CNMAI Congresso Nacional de Matemática Aplicada à Indústria 2014
18 a 21 de novembro de 2014, Caldas Novas - Goiás

Aplicação de Testes Adaptativos Computadorizados em Modelos de Desdobramento Graduado Generalizados

Augusto Sousa da Silva Filho, augustofilho@yahoo.com.br [1]
Marcos Antônio da Cunha Santos, msantos@est.ufmg.br [2]
Rodrigo Tómas Nogueira Cardoso, rodrigoc@des.cefetmg.br[3]

[1]Faculdade IBS - Fundação Getúlio Vargas

[2]Universidade Federal de Minas Gerais - Departamento de Estatística

[3]Centro Federal de Educação Tecnológica de Minas Gerais - Dep. de Modelagem Matemática e Computacional

Resumo. *Quando um pesquisador deseja determinar o grau ou a intensidade dos sintomas de um paciente aplica um questionário de papel e caneta com uma série de perguntas pré-definidas, o mesmo acontece quando um pedagogo busca determinar a habilidade de um aluno em uma determinada área do conhecimento. Na quase totalidade dos casos, tais questionários são demasiadamente extenuantes para o entrevistado e, devido a isso, correm o risco de não representar de forma significativa o estado do paciente ou o grau de conhecimento do aluno. Uma alternativa para este cenário é a construção de um Computerized Adaptive Testing - CAT. O CAT procura otimizar o teste para cada respondente, buscando estimar a característica predominante de um indivíduo a cada item respondido. Neste artigo, o CAT desenvolvido não utilizará a Teoria Clássica de Medidas, que considera apenas a proporção de acertos e erros em uma prova. Em vez disso, será utilizada a Teoria de Resposta ao Item - (TRI), que apresenta a vantagem de fazer comparações entre traços latentes de indivíduos de populações diferentes quando são submetidos a testes ou questionários que tenham alguns itens comuns e permite, ainda, a comparação de indivíduos da mesma população submetidos a testes totalmente diferentes. Isto é possível porque a TRI tem como elementos centrais os itens e não o teste ou questionários como um todo. A TRI apresenta modelos matemáticos que já foram amplamente abordados pela literatura, tais como os modelos acumulativos, devido à sua facilidade de implementação. No entanto, outros modelos não apresentam a mesma facilidade de implementação, como os modelos de desdobramentos. Este trabalho propõe a construção de um CAT utilizando um modelo de desdobramento graduado generalizado, visto que este modelo pode ser utilizado tanto para respostas dicotômicas como para respostas politômicas ordinais.*

Palavras-chave: *Teoria de Resposta ao Item, Testes Adaptativos Computadorizados, Modelos de Desdobramento Graduado Generalizado, Modelagem Matemática e Computacional*

1. TESTES ADAPTATIVOS COMPUTADORIZADOS E TEORIA DE RESPOSTA AO ITEM

Em um mundo globalizado, onde a rapidez é a marca registrada em quase todos os segmentos de nossas vidas, é possível afirmar que somos ao mesmo tempo espectadores e protagonistas dos diversos segmentos do conhecimento humano, rapidez esta representada especialmente pelos sucessivos aprimoramentos e inovações nos campos científicos e tecnológicos. Inseridos nesse contexto de mudanças e transformações técnicas, sociais e econômicas, acentua-se a importância de descobrir novas metodologias que forneçam condições para que essas áreas se desenvolvam e a avaliação, certamente, tem sido um instrumento fiel que irá mensurar estas transformações. A avaliação é um sistema de informação que tem como objetivo fornecer um diagnóstico e subsídios para manutenção e para prover um contínuo monitoramento de um sistema, como por exemplo, sistema educacional, sistema econômico, sistema religioso, sistema político, com vistas a detectar os efeitos positivos ou negativos de políticas adotadas". A forma tradicional usada para avaliação de desempenho de um estudante é denominada de Teoria Clássica de Testes (TCT), que em leva em consideração o total de questões corretas dentro de um conjunto total de questões, ou seja, obtém-se o escore do teste. Uma vantagem desta

teoria é que ela é relativamente simples de ser interpretada e exige poucas ou nenhuma suposições sobre os dados. Por outro lado, a teoria clássica de Testes sofre de algumas limitações ou deficiências.

As principais deficiências da Teoria Clássica do Teste - TCT são:

a) O escore do estudante não é uma medida absoluta, pois pode variar de teste para teste, dependendo do conteúdo do teste;
b) É difícil comparar o desempenho de alunos aplicando-se testes diferentes;
c) A avaliação de desempenho dos estudantes é influenciada pela amostra analisada.

A Teoria de Resposta ao Item - (TRI), vem sendo estudada nas últimas décadas com sucesso para a construção e análise de testes e vem tendo muito sucesso exatamente nas limitações da TCT.

Nos trabalhos de Baker e Kim (1992) e de Hambleton e Cook (1977), é possível destacar as principais vantagens da TRI em relação à TCT:

a) Permite construir uma escala para medir o conhecimento dos estudantes, de tal forma que se possa avaliar a proficiência dos estudantes tornando-a independente;
b) Possibilita obter características das questões ("itens", no vocabulário da TRI), identificando as questões que realmente contribuem para avaliação do conhecimento;
c) Permite acompanhar o desenvolvimento de um aluno ao longo do tempo;
d) Permite comparar resultados de testes aplicados em classes de alunos diferentes;
e) Permite comparar a dificuldade das questões.

