Testes Adaptativos Computadorizados baseados na Teoria de Resposta ao Item em Sistemas *e-learning*: uma revisão sistemática da literatura

Victor M. G. Jatobá[1], **Karina V. Delgado**[1], **Jorge S. Farias**[2], **Valdinei Freire**[1]

[1]Escola de Artes Ciências e Humanidades – Universidade de São Paulo (USP)
Caixa Postal 03.828-000 – São Paulo – SP – Brazil

[2]Departamento de Ciências exatas e da Terra – Universidade do Estado da Bahia (UNEB)
Caixa Postal 41.150-000 – Salvador – BA – Brazil.

{victorjatoba,kvd,valdinei.freire}@usp.br, jfarias@uneb.br

Abstract. *Computerized Adaptive Testing (CAT) based on Item Response Theory (IRT) allows more accurate assessments and shows advantages when incorporated into e-learning environments. However, due the solutions multiplicity, this knowledge is fragmented in the literature, what makes it difficult to understand the challenges faced in the area. In this sense, the goal of this work is to understand and to characterize e-learning systems based on CAT and IRT. To this, a Systematic Review of the literature was performed where four research questions were analyzed in 9 papers selected by the defined protocol criteria.*

Resumo. *Testes Adaptativos Computadorizados (CAT) baseados na Teoria de Resposta ao Item (IRT) permitem realizar avaliações mais precisas e mostram vantagens quando incorporados a ambientes de aprendizado virtual (*e-learning*). Entretanto, devido à multiplicidade de soluções, este conhecimento encontra-se fragmentado na literatura, o que dificulta o entendimento dos dissensos e desafios enfrentados na área. Neste sentido, este trabalho buscou entender e caracterizar os estudos que usam CAT e IRT que estão inseridos em sistemas* e-learning. *Para tal, uma revisão sistemática da literatura foi realizada, na qual quatro questões de pesquisa foram analisadas em 9 artigos previamente selecionados pelos critérios estabelecidos no protocolo de revisão.*

1. INTRODUÇÃO

A etapa de avaliação de estudantes sempre foi muito importante no processo de aprendizagem. Na computação, os sistemas de aprendizado geradores de testes ajudam os estudantes a saberem se atingiram o nível adequado de conhecimento aprendido [Guzmán and Conejo 2004]. Um exemplo desse tipo de sistema é o Teste Adaptativo Computadorizado (do inglês, *Computerized Adaptive Testing* – CAT).

CATs são testes administrados por computadores que, de forma eficiente, reduzem o número de itens (questões) mantendo um melhor diagnóstico do desempenho do respondente [Kovatcheva and Nikolov 2009, Spenassato et al. 2016]. No CAT clássico, inicialmente é selecionada uma questão e, a cada nova, é estimado o nível de proficiência do estudante. Caso o critério de parada não seja atendido, outra questão é selecionada.

DOI: 10.5753/cbie.sbie.2017.273

Existem diversas formas de avaliar o nível de proeficiência do respondente. Um modelo bastante utilizado atualmente é o da Teoria de Resposta ao Item (do inglês, *Item Response Theory* – IRT) [Lord 1980]. Esta teoria é composta por modelos matemáticos que procuram estabelecer a probabilidade de um respondente qualquer acertar uma determinada questão, dadas as características do item e as habilidades do avaliado [de Andrade et al. 2000].

Três modelos logísticos são comumente utilizados na IRT, para calcular a probabilidade de acerto. Eles diferem entre si no número de parâmetros que consideram para descrever um item. São eles: (i) o modelo Rasch, de uma dimensão, que considera apenas o parâmetro de dificuldade do item (ML1) [Rasch 1960]; (ii) o modelo com os parâmetros de dificuldade e discriminação do item (ML2), originado na década de 1950 [Baker and Kim 2004]; e (iii) o modelo com três parâmetros (ML3), que acrescenta a probabilidade de acerto ao acaso ao ML2 [Birnbaum 1968].

