction of evaluation instruments domain.

There is an operational difficulty in the sense that for the implementation of a CAT procedure the following are needed: financial resources, development of statistical programs, and the construction of a proper item pool. Finally, the practical difficulty is related to explaining this kind of technology to society; many questions may be raised related to the fact that not everyone answers the same number of items and the items are different according to the examinee. Thus, in the end, it is not easy to explain to society that a procedure in which an examinee answered 10 items and the other answered 15 items is more precise for both persons than a procedure in which 40 items are administered to both persons. The difficulties of acceptance are even higher when the assessment is a test of selection and/or certification.

It will be a challenge to consolidate and to establish such a technology. There are advantages of this kind of testing and many social barriers to be broken. Having this in mind, this study is one of the first initiatives of implementing this technology in Brazil. Specifically, this study attempts to evaluate only one of the main requirements for CAT implementing: The Item Pool.

The Item Pool

The first step in the creation of an adaptive test consists in the development of an item pool. According to Flaugher (2000), the following is necessary for the construction of a good item pool: (1) a sufficient number of items for each level of ability which is being measured; (2) a pedagogical revision of the quality of the items, and (3) pre-testing of the items and psychometric item analysis. All these steps aim to ensure that the item pool will contain a wide variety of items, with diverse difficulties, and that it covers the complete matrix of the defined ability. Furthermore, the diversity of items is very important for the adaptation algorithm to have a higher possibility of item choice so that it better accomplishes its function.

Considering the importance of the item pool, a psychometric item analysis after the pre-test phase is fundamental to guarantee the quality of the pool. Two main theories were used for the analysis of pre-tested items: classical test theory (CTT) and item response theory (IRT).

The focus of CTT is the assessment of persons on basis of the results obtained in tests through scores (raw or standardized). That is, an examinee answers a set of questions and obtains a score, which is the sum of all correct answers. Through CTT, this score expresses the magnitude of what was intended to be measured of the person, that is, individuals with higher scores are more able in that knowledge area than individuals with lower scores. CTT has some limitations, for instance, being dependent of the particular group of items which compose the test. By contrast, IRT brings forth a new type of statistical analysis, centered on the items and not on the test as a whole.

IRT starts from the assumption that the examinee has a latent trait, which is an individual proficiency or ability characteristic that determines how the items in a test are answered. The latent trait has a probabilistic relation with each of the items. Consequently, IRT models the probability for an individual of giving the correct answer to an item based on the characteristics of the items and the latent trait of the examinee. Therefore, it is possible to compare individuals among themselves and to evaluate the quality of the items because the items and the individuals are expressed on the same measurement scale.

At the construction stage of the item pool, the main interest consists in evaluating the quality of the items based on theoretical considerations. As a result of these analyses, items are excluded, the quality of the item bank is evaluated, and the development and pretesting of new items is suggested.

Taking the above into consideration, this study aimed to verify whether the psychometric quality of the item pool was sufficient to implement an adaptive test procedure. The item pool used in this study is a database of the proficiency exam in English language that has been used since 2004 by the University of Brasilia.

Method

Context

The proficiency test in English was introduced in 2004 at the University of Brasília (UnB). The assessment takes place twice a year and has as its fundamental goal to introduce and guide the student to practice strategies of reading comprehension that favor more efficient reading of English Language texts. Any regular student of UnB can take the exam.

At the time of this study, the proficiency exam had been applied nine times to a total of 2,969 students. On average, 330 students were examined each semester. The exam is a paper-and-pencil test which is composed of 50 multiple-choice items.

Students take different tests each semester, although some common items are administered from one semester to another. It is the common items that make the equating among different tests possible. As quoted by Andrade, Tavares and Valle (2000), equating through common items means putting item parameters which come from partially distinct tests or groups of examinees from different populations onto a common scale, making the items and abilities comparable.

The ability of each student is measured on a scale with an average of 50 and standard deviation of 16. A student is approved with a score higher than 50 on this scale. It is noteworthy that, despite IRT being utilized since the first year of the exam, no procedure of group estimation among the items across the different periods has been implemented. Therefore, despite the fact that the different test forms contained common items, the analyses and the publishing of the results using IRT have been done separately each semester.