A TRI utiliza um modelo matemático para extrair informações e realizar estimativas das questões e dos estudantes. As estimativas buscam explicar o efeito entre as respostas dos estudantes e seus traços latentes (habilidade/proficiência). Portanto, o modelo matemático obtido irá expressar tal relação em forma de uma equação matemática.

No Brasil, a teoria de resposta ao item vem sendo utilizada com sucesso na análise dos dados do Sistema Nacional de Ensino Básico (SAEB) e no Exame Nacional do Ensino Médio (ENEM), ambos realizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) do Ministério da Educação (MEC). No exterior, pode-se citar o Programme for International Student Assessment (PISA).

As avaliações realizadas pelo MEC, normalmente são aplicadas em larga escala em todo o território nacional e do tipo "papel e lápis" (do inglês, *paper-and-pencil*) - que é a forma usual de uma avaliação. Tal forma de avaliação exige uma grande logística e elevados custos com transporte, fiscalização, aplicação, correção, etc,. Nesta operação gigantesca é comum erros pontuais, como o que ocorreu no ano de 2009, com o furto de apenas uma prova do ENEM, o que ocasionou um prejuízo de R$ 34 milhões para o governo, conforme foi noticiado no jornal "A Folha de São Paulo", pela jornalista Sófia Fernandes:

"O ministro da Educação, Fernando Haddad, afirmou nesta quinta-feira, em Brasília, que o prejuízo com o vazamento da prova do Enem será de 30% do valor total do contrato, que custou R$ 116 milhões, segundo a pasta. Dessa forma, a perda estimada é de aproximadamente R$ 34 milhões – valor apenas das impressões. Não se sabe ainda quem deve assumir esse prejuízo."

Uma alternativa para a realização destas avaliações em larga escala e com um custo significativamente menor seria a aplicação de versões informatizadas dos testes. No Brasil, avaliações desta natureza ainda podem ser consideradas incipientes e apresentam apenas a mudança no processos de aplicação do teste, que passou apenas a ser apresentada diretamente na tela de um computador, desta forma é possível definir um teste informatizado como um teste tradicional do tipo "papel e lápis" realizado por meio do computador. Logo, essa foi considerada a primeira geração de testes informatizados. Uma alternativa para as avaliações do tipo "papel e lápis" tradicional ou informatizada é a utilização dos chamados Testes Adaptativos Computadorizados - CAT.

Um Teste Adaptativo Computadorizado - (do inglês, *Computerized Adaptive Testing* - CAT), é um teste baseado na TRI administrado via computador que tem como objetivo apresentar itens adequados para o indivíduo que está realizando o teste, buscando uma melhor estimação da habilidade ou nota desse indivíduo junto com a redução do tempo de realização do teste".

No Brasil, a Universidade de Brasília tem se destacado na produção de Testes Adaptativos Computacionais, principalmente após o desenvolvimento de um CAT para avaliar a proficiência em língua estrangeira.

Embora possa parecer simples desenvolver um CAT, muitos cuidados devem ser tomados para evitar a construção de um teste inadequado. A maneira como as etapas do algoritmo do CAT devem ser desenvolvidas e os métodos e critérios a serem adotados dependem de diversos fatores, tais como:

a) Tipo de teste;
b) Objetivo do teste;
c) Tamanho e qualidade do banco de itens;
d) Modelo de resposta ao item utilizado;
e) Público alvo para a aplicação do teste, etc.

Os Testes Adaptativos Computadorizados são normalmente desenvolvidos em estudos teóricos, simulações, aplicações práticas e comparações de métodos e voltados para aspectos particulares do CAT em estudo, como

o desenvolvimento de um CAT específico ou o estudo de algum método específico. Existe uma carência em encontrar um método para a implantação de um CAT desde o princípio da sua elaboração até a sua efetiva aplicação, passando por todos os passos necessários e que também possa servir como um guia para a implantação de um CAT mais geral.

Diversos autores utilizam os testes adaptativos computadorizados baseados em teoria de resposta ao item, com uso de modelos acumulativos de 1, 2 e 3 parâmetros, por serem considerados mais simples para implementação. No entanto existe uma carência para uso dos modelos de desdobramento, dentre eles o Modelo Parella, Modelo de desdobramento graduado generalizado e o modelo Cosseno Hiperbólico, este modelos não alcançaram tanto progresso e a razão disso se deve principalmente à compreensão do seu mecanismo de resposta e a falta de programas computacionais para estimar os parâmetros deste tipo de modelo.

Os modelos de desdobramentos da Teoria de Resposta ao Item são modelos onde são estimados os parâmetros dos itens e os parâmetros dos indivíduos, desta forma, os parâmetros dos indivíduos e dos itens são colocados em uma mesma escala. Portanto, os parâmetros do individuo são alocados de acordo com sua opinião e os dos itens são localizados nessa mesma escala de acordo com seu conteúdo. Portanto, estes modelos também são conhecidos como modelos de proximidade, pois as categorias de resposta dos itens mais altos são mais prováveis quando a distância entre a posição do parâmetro do indivíduo e do item na escala diminui. Em um modelo de desdobramento, a probabilidade de um indivíduo dar uma resposta a um item está em função da distância entre a posição do parâmetro do indivíduo e a posição do parâmetro do item na escala. Logo em um modelo de desdobramento para uma medida de atitude o indivíduo concorda com uma categoria de resposta de um item na medida em que o sentimento transmitido pela resposta do item combina com a sua própria opinião.