CAT, baseado na IRT, permite fazer testes mais precisos [Kovatcheva and Nikolov 2009], pois é possível identificar as áreas de carência do estudante e, assim, selecionar uma sequência de itens adaptada ao conhecimento do respondente [Chen et al. 2005].

A IRT também pode ser aplicada em ambientes de aprendizado virtual (*e-learning*) e, nesse caso, podem herdar as mesmas vantagens do uso em CAT [Wauters et al. 2010]. Entretanto, ambientes de aprendizado possuem objetivos diferentes em relação aos de ambientes de teste, como, por exemplo, otimizar a eficiência do aprendizado, e não medir com maior precisão o nível de proficiência do avaliado [Wauters et al. 2010]. Além disso, o conhecimento sobre as diferentes soluções baseadas em CAT e IRT em sistemas *e-learning* encontra-se fragmentado na literatura, o que dificulta o entendimento dos dissensos e desafios enfrentados na área.

Assim, o objetivo deste trabalho é realizar uma Revisão Sistemática da Literatura (RSL), para caracterizar os **estudos que usam CAT e IRT que estão inseridos em sistemas *e-learning***, considerando as técnicas de adaptação implementadas, os métodos propostos nos testes adaptativos, o arquétipo de IRT usado e os tipos de experimentos adotados.

Este texto é composto por cinco seções. A Seção 2 descreve o método de pesquisa. Posteriormente, a Seção 3 apresenta os resultados obtidos e a Seção 4, as ameaças à validade desta RSL. Por fim, na Seção 5 são realizadas as discussões e as considerações finais.

2. MÉTODO

Uma revisão sistemática é uma forma de estudo secundário que utiliza uma metodologia bem definida para identificar, analisar e interpretar todas as evidências disponíveis relacionadas a uma questão de pesquisa específica, de maneira exaustiva, imparcial e, até certo grau, repetível [Kitchenham 2012]. A seguir, são apresentadas as três fases para o desenvolvimento de uma RSL, definidas em [Kitchenham 2012] e que são utilizadas neste trabalho: planejamento, condução e elaboração do relatório com os resultados obtidos.

2.1. Planejamento

Questões de pesquisa. Foram definidas quatro questões de pesquisa e cada uma delas foi decomposta em questões mais específicas. As questões são:

Q-1 Quais são as técnicas de adaptação empregadas nos sistemas *e-learning*?
- Q-1.1 O sistema preocupa-se em proporcionar *feedbacks* para o usuário?
- Q-1.2 O sistema faz sequenciamento de materiais de aprendizado?

A Questão 1 é importante, pois os sistemas de aprendizado adaptativo podem ser classificados de acordo com as técnicas de adaptação implementadas [Wauters et al. 2010].

Q-2 Quais são as características dos CATs?
- Q-2.1 Quais métodos são utilizados para selecionar o próximo item?
- Q-2.2 Quais são os critérios de parada adotados?

Já a Questão 2 buscou identificar dois aspectos fundamentais dos questionamentos-chave, levantados por [Wainer et al. 2000], que estão presentes na construção de um CAT.

Q-3 Quais são as características do uso da IRT?
- Q-3.1 Quais são os métodos usados para estimar os parâmetros do item?
- Q-3.2 Quais são os métodos usados para estimar as habilidades latentes?
- Q-3.3 Qual é o modelo logístico escolhido para a probabilidade de acerto?
- Q-3.4 Qual é o modelo de resposta (dicotômico ou politômico) considerado?

A Questão Q-3 visa caracterizar a IRT utilizada em ambientes de aprendizado virtual.

Q-4 Como são construídos e validados os sistemas *e-learning* baseados em CAT e IRT?
- Q-4.1 Qual é o número de itens e aprendizes utilizados nos experimentos?
- Q-4-2 Quais ferramentas e linguagens de programação são adotadas?
- Q-4.3 A solução utiliza algum *benchmark* ou base de dados para teste?
- Q-4.4 Os testes são realizados com dados reais ou artificiais?
- Q-4.5 Quais métricas são utilizadas para avaliar a qualidade da sequência de itens e da sequência curricular?
- Q-4.6 Os métodos são validados sobre quais tipos de cursos ou aplicações?

