Employment of Item Response Theory to measure change in Children's Analogical Thinking Modifiability Test

Emprego da Teoria de Resposta do Item para medida de mudança no Children's Analogical Thinking Modifiability Test

Odoisa Antunes de **QUEIROZ**[1]
Ricardo **PRIMI**[2]
Lucas de Francisco **CARVALHO**[2]
Sônia Regina Fiorim **ENUMO**[3]

Abstract

Dynamic testing, with an intermediate phase of assistance, measures changes between pretest and post-test assuming a common metric between them. To test this assumption we applied the Item Response Theory in the responses of 69 children to dynamic cognitive testing Children's Analogical Thinking Modifiability Test adapted, with 12 items, totaling 828 responses, with the purpose of verifying if the original scale yields the same results as the equalized scale obtained by Item Response Theory in terms of "changes quantifying". We followed the steps: 1) anchorage of the pre and post-test items through a cognitive analysis, finding 3 common items; 2) estimation of the items' difficulty level parameter and comparison of those; 3) equalization of the items and estimation of "thetas"; 4) comparison of the scales. The Children's Analogical Thinking Modifiability Test metric was similar to that estimated by the TRI, but it is necessary to differentiate the pre and post-test items' difficulty, adjusting it to samples with high and low performance.

Uniterms: Dynamic assessment; Rasch model; Item Response Theory.

Resumo

Provas assistidas, com fase intermediária de ensino, medem mudanças entre pré-teste e pós-teste pressupondo uma métrica comum entre eles. Para testar este pressuposto, aplicou-se a Teoria de Resposta ao Item nas respostas de 69 crianças à prova cognitiva assistida Children's Analogical Thinking Modifiability Test *adaptada, com 12 itens, totalizando 828 respostas, para verificar se a escala original produzia os mesmos resultados em termos de quantificação de mudança que a escala equalizada obtida via Teoria de Resposta ao Item. Seguiram-se os passos: 1) ancoragem dos itens de pré e pós-teste, por uma análise cognitiva, encontrando-se três itens em co-*

▼ ▼ ▼ ▼ ▼

[1] Universidade Federal do Espírito Santo, Centro de Ciências Humanas e Naturais, Programa de Pós-Graduação em Psicologia. Vitória, ES, Brasil.
[2] Universidade São Francisco, Programa de Pós-Graduação em Psicologia. Itatiba, SP, Brasil.
[3] Pontifícia Universidade Católica de Campinas, Centro de Ciências da Vida, Pós-Graduação em Psicologia. Av. Jonh Boyd Dunlop, s/n., Prédio Administrativo, Jd. Ipaussurama, 13059-900, Campinas, SP, Brasil. *Correspondência para*/Correspondence to: S.R.F. ENUMO. *E-mail*: <sonia.enumo@puc-campinas.edu.br>.

Article based on the dissertation of the O.A. QUEIROZ, intitled "*Proposta alternativa para análise do desempenho em provas cognitivas assistidas*". U*niversidade Federal do Espírito Santo*, 2010.

Acknowledgments: We would like to thank the following people for providing us with the research date Ana Cristina B. Cunha (*Universidade Federal do Rio de Janeiro*), Christyne G. T. Oliveira (*Faculdades Salesianas de Vitória*), Margareth R. Santa Maria-Mengel (*Centro Universitário de Franca*) and Maria Beatriz M. Linhares (*Universidade de São Paulo, Ribeirão Preto*).

mum; 2) estimação do parâmetro de dificuldade dos itens e comparação destes; 3) equalização dos itens e estimação dos "thetas"; 4) comparação das escalas. A métrica do Children's Analogical Thinking Modifiability Test *foi semelhante à estimada pela Teoria de Resposta ao Item, mas é preciso diferenciar a dificuldade dos itens de pré e pós-teste, adequando-o a amostras com alto e baixo desempenho.*

Unitermos: *Avaliação assistida; Modelo de Rasch; Teoria de Resposta ao Item.*

The Dynamic Assessment (DA) is a methodology that emerged in the 70s of the twentieth century. It combines assessment and intervention, aiming at measuring human abilities, especially the "learning potential" (Haywood & Tzuriel, 2002; Lidz, 1987). For that, as well as the traditional evaluation, it uses several procedures in the information collection about its object of study, highlighting the dynamic testing.

