# PROPOSTA DE UM ESQUEMA DE ANOTAÇÃO PARA OS ITENS DE TESTES ADAPTATIVOS INFORMATIZADOS BASEADOS NO CBAT-2

## 1. Introdução

Um Teste Informatizado (TI) é um teste que avalia eletronicamente os conhecimentos e habilidades de um indivíduo por meio de itens, que medem os atributos mentais do indivíduo. Para Osterlind (1998), os itens de teste não podem ser chamados de questões, pois um item pode assumir outros formatos, que não são necessariamente interrogativos. Um exemplo são itens *cloze*, criados por Taylor (1953), cuja resposta é dada por um preenchimento de palavra(s) em uma frase.

O trabalho de Scalise e Gifford (2006) introduz a taxonomia *Intermediate Constraint Taxonomy for E-Learning Assessment Questions and Tasks*, que contempla 28 tipos de itens quanto ao formato de resposta. A partir da revisão de 44 artigos e capítulos de livros, os autores classificam os itens para testes informatizados em: (i) Múltipla Escolha, (ii) Seleção/Identificação, (iii) Reordenação/Rearranjo, (iv) Substituição/Correção, (v) Completamento, (vi) Construção e (vii) Apresentação/Portfólio.

Em avaliações educacionais de larga escala, geralmente os testes são de múltipla escolha e os itens são dicotômicos, ou seja, dentre as opções de resposta, apenas uma é a correta (pré-determinada) e as demais são chamadas de distratoras.

Visando avaliações educacionais aplicadas com maior flexibilidade e adaptabilidade, com significativa redução do tempo, com resultados imediatos e maior precisão em relação a testes que apresentam um número fixo de itens, o Teste Adaptativo Informatizado (TAI) seleciona os itens ao examinado segundo o histórico de itens respondidos anteriormente (Olea et al., 1999). Dessa forma, cada examinado pode receber um elenco diferente de itens, que podem variar em quantidade.

Um TAI é minimamente composto por um Banco de Itens (BI) calibrado, um método para estimar as habilidades dos examinados, um critério de seleção de itens, um critério para a seleção do primeiro item e um critério de parada do teste. A Figura 1 mostra o esquema geral de um TAI. O examinado inicia o teste e um critério inicial é aplicado para a seleção do primeiro item. O examinado responde ao item e sua habilidade estimada é calculada, conforme o modelo de resposta adotado. Aplica-se o critério de parada do TAI que, se satisfeito, termina o teste. Caso contrário, um novo item é selecionado e o ciclo é reiniciado.