Seleção de Fontes e String de Busca. Uma vez que o Scopus indexa várias bases de dados, dentre elas: IEEE Xplore[1], Science Direct[2], ACM Digital Library[3] e Engineering Village[4], os motores de busca utilizados foram o Scopus [5] e o Web of Science (WoS)[6] com as seguintes strings:

Scopus: *TITLE-ABS-KEY (("item response theory"OR irt OR "Item calibration"OR "estimation") AND ("computer adaptative testing"OR "computer-adaptative tests"OR "Computer Adaptive Testing"OR "Computerized adaptive testing"OR CAT OR "Intelligent Testing"OR "Item banks"OR "Adaptive algorithms") AND ("Web-based educational system"OR "Web-based learning"OR "e-learning"OR "Web based Learning"OR "E-learning"OR "Student*

[1]http://ieeexplore.ieee.org
[2]http://www.sciencedirect.com
[3]http://dl.acm.org
[4]http://www.engineeringvillage.com
[5]https://www.scopus.com
[6]http://apps.webofknowledge.com

assessment"OR "Item-based learning"OR "Distance education"OR "learning strategies"OR "assessment tools"))
Web of Science: *TS=(((((((("item response theory") OR (irt)) OR ("Item calibration")) OR ("estimation")) AND ((((((("computer adaptative testing") OR ("computer-adaptative tests")) OR ("Computer Adaptive Testing")) OR ("Computerized adaptive testing")) OR (cat)) OR ("Intelligent Testing")) OR ("Item banks")) OR ("Adaptive algorithms"))) AND ((((((((("Web-based educational system") OR ("Web-based learning")) OR ("e-learning")) OR ("Web based Learning")) OR ("E-learning")) OR ("Student assessment")) OR ("Item-based learning")) OR ("Distance education")) OR ("learning strategies")) OR ("assessment tools")))*

Critérios de Inclusão e Exclusão. Os critérios de inclusão (CI) e os de exclusão (CE) considerados nesta RSL são:

- CI-1 Estudos contendo as palavras-chave: IRT, CAT e ambientes de aprendizado, assim como qualquer um de seus sinônimos.
- CE-1 Estudos não revisados por pares, tais como relatórios técnicos e capítulos de livros.
- CE-2 Estudos não disponíveis integralmente na Web ou disponíveis na Web, porém não acessíveis de forma gratuita via instituição dos autores da RSL.
- CE-3 Estudos que não fazem uso de CAT ou ambientes de aprendizado ou IRT.
- CE-4 Estudos não escritos integralmente em inglês.
- CE-5 Estudos incompletos ou mal relatados, a ponto de não ser possível identificar o método proposto nas ferramentas CAT, IRT ou nos ambientes de aprendizado.
- CE-6 Estudos secundários ou terciários.
- CE-7 Estudos não relacionados às áreas de Ciência da Computação, Matemática e Engenharia.

2.2. Condução

A pesquisa foi realizada no dia 5 de fevereiro de 2017. O motor de busca Scopus retornou 99 trabalhos; destes, 56 artigos foram excluídos pelos critérios CE-1, CE-4 e CE-7, através das ferramentas de filtro desse motor de busca. No motor do Web of Science foram retornados 25 estudos com 19 artigos duplicados. Com isso, restaram 49 do total de 68 estudos. Em seguida, os critérios foram aplicados manualmente para cada um dos 49 estudos restantes. Foram analisados o título, o resumo e as palavras-chave. Quando houve dúvida, o estudo foi analisado por completo. No fim, 9 estudos foram selecionados, os quais são relativamente recentes (publicados a partir de 2005). Esta fase foi auxiliada pelo Software StArt *(State of the Art through Systematic Review)*[7].