2. MATERIAL E MÉTODO

2.1. Modelos de Desdobramento Graduado Generalizado - GGUM

O modelo de desdobramento graduado generalizado GGUM é um modelo de Teoria de Resposta ao Item unidimensional, desenvolvido para analisar respostas binárias como as graduadas baseadas numa relação de proximidade.

O modelo de desdobramento graduado generalizado pode ser usado tipicamente em situações de medidas onde os entrevistados são solicitados a indicarem seu nível de concordância, a um conjunto de itens que se situam numa escala bipolar, ou seja, que variam o seu conteúdo do negativo, passando pelo neutro e chegando ao positivo. Este modelo é ideal para analisar dados relacionados a atitudes.

O GGUM generaliza modelos de desdobramentos de resposta ao item em dois modos:

- Primeiro, ele implementa um parâmetro de discriminação que varia através dos itens e logo os itens são utilizados para discriminar entre os respondentes em modos diferentes.
- Segundo, o GGUM permite o uso de categoria de resposta diferencial através dos itens. Dessa forma ele age por implementação no limiar da categoria resposta que varia através dos itens.

O GGUM baseia-se no modelo de Credito Parcial Generalizado proposto por Muraki (1992), que se fundamentou no modelo de créditos parciais de Masters (1982), em que os parâmetros são de locação. Quando é aplicado às respostas subjetivas, o modelo de crédito parcial generalizado é dado por:

$$P\left(Y_i = y|\theta_j\right) = \frac{\exp\left[\alpha_i\left(y\left(\theta_j - \delta_i\right) - \sum_{k=0}^{y}\tau_{ik}\right)\right]}{\sum_{v=0}^{m}\exp\left[\alpha_i\left(v\left(\theta_j - \delta_i\right) - \sum_{k=0}^{v}\tau_{ik}\right)\right]} \tag{1}$$

Com a seguinte restrição:

$$\sum_{k=0}^{M}\tau_{ik} = 0 \tag{2}$$

Onde:

i) Y_i uma resposta subjetiva à declaração de atitude i;
ii) $y = 0, 1, 2, 3, \ldots, M$; $y = 0$ corresponde ao nível mais forte de discordância abaixo do item, enquanto que, $y = M$ corresponde ao nível mais forte de discordância acima do item;
iii) θ_j parâmetro de locação do indivíduo j num *continuum* de atitude;
iv) δ_i parâmetro de locação de um item i num *continuum* de atitude;
v) α_i parâmetro de discriminação de um item i;
vi) τ_{ik} parâmetro de locação do limiar de categoria de resposta subjetiva k num *continuum* de atitude relativa à posição do item i;
vii) M é o numero das categorias de respostas subjetivas menos 1.

Define-se como ψ distância entre os parâmetros limiares, desta forma, o valor de τ_{i0} é arbitrário definido para ser zero na Equação (1), mas poderia ser ajustado para qualquer constante sem afetar o resultado das probabilidades.

O GGUM foi desenvolvido a partir de quatro proposições básicas sobre o processo de resposta.

a) A primeira salienta que quando um indivíduo é solicitado para expressar a sua opinião de aceitação em uma declaração de atitude, o indivíduo tende a concordar com o item à medida que ele é localizado próximo de sua posição pessoal num continuum de atitude latente unidimensional. Dessa forma, o grau para o qual o sentimento de um item reflete a opinião de um indivíduo é dada pela proximidade do indivíduo ao item num continuum de atitude. Se δ_i denotar a locação (posição) do item i num *continuum* e θ_j denotar a locação do indivíduo j no mesmo *continuum* então o indivíduo é mais tendente a concordar com o item à medida que a distância entre θ_j e δ_i se aproxima de zero. Isso é uma característica fundamental de um processo de ponto ideal.

b) A segunda proposição do modelo indica que um indivíduo pode responder uma determinada categoria de resposta por dois motivos distintos. Existem duas possíveis respostas subjetivas associada com a única resposta observável. O Modelo de Desdobramento Graduado Generalizado desdobra, então, duas respostas subjetivas para cada resposta observável numa escala de avaliação.

c) A terceira proposição do GGUM é que as respostas subjetivas às declarações de atitudes seguem um modelo de resposta ao item cumulativo.

A Equação (1) de Muraki, define um modelo de TRI para níveis de respostas subjetivas. Desta forma, cada categoria de resposta observável está associada com duas possibilidades de resposta subjetiva, ou seja, um abaixo do item e outro acima do item. As duas categorias de respostas subjetivas correspondentes a uma dada categoria de resposta observável são disjuntas. Portanto, a probabilidade de um indivíduo responder usando uma categoria observável particular é a soma das probabilidades associadas com as duas respostas subjetivas correspondentes.

$$P\left(Z_i = z|\theta_j\right) = P\left(Y_i = z|\theta_j\right) + P\left(Y_i = (M-z)\,|\theta_j\right) \tag{3}$$

Onde:

i) Z_i é uma resposta observável à declaração de atitude i;

ii) $z = 0, 1, 2, \ldots, D$; $z = 0$ corresponde ao nível de discordância mais forte e $z = D$ corresponde ao nível de concordância mais forte;

iii) D é o número de categorias de respostas observáveis menos $1 \cdot M = 2 \cdot D + 1$.