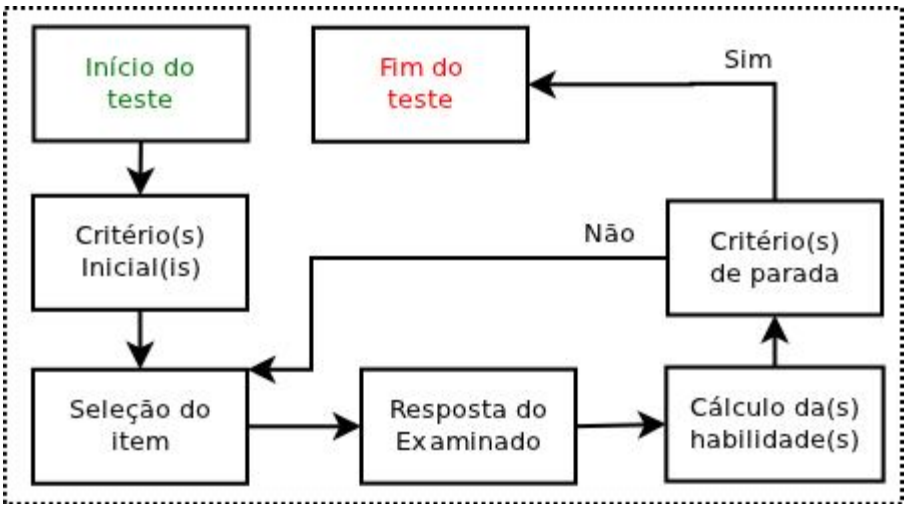


**Figura 1. Esquema de um TAI.**

Quanto ao BI de um TAI, um item deve conter, minimamente, (i) o enunciado, (ii) a questão/afirmação, (iii) as opções/campo de resposta, (iv) as informações psicométricas numéricas, (v) categorizações/classificações que rotulam o item e (vi) o identificador (ID) do item. Weiss (1985) afirma que um TAI que administra de 15 a 20 itens para o examinado, requer pelo menos 100 itens para se ter um teste com boa precisão. Já para Olea et. al. (1999), o banco deve ser dez vezes maior que o número de itens aplicado para um examinado.

Na abordagem psicométrica, necessita-se calibrar o BI, principalmente se o TAI for baseado na Teoria de Resposta ao Item (TRI) (van der Linden e Pashley, 2002). A TRI modela a probabilidade de um aluno responder a um item corretamente em função de sua habilidade (ou proficiência) estimada. Assim, quanto maior a proficiência, maior a probabilidade do aluno acertar a resposta ao item. A TRI é utilizada em exames em avaliações formais, tais como o Exame Nacional do Ensino Médio (ENEM), o Exame Nacional para Certificação de Competências de Jovens (ENCCEJA), Prova Brasil, Sistema de Avaliação da Educação Básica (SAEB), *Test of English as a Foreign Language* (TOEFL), *Graduate Record Examination* (GRE), *Programme for International Student Assessment* (PISA), dentre outros.

A TRI permite a comparação entre diferentes etapas de aplicação do teste, como por exemplo, ser aplicado diversas vezes ao ano. Essa comparação é possível quando uma escala de conhecimentos/habilidades do examinado é determinada. Para Pasquali e Primi (2003), o desempenho do examinado em um item pode ser predito a partir de um conjunto de fatores ou variáveis hipotéticas, ditos aptidões, traços latentes ou habilidades; expressos pelo parâmetro $\theta$, que é interpretado como a “causa” e o desempenho o “efeito”.

O modelo de resposta mais aplicado em TAI baseado na TRI é o Modelo Logístico de Três Parâmetros (ML3P) (Lord, 1980; Baker, 2001), que possui os seguintes parâmetros associados: (i) parâmetro discriminação *a*, que relaciona-se com a inclinação da curva do gráfico gerado pelo

ML3P, (ii) parâmetro dificuldade *b* e (iii) parâmetro "chute" ou acerto aleatório *c*. A calibração dos parâmetros *a*, *b* e *c* pode ser realizada sob duas abordagens (*omitido devido à revisão cega1*): (i) a Calibração Estática, que é realizada a partir de uma grande amostra de examinados que responderam aos itens em um pré-teste e (ii) a Calibração Dinâmica, realizada em tempo real no TAI. O procedimento de Calibração Estática dos itens é um procedimento caro e trabalhoso, pois necessita-se de um grande número de examinados para determinar os parâmetros para cada item do banco. Visando minimizar a necessidade de pré-calibração do BI, Huang (1996) propôs o algoritmo *Content-Balanced Adaptive Testing* (CBAT-2), que elimina a calibração prévia de itens do teste. A medida que o examinado responde os itens e termina o teste, o algoritmo recalcula a dificuldade do item baseado no ML3P. O CBAT-2 é (*omitido devido à revisão cega1*) (i) uma ótima solução para testes que necessitem de um BI de pequeno porte e aplicável em contextos em que a aplicação de pré-testes seja inviável, (ii) um teste de processamento computacional muito rápido e robusto e (iii) a estimação do parâmetro θ do examinado segue os modelos/métodos da TRI.