The dynamic testing differs from traditional psychometric test in three aspects: (a) it is involved with the process rather than with the product, so it uses the subject as its own control; (b) it can be applied in two ways: as a structured method (test-teach-retest), with systematic help in the teaching or assistance phase (the examinee receives feedback on the performance and gradual tips for the problem solution), or as a clinical method, in which help is offered freely (item by item), but targeted to the needs of the learner, and (c) the process is interactive and it has the participation of the examiner and the examinee (Sternberg & Grigorenko, 2002).

There has been an increasing use of the DA in Brazil as from the 90s of the last century, in which we have been verifying that it is possible to measure the existence of learning potential in children with and without Special Educational Needs (SEN) (Cunha, Enumo & Canal, 2011; Enumo, 2005; Linhares, Escolano & Enumo, 2006). However, there are some functional problems related to the change quantification between the pre-test and post-test phases, both unassisted, usually taken by the rate of gain between those phases (Linhares et al., 2006). The main criticisms are: (a) lack of trustworthiness in the differences between the pre-test and post-test scores, where the decreasing correlation between the two measures of the same trait may indicate poor reliability in at least one of the two measures; (b) The presence of the ceiling effects in the sample subgroups is almost null, that is, some participants show a variation in performance close to zero when compared to the pre and post-test phases (c) the scale nature in which change is measured is not well understood, and (d) the problem of equivalence of pre and post-test grades, because many times they are made up of different items aiming to quantify the change. Although the phases consist of different items they must be anchored in the same metric to be able to measure the change, but this is not always explicitly treated (Embretson, 1987; Haywood & Tzuriel, 2002; Sternberg & Grigorenko, 2002).

Thus, seeking a new quantification form of the dynamic testing results, this study examined, through Item Response Theory (IRT), children's cognitive performance in Children's Analogical Thinking Modifiability (CATM) test (Tzuriel & Klein, 1990).

Item Response Theory is a mathematical model, which, unlike the Classical Test Theory (CTT), assesses the probability that a subject respond correctly to a test item, according to the item parameters and its ability (latent trait) and not the total final score of the match (Andrade, Tavares & Valle, 2000; Pasquali, 2009). Therefore, subjects with different abilities show distinct probabilities of issuing a correct answer to the item. This functional relation is represented in the Item Characteristic Curve (ICC) (Baker, 2001; Muñiz, 1997).

The ICC is a mathematical function monotonically increasing that has in its abscissa the values related to the variable ability (theta, θ) and in the ordinate, the probability of setting the item $P(\theta)$ (Baker, 2001; Muñiz, 1997). The use of the ICC for the representation of mathematical models is unlimited, but only three are more applied - they are the logistic models: a) one parameter, which evaluates only the item difficulty; b) two parameters - difficulty and discrimination, and c) three parameters - difficulty, discrimination and correct answer at random (Muñiz, 1997; Pasquali & Primi, 2007).

In this study, we will use the One-Parameter Logistic Model (1LP), whose equation is presented below (Muñiz, 1997, p.38):

$$P_{ij}(\theta j) = \frac{e^{D(\theta j - bi)}}{1 + e^{D(\theta j - bi)}}$$

Where:

Pi(θ) *j*= probability of the subject, with ability θj set the item i

θj= the subject's ability *j* *bi*= item difficulty *i*

D= adjustment constant equal to 1.7 e = mathematical constant equal to 2.72

In One-Parameter Logistic Model the subject's response to the item depends on their *ability* (θj) and on the item *difficulty (bi),* and therefore the higher the ability, the greater the *probability* of success. Otherwise, considering the *ability* (θj) constant, the greater the item *difficulty (bi),* the lower the *probability* of the subject to solve it (Muñiz, 1997; Pasquali, 2007).

The models proposed by IRT take two fundamental assumptions: (a) the unidimensionality and (b) local independence (Muñiz, 1997; Pasquali & Primi, 2007). The unidimensionality is implicit in the model formulations, which seeks to predict the response probability considering only the subject's ability and the item difficulty (in case of the Rasch model), that is, the response probability is predicted by the difference between those two parameters, expressed in only one dimension. In practice, this implies that the model will only work if one dominant dimension explains most of the response probability to the item.