d) A quarta proposição do GGUM, como a existência de simetria dos limiares das categorias subjetivas em torno do ponto $(\theta_j - \delta_i) = 0$, indicando que:

$$\tau_{i(D+1)} = 0 \tag{4}$$

e

$$\tau_{iz} = -\tau_{i(M-Z+1)}, \textit{para } z \neq 0. \tag{5}$$

Desta forma, esta quarta proposição implica que os indivíduos estão bem propensos a concordar com um item localizado tanto em unidades -p ou unidades +p da posição do indivíduo num continuum de atitude. Essa proposição conduz à identidade:

$$\sum_{k=0}^{z} \tau_{ik} = \sum_{k=0}^{M-z} \tau_{ik} \tag{6}$$

Logo, incorporando esta proposição à Equação (3), resulta-se na função de probabilidade do GGUM, descrito por Roberts et al. (2000), da seguinte forma:

$$P\left(Z_i = z|\theta_j\right) = \frac{\exp\left[\alpha_i\left(z\left(\theta_j - \delta_i\right) - \sum_{k=0}^{z}\tau_{ik}\right)\right] + \exp\left[\alpha_i\left((M-z)\left(\theta_j - \delta_i\right) - \sum_{K=0}^{z}\tau_{ik}\right)\right]}{\sum_{v=0}^{H}\left[\exp\left(\alpha_i\left[v\left(\theta_j - \delta_i\right) - \sum_{k=0}^{v}\tau_{ik}\right]\right) + \exp\left(\alpha_i\left[(M-v)\left(\theta_j - \delta_i\right) - \sum_{k=0}^{v}\tau_{ik}\right]\right)\right]} \tag{7}$$

Onde:

i) Z_i é uma resposta observável a um item de atitude i;

ii) $z = 0, 1, 2, 3, \ldots H$; $z = 0$ corresponde ao nível de discordância mais forte e $z = H$ corresponde ao nível de concordância mais forte;

iii) H é o número de categorias de respostas observáveis menos 1 com $M = 2h + 1$;

iv) θ_j é o parâmetro de locação do indivíduo j na escala do traço latente;

v) δ_i é o parâmetro de locação do item i na escala do traço latente;

vi) α_i é o parâmetro de discriminação do item i;

vii) t_{ik} é o parâmetro de posição do limiar de categoria de resposta subjetiva k na escala do traço latente relativa à posição do item i; corresponde ao valor da distância entre θ_j e δ_i que determina o ponto em que a k-ésima categoria de resposta subjetiva passa a ter probabilidade de resposta sobre $(k-1)$-ésima categoria de resposta subjetiva para o indivíduo j no item i e τ_{i0} é, por definição, igual a zero;

viii) M é o número das categorias de respostas subjetivas menos 1.

Assim o GGUM é um modelo da TRI desenvolvido para analisar tanto respostas binarias quanto respostas graduadas baseadas em uma relação de proximidade. Sua aplicação típica ocorre quando entrevistados são convidados a indicar seu nível ou grau de concordância em relação a um conjunto de itens que se situam em uma escala bipolar, ou seja, onde existe uma variação do seu conteúdo do positivo para o negativo, passando pelo neutro.

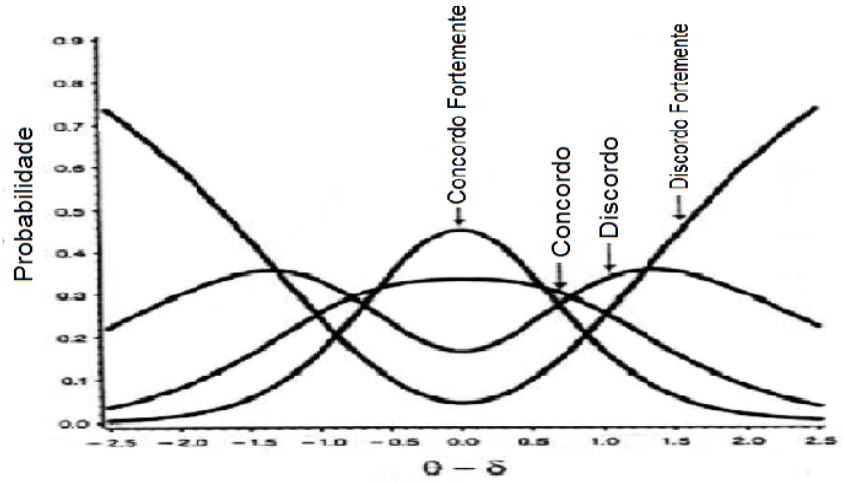

Figura 1. Função de Probabilidade de um item com 4 categorias de respostas observáveis - Fonte: ?

A Figura (1), mostra estas funções de probabilidade de respostas observáveis das categorias de respostas para um mesmo item referenciado. As funções de probabilidades das respostas observáveis não se interceptam em $\tau_{i1}, \tau_{i2}, \tau_{i3}, \ldots, \tau_{iD}$. Assim é possível perceber que os parâmetros τ_{ik} perdem sua simples interpretação ao nível de resposta observável. Da mesma maneira o parâmetro α_i indexa a discriminação para um nível de resposta subjetiva. Em contrapartida, a média $\theta_j - \delta_i$ não mudam quando se deslocam de um nível de resposta subjetiva para observável.