Item Pool

In this study, the database of the proficiency exam in English Language developed by the organization CESPE—Centro de Seleção e Promoção de Eventos ,which is a part of UnB, was used. This database is composed of all items used in the last nine proficiency exams.

The complete item pool consisted of 384 different items. Each test form was composed of 50 items with on average of eight common items among test forms. Table 1 shows the number of common items among the nine test forms.

Table 1. Composition of the Nine Test Forms of the Proficiency Exam

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9
Test 1	50	8	4	2	2	8	-	1	1
Test 2	-	42	4	2	-	-	-	-	-
Test 3	-	-	42	4	-	-	-	-	-
Test 4	-	-	-	42	6	-	-	-	-
Test 5	-	-	-	-	42	-	-	-	-
Test 6	-	-	-	-	-	42	8	8	6
Test 7	-	-	-	-	-	-	42	-	-
Test 8	-	-	-	-	-	-	-	41	2
Test 9	-	-	-	-	-	-	-	-	41
Total	50	50	50	50	50	50	50	50	50

Procedure of Analysis

Item analysis was based upon CTT and IRT. Each test was first analyzed separately by CTT and IRT, and in a second stage the nine test forms were analyzed jointly by IRT.

In the separate CTT and IRT analyses, the criterion for eliminating items was exclusively based on item discrimination using the biserial correlation and the IRT *a* parameter. Items were eliminated stepwise using the following criteria: (1) items with a negative biserial correlation; (2) items with a biserial correlation below .10; (3) items with a biserial correlation below .20; (4) items with an *a* parameter below .40; and, finally (5) items showing an *a* parameter below .50. These psychometric criteria were established based upon Baker (2001) and also based on frequent use in practice. It was decided to eliminate the items using this stepwise procedure because the elimination of the worst items could improve the quality of other items and, as a result, fewer items would be eliminated.

In the joint analysis, all items were considered new, submitted to different populations, with common items between the test forms (Andrade, Tavares, & Valle, 2000). As a result, an IRT model was used for multiple populations and dichotomous answers. The normal model was used and the estimation method was expectation a posteriori (EAP). All the item parameters were calibrated by the software BILOG-MG (Zimowski et al., 1996).

As a second step, an information function analysis of the test was implemented considering the application of 15 items for different ability levels of the scale. Additionally, the conditional standard error was also analyzed.

After the psychometric analysis of the items, a pedagogical analysis for the construction of a proficiency scale was implemented. With the joint analysis of IRT, the items of the several tests were placed on a common scale (with mean 0 and standard deviation 1) that has a pedagogic meaning. However, the arbitrary IRT metric (0,1) provided by the program might generate interpretation difficulties. As Valle (2001) indicated, an examinee with a score of zero might think that he/she has no knowledge in English, whereas the zero only indicates the average of the scale. the scale used previously by the exam had a mean of 50 and standard deviation of 16. This scale also had led in some cases to inadequate interpretations, for it suggests that the proficiency values are located on a 0 to 100 interval, which students commonly associated with the "percentage of correct items". In order to avoid such inaccuracies with the interpretation of the

common scale, a new proficiency scale in English was suggested, with 100 as the average and 25 as the standard deviation.

For the interpretation of the scale, a technique based on anchor items was utilized. At first, intervals that were selected by analysts for pedagogic interpretation were defined for the scale. After establishing these levels, the items that characterize each one of the levels (anchor items) were identified. After that, successive points of the scale were fixed to the anchor items and each of these levels was pedagogically interpreted by specialists.

The criteria used for the choice of the anchor items were the ones defined by Andrade, Tavares, & Valle (2000). Considering two consecutive anchor levels Y and Z, with Y <Z, a determined anchor item is from Z level if and only if it obeys three conditions simultaneously: (1) The probability of individuals who are on level Z answering the item correctly is equal to or greater than .65; (2) the probability of individuals who are on the previous level (Y) answering the item correctly is lower than .50 and (3) the probability of a correct response of individuals who are on level Z subtracted from the probability of a correct of individuals who are on level Y is greater than or equal to .30.