To accept this premise, although we understand that human behavior is not governed by a single latent trait, we established the existence of a *dominant aptitude* (Pasquali, 2007; Primi, 2004). As for local independence, IRT assumes that the answer to an item (correct or not), on the basis of a particular ability (θj), does not interfere in the responses given to the other test items, that is, the items are independent (Pasquali, 2007; Primi, 2004).

An important application of IRT is the creation of equivalent scales. In classic psychometrics the final metric of the test is based on the number of hits. However, it will vary with the number and difficulty of the items, so that, for example, the same numerical value of 10 hits does not mean the same thing in two different tests. In IRT, by the functions of the ICC, we can create the Characteristic Curve of the Test (CCT) by simply adding the ICC of the items that make it up. The CCT, as the ICC, will show the relation between the test score *t* and scale of theta. If we have two different tests on the difficulty through the CCT plotted together, we can convert the original metric (based on the number of hits) in a common metric for theta. The CCT illustrates an inherent property of the IRT, which, for not using the metric of the total score, but basing on the relation between hit and theta, naturally creates the possibility of equating grades of different tests that measure the same construct. We call this procedure equalization (Andrade et al., 2000; Wolfe, 2004).

Generally, the dynamic tests use the total score, and by the difficulty control and the number of items assume the existence of a common metric between the pre-and post-test phases. However, they rarely use IRT to equate the scores or to test this assumption, to see if, in fact, they get an equivalent scale (Sternberg & Grigorenko, 2002). Furthermore, in these situations it would be ideal that the post-test be harder than the pre-test, since we expect the subjects to modify in the intervention. However, this situation creates another problem for the equation of the grades.

Faced with these problems, this study applied IRT to equate the grades of a dynamic testing, in order to compare the metric obtained via IRT with the original metric, checking if the scales assumed as similar are so indeed. This is intended to contribute to improve the psychometric aspects of the dynamic assessment procedures.

Method

This study was not forwarded to the Ethics Committee because we only used data from researches already completed (period 1999-2008) and published, which had been previously approved by the Psychology Post Graduate Program from the University. However, the consent forms were signed by those responsible for the children, according to the standard 196/96 of the Ministry of Health. Therefore, in this work, besides the evaluation of ethical issues by the Examining Committee, we only asked for the authorization from the research group coordinators to use the material.

Data

We used data from three researches on dynamic assessment in the country (Cunha et al., 2011; Oliveira, 2008; Santa Maria & Linhares, 1999), which included responses from 69 children on the dynamic cognitive testing CATM adapted, with 12 items, total of 828 results. Participants had a mean age of 8.5 years old (5-12 years), and differed according to prematurity and low birth weight (34), learning disabilities (29) and visual impairment (6), and attending Elementary School (27), Kindergarten (39), special class (3), and one that did not attend to school at all.

Dynamic cognitive testing

In these researches we used a version adapted by Santa Maria and Linhares (1999) for the CATM (Tzuriel & Klein, 1990) - they applied the CATM, with modifications in the original test including another phase, of transfer, in order to verify the generalization of the child's performance, consisting of more complex items than the original test. The CATM is a nonverbal test that evaluates the analogical reasoning (A : B :: C : ?), in which the child must indicate the solution (C : D relation). It contains cards with tasks and pads with geometric figures with the attributes: shape (triangle, circle and square), color (blue, yellow and red) and size (smaller and larger). The adapted version has 32 questions organized into four phases, in an increasing order of difficulty, depending on the number of the response attributes (Board 1 only requires the recognition of one attribute; gradually, until the end of phase two, attributes will be required, and so on). For this study, we used data from 22 problems of each research, divided into three phases: pre-test (6 items) - assistance (10 items) - post-test (6 items) (Santa Maria & Linhares, 1999).

Procedure

Initially, we tried to equalize the pre-and post-test items. For this, there must be some kind of "anchor", from which we fix the metric scale (Wolfe, 2004). This anchor can be based on: (a) people's ability, in case the same group is responding to the two tests; a situation in which the average of the group ability is the anchor or (b) the difficulty of the items, if there are common items in the pre and post-test phases, when *b's* (difficulty indexes), in this case, become the anchor.