3. Resultados

Os estudos selecionados estão listados, cronologicamente, na Tabela 1. Nesta tabela, estão presentes: o identificador (ID), contendo a classe de veículo de publicação (C–Conferência e J–Periódico); a referência; o título; e ano de publicação. Nota-se uma maior concentração de estudos publicados nos últimos 6 anos. A seguir são apresentadas as respostas às questões de pesquisa.

[7]Disponível em: http://lapes.dc.ufscar.br/tools/start_tool

Tabela 1. Estudos selecionados nesta RSL

ID	Referência	Título	Ano
C-1	[Lilley et al. 2005]	Learners' and tutors' perspectives on the usefulness of an automated tool for feedback on test performance	2005
J-2	[Chen et al. 2006]	Personalized curriculum sequencing utilizing modified item response theory for web-based instruction	2006
C-3	[Lee et al. 2010]	A personalized assessment system based on item response theory	2010
C-4	[Wong et al. 2010]	E-learning: Developing a Simple Web-Based Intelligent Tutoring System Using Cognitive Diagnostic Assessment and Adaptive Testing Technology	2010
J-5	[Özyurt et al. 2012]	Integrating computerized adaptive testing into UZWEBMAT: Implementation of individualized assessment module in an e-learning system	2012
C-6	[Lendyuk et al. 2013]	Simulation of computer adaptive learning and improved algorithm of pyramidal testing	2013
C-7	[Rajamani and Kathiravan 2013]	An Adaptive Assessment System to compose Serial Test Sheets using Item Response Theory	2013
J-8	[Jeong and Hong 2013]	A service component based CAT system with SCORM for advanced learning effects	2013
C-9	[Heitink and Veldkamp 2015]	Computer adaptive assessment for learning in a virtual learning environment	2015

Legenda: *C = Conferência; J = Periódico*

3.1. Técnicas de adaptação empregadas nos sistemas e-learning: Q-1

A Tabela 2 exibe quais trabalhos fazem adaptação de materiais de aprendizado e quais fornecem algum tipo de *feedback*. A seguir, esses questionamentos são discutidos com mais detalhes.

Tabela 2. Técnicas de adaptação empregadas nos sistemas de aprendizado

ID	Adaptação do material	*Feedback*
C-1	Não	Sim
J-2	Sim	Não
C-3	Sim	Sim
C-4	Sim	Sim
J-5	Sim	Não
C-6	Sim	Não
C-7	Não	Não
J-8	Não	Sim
C-9	Não	Sim

O J-2 diferente de todos: ao invés de usar questões, coleta do usuário qual foi o nível de dificuldade e o grau de compreensão do conteúdo aprendido e usa a IRT para realizar apenas a recomendação dos materiais de aprendizado.

O C-3 e o C-4 criam um CAT baseado em IRT e o incorporam a um Sistema Tutor Inteligente que seleciona os materiais de aprendizado, a partir da estimativa das habilidades realizadas pelo CAT.

O J-5 seleciona os conteúdos de forma adaptada, utilizando um componente chamado de *Expert System Module*, que se baseia em regras internas e no estilo de aprendizado do aluno.

O trabalho C-6 propõe uma variação do CAT denominada Piramidal. Esse modelo é disposto por uma árvore em formato de pirâmide, que separa os materiais de aprendizado

por blocos. Caso o usuário erre a maioria das questões de um bloco de conhecimento, por exemplo, são sugeridos materiais para reforçar o aprendizado daquele bloco.

Outra característica em sistemas *e-learning* é a capacidade de promover *feedbacks*. O C-1, por exemplo, informa o nível de proficiência geral e por tópico do aprendiz e recomenda revisões. O C-4 fornece, além da proficiência geral, as soluções para os itens. O C-3 afirma que pode fornecer comentários sobre o progresso da aprendizagem e revelar áreas de carência, porém não deixa claro como isso é feito. O J-8 também não fornece detalhes, apenas afirma que após o respondente terminar o teste, o sistema apresenta um parecer sobre os dados da prova. Já o C-9 foi além: para cada questão incorreta, o sistema fornece dicas e conteúdos e o usuário tem outra chance de responder. Com isso, foi construído o Modelo Politômico de Crédito Parcial (do inglês, *Parcial Credit Model* – PCM) [Ostini and Nering 2006], que considera 0 para errou, 1 para acertou, sem *feedback*, e 2 para acertou, com *feedback*. O trabalho mostrou que os *feedbacks* influenciaram de forma positiva o comportamento das respostas.