No CBAT-2 os parâmetros de item *a* e *c* do ML3P são fixados, respectivamente, em *1.2* e *1/z*, em que *z* é o número de opções de resposta para itens dicotômicos. Por exemplo, para um item de múltipla escolha com quatro opções de resposta, *z=0.25*. Além dos parâmetros da TRI, o algoritmo possui outros parâmetros associados: *W, R* e Φ. *W* e *R* são, respectivamente, o número de vezes que o item foi respondido correta e incorretamente, dentre todos examinados que responderam-na. Φ é chamado de dificuldade acumulada do item.

A revisão da literatura apresenta apenas uma proposta de anotação de BI para TRI, no melhor do nosso conhecimento, disponibilizada pela organização IMS *Global Learning Consortium*, e que é concebida para a TRI e não para o TAI. Diante desse cenário, esse artigo propõe um esquema de anotação de corpus de itens em XML para testes adaptativos baseados no CBAT-2, visando apoiar o desenvolvimento e aplicação de testes informatizados, facilitando o intercâmbio de dados e a disponibilização pública para utilização em *benchmarks*.

A Seção 2 trata de um ambiente computacional desenvolvido que utiliza e disponibiliza um BI implementado sob um esquema de anotação de corpus de itens para um sistema tutor inteligente que prepara estudantes de pós-graduação quanto aos conhecimentos sobre gênero de textos científicos em inglês. A Seção 3 traz as conclusões desta pesquisa.

## 2. Proposta de um esquema para bancos de itens baseados no CBAT-2

Com o objetivo de apoiar o estudante de pós-graduação em suas leituras e escritas no gênero de textos científicos em inglês, e também prepará-lo para a avaliação formal do Exame de Proficiência em Inglês (EPI), foi desenvolvido no ICMC-USP o ambiente computacional de aprendizagem

*Computer-Aided Learning of English for Academic Purposes* (CALEAP-Web). O CALEAP-Web foi desenvolvido a partir da integração de um TAI baseado no CBAT-2 e um ambiente computacional que possui tarefas de aprendizagem elaboradas a partir das convenções do gênero de textos científicos (*omitido devido à revisão cega2*).

O ambiente de avaliação é o *Adaptive English Proficiency Test for the Web* (ADEPT), que avalia o estudante de pós-graduação de maneira formativa (Haydt, 1988) em diferentes habilidades linguísticas, distribuídas em quatro módulos dentro do conhecimento de inglês acadêmico a saber: (1) convenções da língua inglesa para artigos científicos, (2) estrutura esquemática de artigos científicos, (3) compreensão de texto e (4) estratégias de escrita do gênero em questão. O BI é composto por 191 itens distribuídos em 99 itens do Módulo 1, 60 itens do Módulo 2, 18 itens do Módulo 3 e 14 itens do Módulo 4; formando um corpus para testes adaptativos com propósitos de avaliar o conhecimento do gênero de textos acadêmicos. A Figura 2 mostra um item tipo *testlet*, que depende de um único enunciado, de forma que os itens associados formem um subconjunto de itens do BI.