However, since the items of the pre and post-test are apparently different, we may consider them similar from the demand point of view of the underlying cognitive processes. In more recent literature on cognitive processes in analogical reasoning items, such as in this case, these demands are called factors of complexity and are responsible for the explanation of the difficulty indexes of the items (Embretson, 1994; Primi, 1998, 2002). Therefore, if it were possible to find items with the same demand of cognitive processes (characterized by the same set of complexity factors) between the pre and post-test phases, it would be possible to consider them as common items which could, thus, be used as an anchor.

After the construction of the CATM database in Excel®, we proceeded to a cognitive analysis of the pre-test boards (6 boards) and post-test (6 boards), to establish the difficulty level of the items on these phases. Then, from the relation between figures (A : B :: C : ?), we quantified the different attributes (color, shape and size) involved in the relations AB and AC. For the variable AB, we assigned value zero when there was no change in the attribute from stimulus A to B; value 1 when there was a change in the attribute; and value 2, when there were two changes. We used the same logic for the variable AC. This procedure was performed on the six items of the pre and post-test phases, which allowed us to find four common items.

After this step, we tested empirically whether items with the same cognitive demand could be indeed considered identical. For this, we estimated the parameters of difficulty of the pre-and post-test items, separately, as if they were different subtests through Winsteps® - Rasch Model Analysis Software focusing on the difficulty on the difficulty scale of the items. This procedure is consistent with the idea that the difficulty levels of pre-and post-test are equivalent, allowing the comparison of *b's* from the common items. Afterwards, we compared the difficulties of these items considered as common, and we verified they were the same, as predicted by cognitive analysis.

Next, we went up to the second stage of the analysis, which was the equalization itself, that is, putting the pre and post-test grades in a common metric. This procedure could be done in two ways: 1) we could estimate the difficulty indexes of the items and the subjects' abilities in the pre-test and then fix the index values of the equivalent items in the post-test and estimate the difficulty indexes of the remaining

Items and the subjects' abilities in the post-test, or 2) we could do the reverse procedure, starting with the post-test and setting the items in the pre-test. Option number 2 was more effective because it showed little variation in the *b's* when they were fixed (displacement), except for one item, which caused its abandonment as a common item, leaving only three common items to the equalization.

Continuing with the analysis, we performed the estimating of the subjects' abilities in both phases of CATM. From this, we compared the grades estimated by IRT with the traditional test grades in order to verify the main question of this article, that asks whether it would be correct to consider the scale of hits as equivalent, that is, if the original scale of CATM produces the same results in terms of change quantification as the equalized scale obtained via IRT.

Results

First phase: Definition of structural identity due to the factors of the item complexity

We analyzed the cognitive variables of the items of pre-and post-test phases, that is, the relation between stimulus A and B (AB) and between stimulus A and C (AC), the sum of AB and AC variables (information number) and the synthesis of the three prior variables (AB and BC), that is, [AB + AC + (AB and BC)]. With this analysis, we found that the following pairs of items can be considered to be identical from the viewpoint of cognitive demand: pre_1/post_2; pre_2/post_4, pre_4/post_3; pre_5/post_5.

After finding the equivalent items between the pre and post-test phases, we estimated the difficulty of these items separately in the pre and post-test, focusing on the difficulty of the items (b). We found that the difficulties reported were the same in both sets. From this, the difficulty indexes of pre-and post-test items were correlated with the factors of complexity, seeking to verify to what extent it is possible to predict them through the indicative variables of the cognitive demands underlying to the items.

The correlations between complexity factors and difficulty index were all positive (from 0.37 to 0.93) and two of them significant (due to the small number of items - 12, only very high magnitudes show significance). An important result is that the variable information number that summarizes the three previous [AB + AC + (AB and AC)], practically foresee the difficulty of the items, indicating that the more difficult a test item is, there is also growth in the variables that represent the AB with AC sum and the total amount of changes between the item stimulus. Separating only the four pairs of the difficulty indexes of the items considered cognitively identical, the correlation between their difficulty indexes is equal to 0.84. In summary, these results support the use of the complexity structure as a class definer of items with the same difficulty. Therefore, although the pre and post-test do not share common items, the items previously shown can be considered equivalent in terms of cognitive point of view.