In practice, it is difficult to find items that satisfied all three conditions. This being so, the items were classified according to the quantity of criteria that are satisfied, with the best items for defining the level those that were capable of satisfying the three criteria.

Results

Separate Analysis

In total, 46 items with biserial correlations below .20, and 34 items with IRT *a* parameters less than .50 were eliminated. Table 2 shows the number of items eliminated by each analysis.

Table 2. Number of Eliminated Items By CTT and IRT Analyses By Test

Analysis	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9
CTT	8	5	3	0	9	7	5	3	6
IRT	7	4	3	0	4	3	5	5	3
Total	15	9	6	0	13	10	10	8	9

For Test 4, administered in the second semester of 2006, no items needed to be eliminated. In the individual analyses, the item selection criterion was based on the item discrimination criterion, as when items were eliminated by this criterion other parameters (difficulty and guessing) presented acceptable estimates of the standard error.

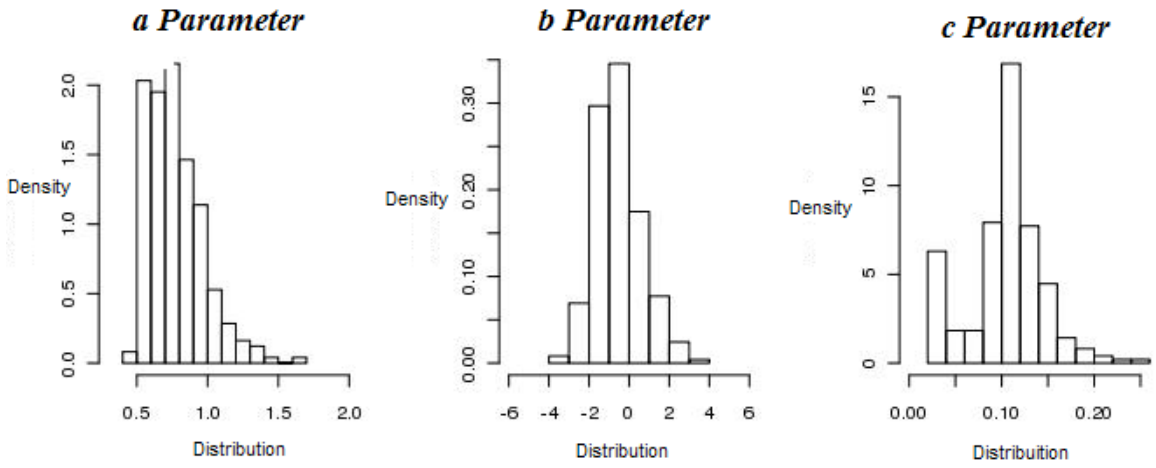
Joint Analysis

In the joint analysis, another 58 items were eliminated due to the criterion of discrimination parameter in the normal IRT metric less than .50. As in the separate analyses, the program was executed many times until the *a* parameter of all items was above the pre-established criterion.

After the elimination of the items, the final item pool (n = 246 items) had a mean *a* parameter of .77 (SD = .20), varying between .49 and 1.67. For the *b* parameter, a substantial variation in

difficulty level of the items was observed, varying between –3.56 and 3.23. However, the majority (75%) of the items showed a *b* parameter below .10. The median value of the *c* parameter was .11 (SD =.04) with an range of .03 to .24. Figure 1 shows the distribution of the three parameters of the items.

Figure 1. Frequency Distributions of the Three IRT Parameters



A good measure of the adequacy of the item bank can be obtained from the analysis of the test information function (TIF). The TIF is calculated based on the item parameters and permits analyzing how much information a test contains for each ability level. In Table 3, the measures of information and standard errors associated with a simulation of an adaptive test with 15 items for several ability values are presented. Once the test procedure is adaptive, the items that compose the test of each ability level are not necessarily the same.