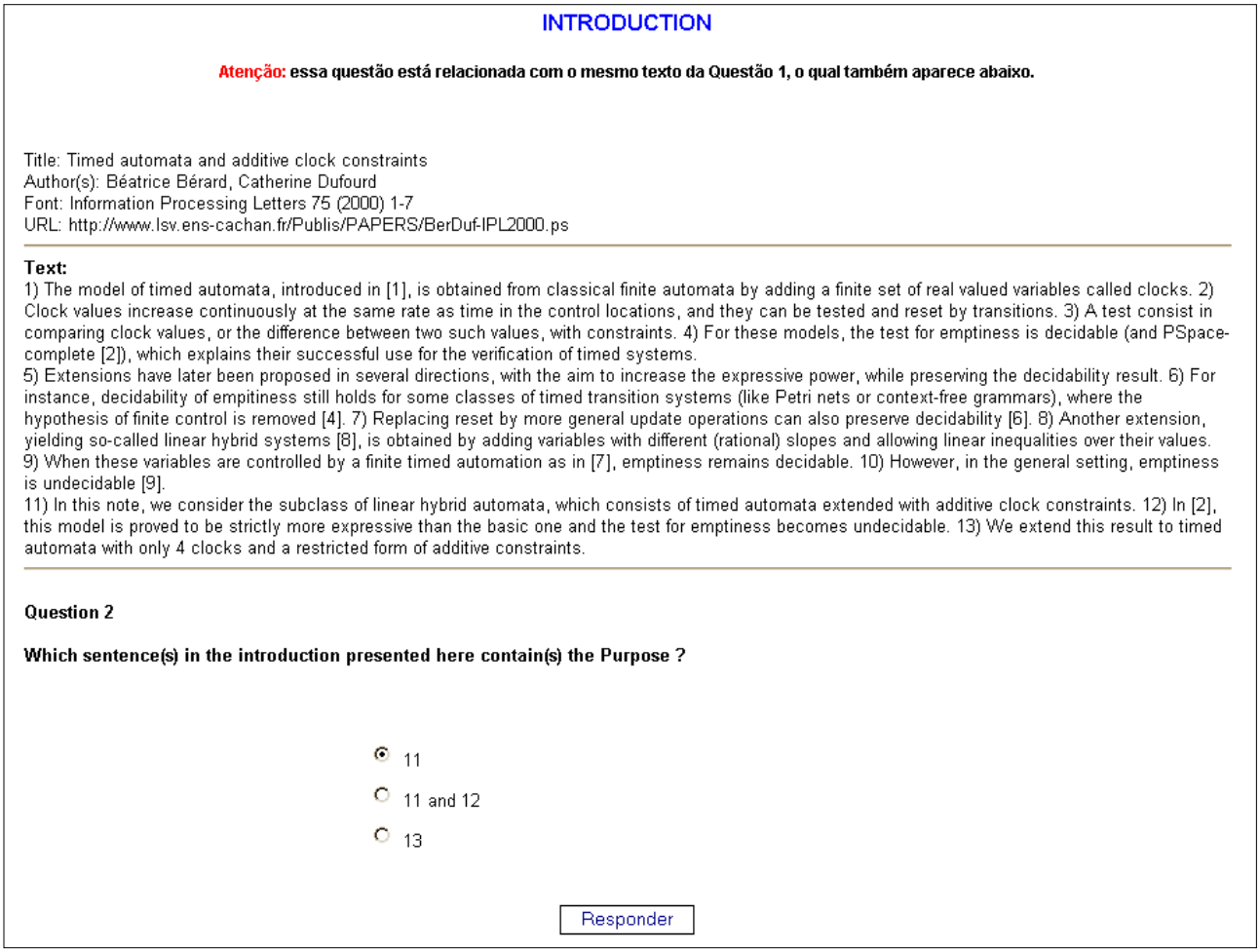


**Figura 2. Tela de um item tipo *testlet* do ADEPT.**

A Figura 3 mostra o esquema XML instanciado para um item do Módulo 1. O marcador (<showuser>) denota as informações que serão exibidas para o examinado, que são o enunciado (<wording>), a questão/afirmação (<question>) e as opções de resposta <answer>, que marca a resposta correta. O marcador <values> denota os dados psicométricos do item, que são (i) os parâmetros de item da TRI, *a* (<a-par>), *b* (<b-par>) e *c* (<c-par>), e (ii) os valores inerentes ao CBAT-2, que são o número de vezes que o item foi respondido (<exposed>), o número de vezes que o item foi respondido incorreta (<wrong>) e corretamente (<right>) e o valor Φ (<phi>).

```
<?xml version="1.0" encoding="UTF-8"?>
<item id="23" is-testlet="no">
<module>gap</module>
<showuser>
 <wording><text>_______the mouse is indisputably a good device for 2D
interaction, it performs only adequately in 3D tasks. </text></wording>
 <question><text>Which is the word that fills in the blank of the
example?</text>
 </question>
 <answer correct="0"><text>But</text></answer>
 <answer correct="1"><text>While</text></answer>
 <answer correct="0"><text>However</text></answer>
</showuser>
<values>
 <a-par>
  <fixedfloat>1.2</fixedfloat>
 </a-par>
 <b-par>
  <float>-0.972526</float>
 </b-par>
 <c-par>
  <fixedfloat>0.333333</fixedfloat>
 </c-par>
 <exposed>
  <integer>20</integer>
 </exposed>
 <wrong>
  <integer>34</integer>
 </wrong>
 <right>
 <integer>10</integer>
 </right>
 <phi>
 <float>1.299028</float>
 </phi>
</values>
</item>
```