Second phase: Equalization data of pre and post-test grades using the common items

Once we found the equivalent items in the pre and post-test phases, we started to estimate the subjects' parameters, testing the two procedures previously described in the process: 1) we could estimate the difficulty indexes of the items and the subjects' abilities in the pre-test and then fix the indexes values of the equivalent items in the post-test and estimate the difficulty index of the remaining items and the subjects' abilities in the post-test, or 2) we could do the reverse procedure, starting from the post-test and setting the items in the pre-test. After testing the two processes, we found that, in the second, the difficulty estimations were more accurate, because of the greater proximity of the item difficulty with the subjects' ability.

The difficulty of the items ranged from -3.17 to 2.14, suggesting variability in the level of ability requirement from the respondents to the items. In relation to the adjustment indexes, two items in the pre-test (pre_5 and pre_6) and one post-test item (post_1) showed infit and outfit indexes above the expected, according to the parameters suggested by Linacre and Wright (1994), that is, lower than 1.20. Furthermore, the item-total correlation observed for the post-test items was greater than or equal to 0.21 for five of the six items, which can be considered adequate. For the pre-test items, we obtained indexes greater than or equal to 0.13 for all the items, except for one item (pre_6), which showed negative magnitude. This indicates that this item does not contribute to the construct evaluation and that it is impairing the reliability of the instrument. The subjects' theta average in the pre-test, after equalization, was clearly lower than theta average in the post-test phase, suggesting that the difficulty level of the set of items in the pre-test phase is not adequate to assess the initial stage of the process (Table 1).

To test this assumption we performed a descriptive analysis of the difficulty indexes and adjustment of the items (Table 2). The results complement the data indicating that the difficulty amplitude was similar for the two sets of items, however, there are more items with difficulty superior to the level of theta from the participants in the pre-test in relation to the post-test phase.

We have made the equalizing of the scores, so we were able to compare the participants' grades in situations of pre and post-test to verify if the metric of the raw score (total score) was similar to that score obtained via IRT (theta), which was confirmed - the mean difference is positive, indicating a gain in the post-test. This gain is significant, as we can see in the analysis of the test t for paired samples (t=5.9, p<0.001, for theta, and t=6.4 p<0.001 for raw score) (Table 3).

We can view these comparisons between metrics in Figures 1 and 2. In the Figure 1 we show the Characteristic Curves of the tests on pre-test (crosses) and post-test (circles) phases, showing the conversion of the raw score to theta. In the graph we can see that the pre and post-test are, as a matter of fact, very similar in terms of difficulty. The curves practically coincide, meaning that, for example, a raw score 3, both pre and post-test equates the same result in theta.

Table 1

Descriptive statistics of the respondents theta values and data fit indices *(infit and outfit)*

Phase		*n*	θ	Standard Error	Infit	Outfit
Post-Test	Average	6	-0.07	1.21	1.01	1.17
	SD	0	1.92	0.26	0.19	0.81
	Maximum	6	4.00	1.95	1.25	2.86
	Minimum	6	-3.95	1.04	0.27	0.22
Pre-Test	Average	6	-1.61	1.32	0.99	1.16
	SD	0	1.60	0.34	0.85	1.76
	Maximum	6	2.26	2.04	2.91	9.90
	Minimum	6	-4.27	0.97	0.25	0.14

Note: SD: Standard Deviation.

Table 2

Descriptive statistics of the indices of difficulty (b) and adjustment of the items (Infit and outfit)

Phase		*n*	b	Standard Error	Infit	Outfit
Post-Test	Average	69	0	0.35	1.01	1.17
	SD	0	1.68	0.03	0.19	0.81
	Maximum	69	2.14	0.40	1.25	2.86
	Minimum	69	-2.49	0.31	0.70	0.51
Pre-Test	Average	69	-0.03	0.39	1.14	1.26
	SD	0	1.65	0.09	0.34	0.63
	Maximum	69	2.14	0.56	1.75	2.46
	Minimum	69	-3.17	0.32	0.69	0.53

Note: SD: Standard Deviation.

Table 3
Comparisons between metrics theta and total score

	n	Minimum	Maximum	*M*	*SD*
Theta_pre-test	69	-4.27	2.26	-1.61	1.61
Theta _post-test	69	-3.95	4.00	-0.07	1.93
Theta_difference	69	-3.04	8.27	1.54	2.17
Pre-Test total score	69	0.00	5.00	1.67	1.18
Post-Test total score	69	0.00	6.00	2.96	1.57
Total score_difference	69	-2.00	6.00	1.29	1.66

Note: M: Mean; SD: Standard Deviation.