3.2. Características do CAT: Q-2

Dentre todos os artigos selecionados, apenas o J-2 não faz uso de avaliações adaptativas. A Tabela 3 traz um levantamento das características dos CATs presentes nos trabalhos selecionados e que são discutidas a seguir.

Tabela 3. Características do CAT utilizadas nos estudos selecionados nesta RSL

ID	Seleção de itens	Critério de parada
C-1	Clássico	NI
J-2	Não se aplica	Não se aplica
C-3	IIF	NI
C-4	Aleatório	Limite inferior do domínio.
J-5	IIF	Regras não informadas; limite de questões
C-6	Piramidal	Nível de proficiência mínimo; usuário parou
C-7	IIF com verossimilhança.	Erro padrão; limite de questões
J-8	IIF	Usuário parou; limite de questões
C-9	PCM + IIF	Usuário parou; limite de questões

Critério de Seleção de itens. Para a recomendação apropriada do próximo item, a IRT faz uso de algumas estratégias. A mais adotada, dentre os artigos selecionados, foi a de Máxima Informação, conhecida também como Função de Informação do Item (do inglês, *Item Information Function* – IIF) [Lord 1980] ou Seleção da Informação Máxima (do inglês, *Maximum Information Selection* – MIS). Essa estratégia foi empregada nos trabalhos C-3, J-5, C-7, J-8 e C-9. Já o C-1 usa exclusivamente o modelo clássico, que seleciona a questão mais fácil, se o respondente erra, ou mais difícil, caso acerte.

Existem também trabalhos que utilizam uma combinação de estratégias ou uma própria, dentre eles: (i) o C-4, que faz seleção aleatória e utiliza um outro modelo CAT, o Modelo de Teste de Taxa de Probabilidade Sequencial (do inglês, *Sequential Probability Ratio Test* – SPRT), que não precisa de calibração nem de um grande número de respondentes; (ii) o C-6, que usa a abordagem Piramidal. Quando as perguntas são respondidas erroneamente, o aluno move-se pela árvore para níveis de dificuldade mais baixos; caso acerte, move-se para níveis mais difíceis; (iii) o C-7, em que uma questão pode estar relacionada a um ou vários conceitos em uma escala de 0 a 4, na qual 0 indica a inexistência de relação; e (iv) o C-9, que utiliza o PCM junto com a IIF.

Critério de parada. Vários trabalhos utilizaram como critério de parada o limite de questões (J-5, C-7, J-8 e C-9), ou seja, quando não há mais itens na base de questões ou quando atingiu o limite de questões a serem exibidas naquele dado momento. O critério de parada adotado pelo C-6 considera que o nível de proficiência do aluno ultrapassou o valor mínimo previamente estipulado. Já no C-7, a condição de parada é quando o erro padrão atingiu um certo grau especificado. No C-4, é usada a probabilidade da resposta correta ser maior ou igual a um dado limite inferior do domínio. O critério de parada, quando o respondente resolve interromper o teste por escolha própria, é utilizado por C-6, J-8 e C-9. Os artigos C-1, J-2 e C-3 não deixaram claro quais critérios de parada foram utilizados.

3.3. Características utilizadas da IRT: Q-3

O primeiro fator analisado na implementação da IRT foi o método adotado para estimar os parâmetros dos itens, em que o C-3, o C-7 e o C-9 utilizaram métodos próprios. Já em relação à estimativa das habilidades latentes, o C-7 e o J-2 utilizaram a abordagem Bayesiana. O J-5 e o C-1 utilizaram o método de máxima verossimilhança (do inglês, *Maximum Likelihood Estimator*), para estimar ambos os parâmetros e as habilidades. Para estimar o nível de proficiência, o C-4 utiliza o Modelo de Traço Logístico Linear (do inglês, *Linear Logistic Trait Model*, e que é baseado na IRT) em conjunto com o modelo de espaço do conhecimento (do inglês, *Knowledge Space Model*). Por fim, os trabalhos C-6 e J-8 não deixaram claro como estimaram os parâmetros e/ou as habilidades.