2.2. Algoritmo de um Teste Adaptativo Computadorizado - CAT

Para (Moreira, 2011, p. 98), "a maior vantagem de um CAT em relação a um teste tradicional consiste na maior eficiência na estimação do nível de habilidade do respondente com um menor número de itens do que os testes tradicionais". Para atingir esse objetivo, dois aspectos básicos são necessários: o método de estimação das habilidades dos respondentes e o critério de seleção dos itens.

A maioria dos CATs utiliza a seguinte estratégia:

a) Um critério de partida, para determinar o primeiro item a ser apresentado;
b) Um método estatístico (geralmente Bayesiano ou Máxima Verossimilhança) para estimar a proficiência do indivíduo e a precisão associada;
c) Um procedimento para selecionar o próximo item;
d) Um critério para finalizar o teste;

A lógica da seleção dos itens no teste, em geral, acontece da seguinte forma: se o indivíduo acerta o item atual, o próximo item deverá ser de um nível mais difícil; se o indivíduo erra o item atual, o próximo item deverá ser de um nível mais fácil.

O desenvolvimento de um CAT é um processo trabalhoso e exige conhecimentos e técnicas importantes. Em primeiro lugar, necessita de um grande banco de itens devidamente calibrado através da TRI. Em segundo lugar, deve-se programar um conjunto de algoritmos para a seleção progressiva dos itens, para a estimação dos níveis de habilidade e sua respectiva precisão. Em terceiro lugar, um CAT deve submeter-se aos testes aplicados para garantir as propriedades desejáveis das estimações, assim como a sua precisão e validade. Se a aplicação do CAT for através da internet, ainda existe um trabalho adicional de programação para preservar a segurança do banco de itens e para realizar o processo de seleção do próximo item em um tempo imperceptível para o avaliado.

A lógica de um teste adaptativo informatizado pode ser descrito da seguinte forma:

1. Iniciar com uma Estimativa Aleatória Provisória de Proficiência: esse é o nível de conhecimento inicial;
2. Selecionar e apresentar um item: compreende os critérios de seleção de itens considerando as restrições, quando existentes;
3. Observar a resposta: o examinando fornece uma resposta ao item;
4. Revisar a estimativa da proficiência: reestimar a proficiência utilizando a resposta observada no passo 3;

5. A regra de parada foi satisfeita (Sim/Não)? Verificar se o critério de parada foi alcançado. Se "Sim", vai para o passo 6. Se "Não", volta para o passo 2;
6. Fim do teste: Finaliza o teste se a resposta do passo 5 for "Sim".

Os seguintes elementos básicos que devem compor os CATs:

- Modelo de Resposta ao Item: é o modelo de probabilidade que será ajustado aos itens do teste;
- Banco de Itens: é o conjunto de itens que contém todo o domínio do conhecimento abordado pelo teste.
- Nível de Conhecimento Inicial: é a estimativa inicial provisória da proficiência, necessária para o início do teste e relacionada com o nível de dificuldade da primeira questão.
- Método de Seleção dos Itens: é um algoritmo que deve selecionar o próximo item em função do nível estimado provisório da proficiência e da sua resposta dada ao item anterior.
- Critério de Parada: é uma regra para finalizar o teste.

A Figura 2, apresenta um fluxograma para a construção de um teste adaptativo informatizado.

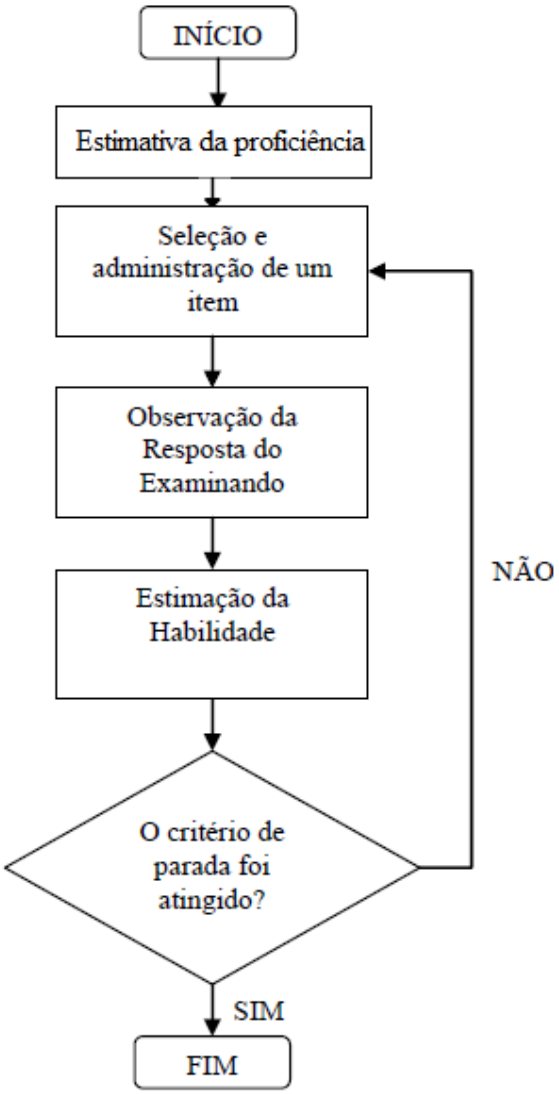

Figura 2. Fluxograma para um CAT - Fonte: Moreira (2011)

2.3. Construção de um banco de itens em um CAT

Para a construção de um teste adaptativo computadorizado é necessário primeiramente a organização de um banco de itens. É possível a criação de testes individualizados, visto que os CATs se adequam as capacidades de cada examinando, para que isso seja alcançado, é necessário que o banco de itens construído tenha uma ampla diversidade de itens.