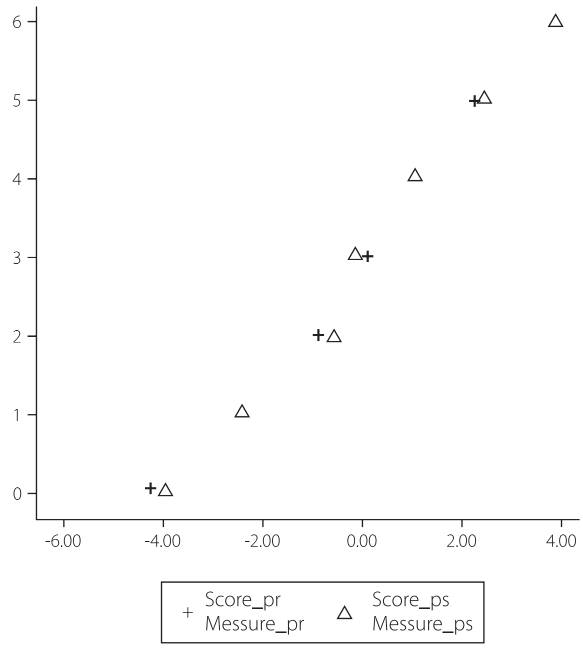

Figure 1. Characteristic curves of the test.

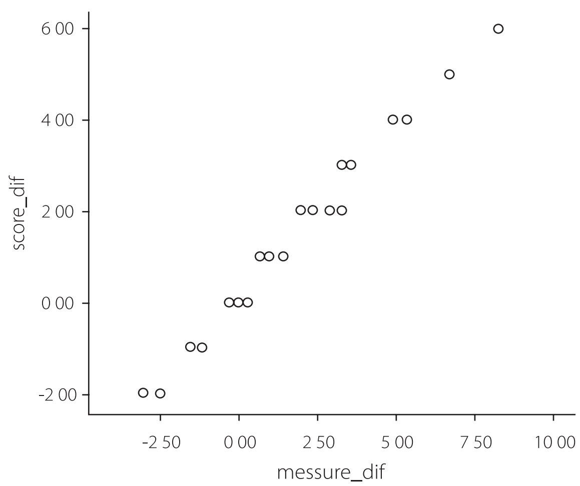

Figure 2. Correlations between theta and total score.

The Figure 2 shows the correlation between the scores of difference calculated by the raw score with the same score calculated from theta. The correlation between them is very high (r=0.99, p<0.001). These comparisons certify that the metric of the scores of differences that measure gains practically bring out the same information.

Discussion

The analysis made by the IRT allows us to conclude that the metric based on CATM raw score is practically the same that was estimated by the IRT, and so we can use the form proposed by CATM (number of hits) to measure the performance change between the pre and post-test.

The application of IRT also showed that the fact that the items of the CATM pre and post-test phases have the same average difficulty creates a problem in the measurement accuracy in the pre-test. The low average of theta in the pre-test phase shows that the difficulty level of the set of items from that step is not adequate to the process initial measurement. The analysis pointed the need of pre and post-test having items with different difficulty levels, with easy items in the pre-test phase (before the mediator's assistance) and difficult items in the post-test phase (after assistance). This arrangement and the use of items in the same metric could help in comparing samples with low and high performance capacity, because easy tests in the pre-test and post-test could benefit low capacity subjects, while difficult tests would favor only subjects with high performance capacity, that is, the use of one of the two ways would generate a problem for the change evaluation. Now, the composition with easy items in the pre-test and difficult items in the post-test would imply distinct raw score metrics.

This methodology does not reduce the effectiveness of CATM and it opens the way to make adaptive assessment of dynamic testing, that is, by knowing the parameters of the items, we can choose from the examining profile the most appropriate items to measure its learning potential. Moreover, it would contribute to reducing the "cost" of the test application, using easy items for children with learning difficulties or disabilities, and more difficult items for "normal" children, who do not have these problems. The comparison is possible, since all the subjects may be put on the same scale, and be quantified by the use of the raw score.