A resposta a um item pode ser avaliada pela perspectiva dicotômica, quando é considerado apenas se o respondente acertou ou errou, e pela perspectiva politômica, que pode considerar valores intermediários. O modelo de resposta às questões mais utilizado foi o dicotômico. Apenas o trabalho C-9 fez uso de itens politômicos.

O número de parâmetros utilizados pelos trabalhos foi bastante variado. O ML2 foi o mais adotado (usado em C-3, C-6, C-7, J-8 e C-9), seguido do ML3 (usado em C-1, J-5 e C-6) e do ML1 (usado em J-2, C-4 e C-6). O destaque foi para o C-6, que analisou os três modelos.

3.4. Características dos experimentos: Q-4

Na Tabela 4, são apresentadas as características gerais dos experimentos realizados em cada um dos estudos selecionados. A segunda e a terceira coluna tratam do tamanho da amostra. A quantidade de itens variou entre 10 e 880. Já o número de alunos selecionados foi entre 20 e 3146. Apenas o C-9 utilizou dados artificiais e nenhum dos trabalhos fez uso de banco de itens de terceiros.

A quarta coluna contém as tecnologias utilizadas para a construção das soluções, a maioria *WEB* (J-2, C-4 e J-5). No J-5, um módulo CAT foi incorporado ao sistema *e-learning* denominado UZWEBMAT. Para isso, primeiro foi criado um banco de questões, que passou por uma série de etapas, desde a coleta das respostas dos alunos até a eliminação de questões que não estavam de acordo com o ML3. De forma similar, o trabalho J-8 construiu um CAT e o incorporou a um sistema *e-learning*. Além disso, no processo de aprendizagem foram utilizadas três estratégias em conjunto, são elas: (i) um sistema de gerenciamento de aprendizado; (ii) um sistema de gerenciamento de conteúdo de aprendizagem; e (iii) o SCORM (Padrão internacional para operacionalizar e gerir os

Tabela 4. Características dos experimentos dos estudos selecionados nesta RSL

ID	Nº de itens	Nº de alunos	Ferramentas e linguagens de programação	Banco de itens	Dados Reais (R) ou Artificiais(A)	Teste t	Correlação	Curso ou Aplicação
C-1	24	115	NI	Não	R	Não	Não	Interação Homem-Computador
J-2	NI	117	PHP 4.3; MySQL Server	Não	R	Não	Sim	Linguagem de programação C
C-3	NI	20	NI	Não	R	Sim	Não	Inglês
C-4	100	NI	NI	Não	R	Não	Não	Redação empresarial
J-5	880	3146	UZWEBMAT; MULTILOG 7.0 [8]	Não	R	Não	Não	Probabilidade
C-6	40	24	Moodle	Não	R	Não	Não	Algoritmos
C-7	10	NI	CBSTR	Não	R	Não	Não	Mineração de dados
J-8	NI	40	SCORM; SPSS	Não	R	Sim	Não	TOEIC
C-9	NI	NI	Moodle	Não	A	Não	Não	Aprendizado Numérico

conteúdos dos objetos de aprendizagem), que operacionalizou e gerenciou o material a ser aprendido de forma eficiente.

Foram identificados quais trabalhos utilizaram o teste t (teste t de *Student*) e quais utilizaram a correlação para avaliar a sequência curricular e a sequência das questões recomendadas. O C-3 utilizou o test t para fazer um teste de amostras pareadas e para realizar uma análise de amostras independentes. No J-8, também foram realizados testes t em amostras independentes, para analisar os resultados do grupo de controle em relação ao grupo de tratamento. Já o J-2, verificou a correlação entre o nível de habilidade e o nível do parâmetro de dificuldade do material de aprendizado recomendado, para avaliar a sequência curricular. Os trabalhos C-1, C-4, C-6, C-7 e C-9 não realizaram a avaliação do sequenciamento de itens ou curricular.