Um plano geral para a construção de um banco de itens é descrito na figura 3.

Segundo Wainer et al. (1990), um banco de itens deve conter:

a) Criação de um número suficientes de itens para cada categoria de competências, baseando-se nas especificações do teste estabelecida previamente;

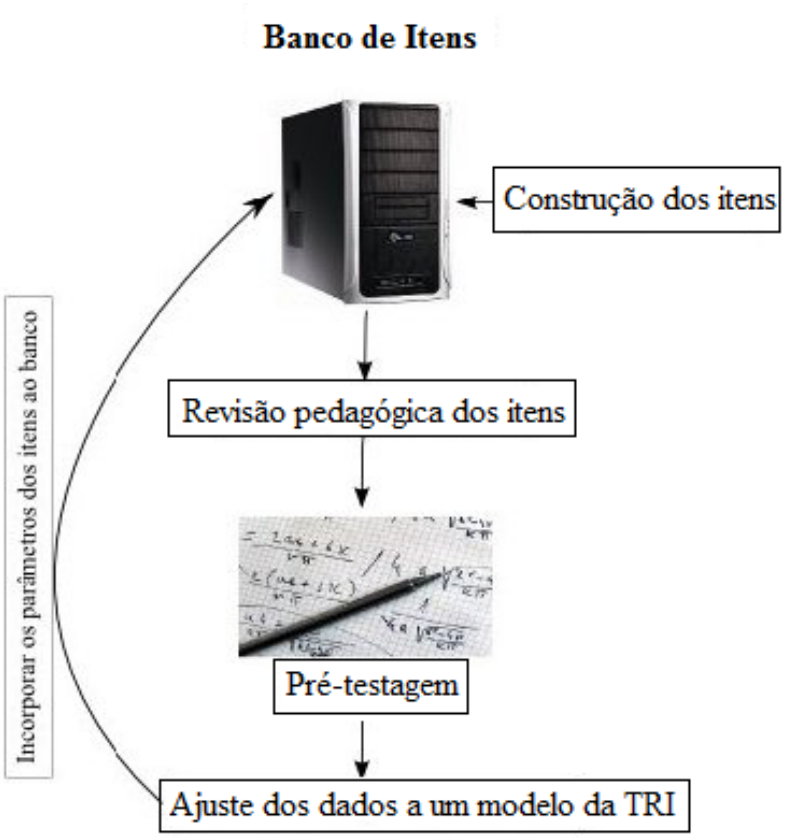

Figura 3. Construção de um banco de itens, Fonte:Costa (2009)

b) Realizações de revisões pedagógicas da qualidade dos itens. Observar, por exemplo, se os itens não apresentam funcionamento diferenciado, tais como presença de DIF, baseado em características específicas do examinado do que a habilidade mensurada pelo teste, tais como gênero ou etnia;

c) Pré-teste dos itens, para verificação da qualidade dos itens envolvidos no teste.

Durante a construção do banco de itens, o interesse principal consiste na estimação dos parâmetros dos itens pela TRI. Com o uso de um banco de dados pré-calibrado, ou seja, um banco de dados que possua estimativas para os parâmetros de interesse, pois um CAT usando a TRI só funcionará se existir um banco de dados calibrado previamente. Portanto, dado um banco de itens pré-calibrado, a estimação das proficiências dos examinados pela CAT depende fundamentalmente dos métodos de seleção dos itens. Uma vez assumido como verdadeiro o modelo proposto, a partir das respostas dos examinandos aos itens, faz-se possível a calibração dos itens

3. RESULTADOS PRELIMINARES

3.1. Desenvolvimento do Banco de Itens

Com o uso do software R Team (2014), foi simulado um banco de dados de tamanho 100, para 25 itens com 4 categorias de respostas objetivas que formam 9 categorias de respostas subjetivas, seguindo o Modelo de Desdobramento Graduado Generalizado - GGUM. O parâmetro α de cada item foi gerado de acordo com uma distribuição aleatória uniforme variando no intervalo de $min = 0$ e $max = 1$, o parâmetro δ foi gerado aleatoriamente de acordo com uma distribuição normal com média zero e variância 1. O vetor τ, é simétrico e os seus valores foram gerados aleatoriamente de acordo com uma distribuição uniforme. A seguir é apresentado somente os seis primeiros parâmetros simulados para um tamanho de amostra igual a 100 e 25 itens.

```
head(montaItens(25))
```

```
Parametros dos itens utilizados na primeira simulacao com números de taus igual a 7
item  alpha    delta    tau1     tau2     tau3        tau4   tau5      tau6     tau7
1     0.91315  2.63809 -2.00644 -1.52803 -0.676276    0      0.676276  1.528034 2.00644
2     0.56269  0.14471 -2.81151 -1.57807 -0.966940    0      0.966940  1.578077 2.81151
3     0.85518  1.53227 -2.89484 -1.58782 -0.699936    0      0.699936  1.587827 2.89484
4     0.58807  0.65615 -2.80953 -1.47251 -0.862249    0      0.862249  1.472513 2.80953
5     0.61937  0.31107 -2.41056 -1.36871 -0.771739    0      0.771739  1.368712 2.41056
6     0.99434 -2.31887 -2.26226 -1.95044 -0.920469    0      0.920469  1.950440 2.26226
```