The problem of the change quantifying in dynamic testing is frequent in the works from this area, but it has not prevented the use of DA. The IRT has already been emphasized by researchers of Psychometrics (Embretson, 1987; Sternberg & Grigorenko, 2002) as a tool able to solve these issues, however, there is still much to learn about its use in dynamic testing.

References

Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). *Teoria da Resposta ao Item*: *conceitos e aplicações*. São Paulo: ABE.

Baker, F. B. (2001). *The basics of Item Response Theory*. Washington, DC: Eric.

Cunha, A. C. B., Enumo, S. R. F., & Canal, C. P. P. (2011). Avaliação cognitiva psicométrica e assistida de crianças com baixa visão moderada. *Paidéia*, *21*(48), 29-39.

Embretson, S. E. (1987). Toward development of a psychometric approach. In C. S. Lidz (Ed.), *Dynamic assessment*: *An interactional approach to evaluating learning potential* (pp.141-170). New York: The Guilford.

Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment*: *A multidisciplinary perspective* (pp.107-135). New York: Plenum.

Enumo, S. R. F. (2005). Avaliação assistida para crianças com necessidades educacionais especiais: um recurso auxiliar na inclusão escolar. *Revista Brasileira de Educação Especial*, *11*(3), 335-354.

Haywood, H. C., & Tzuriel, D. (2002). Applications and challenges in dynamic assessment. *Peabody Journal of Education*, *77*(2), 40-63.

Lidz, C. S. (Ed.). (1987). Historical perspectives. In *Dynamic assessment*: *An interactional approach to evaluating learning potential* (pp.3-32). New York: The Guilford.

Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(2), 370.

Linhares, M. B. M. Escolano, A. C. M., & Enumo, S. R. F. (Orgs.). (2006). *Avaliação assistida*: *fundamentos, procedimentos e aplicabilidade*. São Paulo: Casa do Psicólogo.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Madrid: Ediciones Pirâmide.

Oliveira, C. G. T. (2008). *Indicadores cognitivos, lingüísticos, comportamentais e acadêmicos em pré-escolares prematuros e nascidos a termo* (Dissertação de mestrado não-publicada). Programa de Pós-Graduação em Psicologia, Universidade Federal do Espírito Santo, Vitória.

Pasquali, L. (2007). Os modelos da Teoria de Resposta ao Item - TRI. In L. Pasquali (Org.), *Teoria de Resposta ao Item - TRI*: *teoria, procedimentos e aplicações* (pp.29-52). Brasília: UnB.

Pasquali, L. (2009). Psicometria. *Revista da Escola de Enfermagem USP*, *43*(Esp.), 992-999.

Pasquali, L., & Primi, R (2007). Fundamentos da Teoria da Resposta ao Item - TRI. In L. Pasquali (Org.), *Teoria de Resposta ao Item-TRI*: *teoria, procedimentos e aplicações* (pp.11-28). Brasília: UnB.

Primi, R. (1998). *Desenvolvimento de um instrumento informatizado para avaliação do raciocínio analítico* (Tese de doutorado não-publicada). Programa de Pós-Graduação em Psicologia Escolar, do Desenvolvimento e da Personalidade, Universidade de São Paulo.

Primi, R. (2002). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence*, *30*(1), 41-70.

Primi, R. (2004). Avanços na interpretação de escalas com a aplicação da Teoria de Resposta ao Item. *Avaliação Psicológica*, *3*(1), 53-58.

Santa Maria, M. R., & Linhares, M. B. M. (1999). Avaliação cognitiva assistida de crianças com indicações de dificuldades de aprendizagem escolar e deficiência mental leve. *Psicologia*: *Reflexão e Crítica*, *12*(2), 395-417.

Sternberg, R. J., & Grigorenko, E. L (2002). *Dynamic testing*: *The nature and measurement of learning potential*. New York: Cambridge University Press.

Tzuriel, D., & Klein, P. S. (1990). *The Children's Analogical Thinking Modifiability Test*: *Instruction manual*. Ramat-Gan: School of Education Bar Ilan University.

Wolfe, E. W. (2004). Equating and item banking with the Rasch model. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement*: *Theory, models, and applications* (pp. 366-390). Maple Grove: Jam Press.

Received on: 16/11/2011
Approved on: 5/11/2012