Na última coluna da Tabela 4, observamos que os sistemas foram inseridos e validados em cursos das mais diversas áreas do conhecimento.

4. Ameaças à Validade da Revisão Sistemática

Apesar de definir o protocolo de pesquisa e de ter empregado recursos eletrônicos na identificação e extração dos dados, não é possível garantir que todos os estudos relevantes foram coletados. É provável que alguns estudos possam ter sido omitidos, devido, primeiro, às diversas variações e sinônimos existentes nos termos de busca e também devido ao uso, principalmente, dos critérios de exclusão 1, 2, 4 e 5, que, apesar de importantes, também podem ter omitido potenciais trabalhos. Esses aspectos e o uso limitado do conjunto de fontes de dados influenciaram também no número final de estudos selecionados.

5. Discussão e Conclusão

Por meio desta RSL, foram analisados 9 estudos primários na área de sistemas de aprendizado *e-learning*, que fazem uso de CAT e IRT. A análise identificou, principalmente, as técnicas de adaptação implementadas, os métodos propostos nas ferramentas CAT, o arquétipo de IRT usado e os tipos de experimentos realizados.

Apesar do uso da IRT estar bastante consolidado em ambientes de teste, sua implantação em sistemas de aprendizado baseados em CAT ainda se encontra bastante limitada, por existir, sobretudo, dificuldades na etapa de calibração, devidas à falta de dados suficientes. Caso exista um banco de itens disponível, uma possível solução é utilizá-lo para executar a calibração antes de iniciar a etapa de avaliação. Esse banco deve possuir entre 5 ou 10 vezes mais questões para uma determinada prova (por exemplo, para uma prova de 30 questões, espera-se um banco em torno de 150 a 300 questões) [Özyurt et al. 2012]. Uma outra solução é utilizar um outro modelo CAT, por exemplo, o Modelo de Teste de Taxa de Probabilidade Sequencial, que não precisa de calibração nem de um grande número de respondentes [Wong et al. 2010].

Outro desafio encontrado em ambientes de aprendizado que realizam testes adaptativos está na estratégia de seleção do próximo item, que, diferentemente de ambientes de testes, deve ter o objetivo de otimizar a eficiência do aprendizado, e não medir com maior precisão o nível de proficiência do avaliado [Wauters et al. 2010]. Dentre as possíveis soluções encontradas, estão: (i) coletar maiores informações sobre o entendimento do aluno em relação aos tópicos de aprendizagem [Chen et al. 2006] e (ii) considerar o nível de aderência entre as questões e os tópicos [Rajamani and Kathiravan 2013].

Além disso, os resultados mostram que:

- A grande maioria dos trabalhos fizeram avaliações adaptativas, combinando-as com o sequenciamento de materiais de aprendizado ou *feedback*;
- Apenas cinco, dos nove estudos selecionados, fornecem algum tipo de feedback ao usuário, os quais indicam, principalmente, os pontos fortes e fracos nas habilidades testadas dos avaliados;
- A estratégia mais adotada, na IRT, para a seleção do próximo item foi a Função de Informação do Item;
- Apenas um dos trabalhos selecionados fez a avaliação politômica;
- Há uma tendência no uso de dados reais para a validação dos trabalhos, pois apenas um utilizou dados artificiais. A quantidade de usuários envolvidos nos experimentos variou entre 20 e 3146 e a quantidade de itens, entre 10 e 880;
- A grande maioria dos trabalhos não realizou a avaliação do sequenciamento de itens ou curricular. Os três trabalhos que avaliaram os sistemas propostos ou utilizaram o teste t de *Student* ou a correlação.