Com a simulação dos parâmetros do Modelo de Desdobramento Graduado Generalizado, usou-se a função de densidade de probabilidade (equação 7) para determinar as probabilidades associadas. É apresentado a seguir, somente as seis primeiras probabilidades encontradas para os dois primeiros itens determinados.

```
> head(probPar(100,c(0,1,2,3,4,5,6)))
[[1]]
     P(Z=0|theta) P(Z=1|theta) P(Z=2|theta) P(Z=3|theta) P(Z=4|theta) P(Z=5|theta) P(Z=6|theta)
[1,] 0.07299141   0.2619921    0.34700170   0.2293066739 7.705834e-02 1.019865e-02 1.451115e-03
[2,] 0.24722873   0.4125435    0.25240464   0.0759137266 1.124452e-02 6.243461e-04 4.048931e-05
[3,] 0.19969509   0.3942258    0.28536602   0.1015813341 1.784934e-02 1.189003e-03 9.346027e-05
[4,] 0.53074909   0.3650631    0.09206530   0.0114114237 6.952029e-04 1.555729e-05 3.684265e-07
[5,] 0.03803135   0.1830611    0.32758012   0.2933041835 1.319768e-01 2.235790e-02 3.688479e-03
[6,] 0.24761117   0.4126468    0.25214093   0.0757361071 1.120353e-02 6.212111e-04 4.022594e-05

[[2]]
     P(Z=0|theta) P(Z=1|theta) P(Z=2|theta) P(Z=3|theta) P(Z=4|theta) P(Z=5|theta) P(Z=6|theta)
[1,] 0.23085530   0.2783757    0.23384593   0.14864068   0.07286309   0.0266275067 8.791798e-03
[2,] 0.11042102   0.1910976    0.23205053   0.21362229   0.14948894   0.0739712002 2.934842e-02
[3,] 0.13281665   0.2112034    0.23619413   0.20144149   0.13203776   0.0622439001 2.406265e-02
[4,] 0.13256570   0.2109872    0.23615734   0.20157833   0.13222551   0.0623678423 2.411808e-02
[5,] 0.33996800   0.3143505    0.20159962   0.09688695   0.03528740   0.0094591084 2.448443e-03
[6,] 0.11026907   0.1909558    0.23201659   0.21370466   0.14961203   0.0740552881 2.938660e-02
```

A simulação acima, apresenta as probabilidades de respostas dos indivíduos para os itens [[1]] e [[2]]. Com estas probabilidades simuladas é possível construirmos uma matriz de respostas para os itens e com isso darmos continuidade para a determinação do teste adaptativo. A seguir são apresentados somente os 81 primeiros resultados.

```
> head(head(bancoItens(100,25)))
Banco de itens Simulados
id X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25
1  1  2  1  1  1  1  1  3  3   1   1   1   2   1   2   2   3   1   1   2   1   2   3   2   2
2  2  2  2  2  2  1  2  2  1   3   1   3   2   2   3   2   1   2   1   3   3   1   3   2   2
3  2  1  2  1  1  3  2  1  3   2   3   1   3   2   3   1   2   2   3   2   1   1   2   5   1
4  4  3  4  3  3  2  4  3  2   4   5   3   6   2   3   3   2   3   3   3   4   4   2   4   2
5  1  1  1  1  1  1  1  3  2   2   1   2   1   1   1   1   1   1   1   1   1   1   2   1   1
6  1  3  1  1  4  1  1  2  3   3   2   1   1   2   2   2   3   2   2   1   3   5   2   2   2
.  .  .  .  .  .  .  .  .  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
.  .  .  .  .  .  .  .  .  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
.  .  .  .  .  .  .  .  .  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
81 1  1  1  1  1  1  1  1  1   1   1   1   1   1   2   2   1   1   1   1   1   1   1   1   1
```

Após a simulação do banco de resposta para o modelo GGUM, é necessário que façamos uma análise da qualidade do banco criado para verificarmos a suposição de unidimensionalidade. Este pressuposto é uma exigência fundamental no processo de estimação dos parâmetros do modelo, uma vez que sua violação invalida as estimativas do parâmetro. Para verificar a suposição de unidimensionalidade foi desenvolvido uma análise Fatorial pelo método dos componentes principais. O algoritmo desenvolvido para verificar a suposição de unidimensionalidade está disponível no Apêndice desta trabalho.

Na tabela 1, foram calculadas na análise fatoriais as comunalidades de cada item para verificar quais itens eram menos favoráveis com a suposição de unidimensionalidade nos dois primeiros componentes. A análise fatorial se realizou com o método dos componentes principais extraindo-se dois fatores. Para que um item se ajuste à suposição de unidimensionalidade é necessário que sua comunalidade seja maior que 0,3, ou seja, que o item tenha mais do que 30% de aspectos em comuns com outros itens. Portanto, os itens 4,9,10,23 e 24 apresentaram comunalidades menor do que 0,30, sendo removidos do banco de Itens.

Embora menos usado, também podemos destacar o Coeficiente Alfa de Cronbach, que serve para estimar a confiabilidade de um questionário aplicado em uma pesquisa. Ele mede a correlação entre as respostas em um questionário através da análise das respostas dadas pelos respondentes, apresentando uma correlação média entre as perguntas. O coeficiente Alfa de Cronbach é calculado a partir da variância dos itens individuais e da variância da soma dos itens de cada avaliador de todos os itens de um questionário que utilizem a mesma escala de medição e o seu valor foi igual a 0.9144242.