Table 3. Information [I(θ)] and Standard Error [SE(θ)] for a 15-Item CAT at Various θ Levels

θ	I(θ)	SE(θ)
–4.00	1.22	0.91
–3.67	1.72	0.76
–3.33	1.84	0.74
–3.00	2.99	0.58
–2.67	3.92	0.50
–2.33	5.56	0.42
–2.00	6.00	0.41
–1.66	7.59	0.36
–1.33	9.83	0.32
–1.00	10.18	0.31
–0.66	10.64	0.31
–0.33	9.47	0.32
0.00	8.85	0.34
0.33	7.09	0.38
0.67	7.35	0.37
1.00	6.52	0.39
1.33	5.71	0.42
1.67	5.24	0.44
2.00	4.29	0.48
2.33	3.36	0.55
2.67	3.08	0.57
3.00	3.38	0.54
3.33	2.84	0.59
3.67	2.00	0.71
4.00	1.28	0.88

Table 3 shows that a 15-item CAT from the English proficiency bank had higher information for students of θ = –1.0 and –.66. That is, students at these θ levels would have their proficiency better estimated; this is because the value of information for these tests was higher than for other θs and, consequently the value associated with the standard error was the lowest found (.31). As expected, θs on the extremes of the distribution had the lowest information with the items of this bank.

Pedagogical Analysis

Successive points of the scale were fixed through anchor items and the levels were interpreted pedagogically by specialists. The scale was divided in intervals of 25 points and the anchor items of each interval were interpreted. Specialists interpreted the levels according to the content in the group of items that defined each anchor level. A summary of the interpretation of levels can be seen in Figure 2.

Figure 2. Interpretation of Scale Levels

Level 25	The student is able to produce inferences and basic deductions, started from ideas contained in the text, through cognates.
Level 50	The student on this level understands basic vocabulary of academic texts, but undergoes limitation regarding daily life vocabulary. Beyond that, the student recognizes main ideas and connects information between items and text. The student also analyzes paraphrases of textual information and establishes relations between the expressions contained in the items and in the text. The candidate makes inferences/deductions and identifies the text's theme. Also, the candidate can recognize main and secondary parts of the text.
Level 75	The student on this level understands basic/intermediate vocabulary of academic texts, but still undergoes limitation regarding daily life vocabulary. He/she recognizes synonym/antonym words or expressions and can relate them and/or adequately replace them in context. In addition, he/she identifies and locates secondary and main ideas in the text. There is, also on this level, the differentiation of facts and opinions and the establishment of relation between thesis and arguments given to sustain it. Inferences/deductions are present and are more accurate. The candidate knows how to recognize paraphrases.
Level 100	The student on this level understands intermediate vocabulary of academic texts, but undergoes some limitations regarding daily life vocabulary. The student establishes cause and consequence ideas between the parts and elements of the text. Beyond this, the student is able to define synonymy between parts of words, as suffixes and/or prefixes. Still, he/she is able to select an option that better sums the idea(s) of the text and the most appropriate title between the alternatives given.
Level 125	The student on this level understands intermediate vocabulary of academic texts, but doesn't undergo limitations regarding everyday vocabulary. He/she is able to recognize the effect of the usage of morphosyntatic resources. The student makes grammatical correspondence between sections of text.
Level 150	The student on this level understands intermediate/advanced vocabulary of academic texts, but doesn't have any limitations whatsoever regarding everyday vocabulary, the grammatical correspondence is more accurate and the deductions / inferences have a higher degree of complexity.

Level 175	The student on this level understands advanced vocabulary of academic texts, but he/she has doesn't have any limitations whatsoever regarding everyday vocabulary. The grammatical correspondence is accurate. Beyond that, the deductions/inferences have a superior degree of complexity.
Level 200	The student on this level understands advanced vocabulary of academic and everyday language texts

Discussion

This study was attempted to evaluate the quality of a Proficiency in English pool of items with the goal of evaluating its adequacy for the implementation of CAT. About 36% of the items of the initial bank needed to be eliminated in order to agree with pre-established psychometric criteria. Exclusively in terms of item quantity, the necessity of an increase in the number of items can be observed in order to augment the possibility of item choice for the adaptive procedure in order to accomplish its function better. The existence of a limited number of items can also generate a problem of over-exposure of an item.