Os resultados do trabalho mostram que o uso de ambientes *e-learning* baseados em CAT e IRT ainda se encontra bastante limitado. Desta forma, a motivação para a realização desta RSL é, além de compor um estado da arte na área, incentivar outros pesquisadores em informática na educação a desenvolverem propostas e modelos de avaliação, bem como divulgarem os resultados de seus processos avaliativos.

Referências

Baker, F. B. and Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Chen, C.-M., Lee, H.-M., and Chen, Y.-H. (2005). Personalized e-learning system using item response theory. *Computers & Education*, 44(3):237–255.

Chen, C.-M., Liu, C.-Y., and Chang, M.-H. (2006). Personalized curriculum sequencing utilizing modified item response theory for web-based instruction. *Expert Systems with Applications*, 30(2):378 – 396.

de Andrade, D. F., Tavares, H. R., and da Cunha Valle, R. (2000). Teoria da resposta ao item: conceitos e aplicações. *ABE, Sao Paulo*.

Guzmán, E. and Conejo, R. (2004). A model for student knowledge diagnosis through adaptive testing. In *International Conference on Intelligent Tutoring Systems*, pages 12–21. Springer.

Heitink, M. and Veldkamp, D. B. P. (2015). Computer adaptive assessment for learning in a virtual learning environment. In *Proceedings of the 18th International Computer Assisted Assessment Conference*, number 571, pages 22 – 26. Springer.

Jeong, H.-Y. and Hong, B.-H. (2013). A service component based CAT system with SCORM for advanced learning effects. *Multimedia Tools Appl.*, 63(1):217–226.

Kitchenham, B. A. (2012). Systematic review in software engineering: Where we are and where we should be going. In *Proceedings of the 2Nd International Workshop on Evidential Assessment of Software Technologies*, EAST '12, pages 1–2. ACM.

Kovatcheva, E. and Nikolov, R. (2009). An adaptive feedback approach for e-learning systems. *IEEE Technology and Engineering Education (ITEE)*, 4(1):55–57.

Lee, Y., Cho, J., Han, S., and Choi, B.-U. (2010). A personalized assessment system based on item response theory. In *Proceedings of the International Conference on Advances in Web-Based Learning (ICWL 2010)*, pages 381–386. Springer Berlin Heidelberg.

Lendyuk, T., Rippa, S., and Sachenko, S. (2013). Simulation of computer adaptive learning and improved algorithm of pyramidal testing. In *2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDACS)*, volume 02, pages 764–769.

Lilley, M., Barker, T., and Britton, C. (2005). Learners' and tutors' perspectives on the usefulness of an automated tool for feedback on test performance. In *Proceedings of the European Conference on Games-based Learning*, volume 2005-January, pages 181–190.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Ostini, R. and Nering, M. L. (2006). *Polytomous item response theory models*. Number 144. Sage.

Rajamani, K. and Kathiravan, V. (2013). An adaptive assessment system to compose serial test sheets using item response theory. In *International Conference on Pattern Recognition, Informatics and Mobile Engineering*, pages 120–124.

Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. *Copenhagen: Danish Institute for Educational Research*.

Spenassato, D., Trierweiller, A. C., de Andrade, D. F., and Bornia, A. C. (2016). Testes adaptativos computadorizados aplicados em avaliações educacionais. *Revista Brasileira de Informática na Educação*, 24(2).

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., and Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

Wauters, K., Desmet, P., and Van Den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *Journal of Computer Assisted Learning*, 26(6):549–562.

Wong, K., Leung, K., Kwan, R., and Tsang, P. (2010). E-learning: Developing a simple web-based intelligent tutoring system using cognitive diagnostic assessment and adaptive testing technology. In *Third International Conference on Hybrid Learning (ICHL)*, pages 23–34. Springer Berlin Heidelberg.

Özyurt, H., Özcan Özyurt, Baki, A., and Güven, B. (2012). Integrating computerized adaptive testing into UZWEBMAT: Implementation of individualized assessment module in an e-learning system. *Expert Systems with Applications*, 39(10):9837 – 9847.