**Figura 3 – Esquema XML de um item para o CBAT-2.**

O ADEPT mostra-se como uma possível solução para instituições que possuem um BI pequeno, e desejam obter o máximo de informação quanto à proficiência do estudante em determinado domínio de conhecimento *(omitido devido à revisão cega1).*

## 3. Conclusões

Para aplicação de avaliações educacionais informatizadas com maior flexibilidade e adaptabilidade, com significativa redução do tempo, com resultados imediatos, o TAI seleciona os itens ao examinado segundo o histórico de itens respondidos anteriormente. Os itens do TAI baseado no CBAT-2 é composto por um BI calibrado enquanto o teste ocorre, eliminando a calibração prévia. Nesse contexto, a revisão da literatura apontou para apenas uma proposta de anotação de BI para a TRI, indicando a ausência de anotação para itens na abordagem CBAT-2 e TAI, o que culminou na proposta de uma anotação de corpus de itens em XML para testes baseados no CBAT-2. O esquema de anotação proposto foi utilizado com todo o BI do ambiente computacional ADEPT, composto por 191 itens sobre o gênero de textos científicos em inglês, e encontra-se disponível para download. Tanto a disponibilização quanto a anotação de corpus de itens para CBAT-2 visam apoiar o desenvolvimento e aplicação de testes informatizados, facilitando o intercâmbio de dados e a disponibilização pública para utilização em *benchmarks*.

## Referências


BAKER, F. The Basics of Item Response. Second. College Park, MD: ERIC Clearinghouse on Assesment and Evaluation, University of Maryland, 2001.

HAYDT, R. C. Avaliação do Processo Ensino-Aprendizagem. São Paulo, Brasil: Editora Ática, S.A., 1988.

HUANG, S. X. On content-balanced adaptive testing algorithm for computer-based training systems. ITS-Intelligent Tutorial Systems, Montréal, Canada, p. 12–14, jun 1996.

LORD, F. M. Application of Item Response Theory to Practical Testing Problems. first. Hilsdale, New Jersey, EUA: Lawrence Erlbaum Associates, 1980.

OLEA, J., PONSODA, V., PRIETO, G. Tests Informatizados Fundamentos y Aplicaciones. [S.l.]: Ediciones Pirámede, 1999.

Omitido devido à revisão cega1. Omitido devido à revisão cega2.

OSTERLIND, S. J. Constructiong Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats. [S.l.]: Kluwer Academic Publishers - New York, Boston, Dordrecht, London, Moscow, 1998.

PASQUALI, L.; PRIMI, R. Fundamentos da Teoria de Resposta ao Item – TRI. Avaliação Psicológica, vol 3, p. 99-110, 2003.

SCALISE, K., GIFFORD, B. Computer-based assessment in e-learning: A framework for constructing intermediate constraint questions and tasks for technology platforms. JTLA (Journal of Technology, Learning and Assessment), Volume 4, n. Number 6, June 2006.

TAYLOR, W. L. Cloze procedure: a new tool for measuring readability. Journalism Quarterly, v. 30, p. 415–433, 1953.

van der LINDEN, W. J., PASHLEY, P. J. Item selection an ability estimation in adaptive testing. In: LINDEN, W. J. van der; GLAS, G. A. W. (Ed.). Computerized Adaptive Testing: Theory and Practice. University of Twente, The Netherlands: Kluwer Academic Publishers, 2002. p. 1–25.

WEISS, D. J. Adaptive testing by computer. Journal of Consulting and Clinical Psychology, v. 53, n. 6, p. 774–789, 1985.