After the necessary eliminations of items, the existence of a sufficiently satisfactory item pool in relation to the distribution of the IRT item parameters was verified. The distribution of the *c* parameter indicated that the majority of items had a value below .20, which means a low probability of pseudo-chance. The *a* parameters (average = .77) showed the existence of rather well discriminating items. The items covered a wide range of difficulties but the majority (75%) of the items had a *b* parameter below .10.

Flaugher (2000) indicated that a good item pool for adaptive testing should contain highly discriminating items (*a* parameter above .60), show a rectangular distribution of the difficulty parameter, and have a low probability of pseudo-chance (*c* parameter less than .20). According to these criteria, the item bank of English proficiency is appropriate in relation to the *a* and *c* parameters, but it not in relation to the *b* parameters: it would be necessary to include more items at the both extremes of the scale.

The analysis of IRT test information also supports this finding. Due to a greater quantity of items in the levels near of the average of the scale, the tests reached higher information levels and, hence, a lower estimation error. But there were higher standard errors toward the extremes of the scale.

Moreover, according to Pasquali (2003), the ideal level of difficulty for the items of a test depends on the purpose of the test. As it is desired that the proficiency test discriminate individuals who master the minimum required ability "Inglês Instrumental I" discipline, it is necessary to establish to which scale level this corresponds. Taking this definition as a starting point, and considering the availability of a large number of items, it is essential for a proper identification of the cutoff point. In the case of a certification examination , it is not important to evaluate how low the proficiency of the student is, nor how high it is. It matters only whether the student's θ estimate exceeds the minimum required to be certified.

To that same end, the construction of a common scale and, especially, the interpretation of the scale, are two major contributions of this study for the Proficiency Examination. As the areas

of each level of scale are known, setting the "cut off" point provides more theoretical and methodological consistency than the arbitrary criteria of midpoint of the scale hitherto used. Although it is still necessary to define the cut off score and new items need to be included in the item bank, the initial analysis of the item pool and the scale interpretation permit initial studies for the implementation of a CAT procedure. The item pool will undoubtedly improve by repeated applications of the English exam using the CAT procedure.

This study is of the utmost importance as a first step in the implementation process of a CAT procedure for the English proficiency exam. It is clear that even the best and most sophisticated CAT procedure does not work properly if the quality of the item pool is not sufficiently high.

References

Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). Item response theory: Concepts and application. Minas Gerais: SINAPE.

Baker, F. B. (2001). *The basics of item response theory* (second edition). ERIC Clearinghouse on Assessment and Evaluation.

Costa, D. R. (2009). Statistical methods at computarized adaptive test. *Master Thesis*, Federal University of Rio de Janeiro, Rio de Janeiro.

Flaugher, R. (2000). Item pool. In H. Wainer (Org.). *Computerized adaptive testing: A primer*. New Jersey: Lawrence Erlbaum Associates.

Gonçalves, J. P. (2004). The integration of computerized adaptive tests and computational environments of task for the learning of the Instrumental English. *Masters Dissertation*, University of São Paulo, São Carlos.

Lord, M. F. (1971). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement, 31*, 3-31.

Oliveira, L. H. M. (2002). Adaptive tests sensitive to the contents of the item pools: An application at exams of English proficiencies for graduate programs. *Masters Dissertation*, University of São Paulo, São Carlos.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

Pasquali, L. (2003). Psychometric theory of the tests in the psychology and in the education. Petropolis: Vozes.

Sands, W. A. & Waters, B. K (1997). Introduction to ASVAB and CAT. In W.A. Sands; B.K. Waters & J.R. McBride (Orgs), *Computerized adaptive testing: From inquiry to operation*. Washington: American Psychological Association.

Valle, R. C. (2001). The construction and the interpretation of the scales of knowledge: general considerations and a vision of what it is used in the SARESP. *Estudos em Avaliação Educacional*, 23.

Wainer, H. (2000). *Computerized adaptive testing: A primer*. New Jersey: Lawrence Erlbaum Associates.

Zimowski, M. F. et al (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software, Inc.