A confiabilidade do questionário simulado segundo o valor de Alfa de Cronbach é dado na tabela a seguir:

A interpretação do coeficiente alfa de Cronbach é quase intuitiva porque os valores variam, em geral, entre zero e 1 e a confiabilidade é tanto maior quanto mais perto de 1 estiver o valor da estatística. De qualquer

Item	Fator 1	Fator 2	Comunalidade
1	0.6709437	-0.019972757	0.4505644
2	0.5568865	-0.174359991	0.3405240
3	0.5494144	0.160132131	0.3274985
4	0.4943004	0.217653457	0.2917059
5	0.6807503	-0.133595368	0.4812686
6	0.6431710	0.195382789	0.4518434
7	0.6684457	0.385591217	0.5955002
8	0.5723217	0.227398978	0.3792624
9	0.5043632	0.203304893	0.2957151
10	0.4992723	0.187657671	0.2844883
11	0.5824116	0.220242640	0.3877101
12	0.5557184	-0.082187498	0.3155777
13	0.6087127	-0.375921658	0.5118482
14	0.5414195	0.213266314	0.3386176
15	0.5676753	-0.294118248	0.4087608
16	0.5758326	0.140512546	0.3513269
17	0.5757812	-0.359739194	0.4609362
18	0.6755634	0.028218051	0.4571821
19	0.5455790	0.434253648	0.4862327
20	0.6403032	-0.276904754	0.4866645
21	0.4274213	-0.411772029	0.3522451
22	0.6368789	-0.251812342	0.4690242
23	0.4264128	-0.030044401	0.1827305
24	0.5046001	-0.211853567	0.2995032
25	0.5947066	-0.004298166	0.3536944

Tabela 1. Matriz de comunalidades dos itens

Valor de Alfa	Confiabilidade
$> 0,90$	Excelente
$0,8 \vdash 0,9$	Bom
$0,7 \vdash 0,8$	Aceitável
$0,6 \vdash 0,7$	Questionável
$0,5 \vdash 0,6$	Pobre
$< 0,50$	Inaceitável

Tabela 2. Tabela de Alfa de Cronbach

modo, é preciso saber que o valor de alfa é afetado não apenas pela correlação entre as respostas obtidas, mas também pelo número de questões feitas e por redundância. Questionários muito longos aumentam o valor de alfa, sem que isso signifique aumento de confiabilidade. Valores muito altos de alfa também podem estar indicando redundância, isto é, a existência de questões praticamente iguais, verbalizadas de forma diferente.

Das duas análises, podemos afirmar que o banco de itens simulado pode ser considerado adequado para a aplicação de um Teste Adaptativo Computadorizado. O algoritmo para a construção do Teste Adaptativo Computadorizado seguirá a seguinte sistemática:

i) Sorteia-se aleatoriamente um item do banco de item simulado;
ii) O Critério de Informação de Fisher determinará qual será o próximo item escolhido;
iii) Para cada passo do Teste Adaptativo Computadorizado, re estima-se o θ_{jt} do indivíduo;
iv) O critério de parada ocorrerá quando um número específico de itens for escolhido.

Seguindo os passos acima, espera-se que o valor de θ_{jt} convirja para o verdadeiro valor de θ para uma quantidade pequena de itens selecionados.

O valor de θ_{jt} é encontrado usando-se a equação 8. Os valores X_q e $A(X_q)$ são respectivamente os nós (nodes) e os pesos (weights) de quadratura. Para o Modelo de Desdobramento Graduado Generalizado, utilizou-se 20 pontos de Quadratura de Gauss-Hermite. Para encontrar os pontos de quadratura, usamos o package statmod, desenvolvido por Smyth et al. (2014).

```
> require(statmod)
> ptosQt=gauss.quad(n=20,kind="hermite",alpha=0,beta=0)
nodes
-5.3874809 -4.6036824 -3.9447640 -3.3478546 -2.7888061
```

```
-2.2549740 -1.7385377 -1.2340762 -0.7374737 -0.2453407
 0.2453407  0.7374737  1.2340762  1.7385377  2.2549740
 2.7888061  3.3478546  3.9447640  4.6036824  5.3874809

weights
2.229394e-13 4.399341e-10 1.086069e-07 7.802556e-06 2.283386e-04
3.243773e-03 2.481052e-02 1.090172e-01 2.866755e-01 4.622437e-01
4.622437e-01 2.866755e-01 1.090172e-01 2.481052e-02 3.243773e-03
2.283386e-04 7.802556e-06 1.086069e-07 4.399341e-10 2.229394e-13
```

A função de verossimilhança será determinada usando-se a resposta do item selecionado dado θ, com isso, será possível determinar um conjunto de resultados para $\hat{\theta}_{jt}$. Para o primeiro passo do algoritmo do Teste Adaptativo Computadorizado, sorteou-se aleatoriamente o octogésimo primeiro indivíduo e selecionou-se o décimo terceiro item. A resposta dada por este individuo é apresentada a seguir.

```
resp.cand=resp0[ID,item0]
resp.cand

#Resposta do item selecionado
X13
  1
```

O valor de $\hat{\theta}_{jt}$ deverá ser determinado para este indivíduo utilizando-se os valores do nó e do peso de quadratura, de acordo com a equação abaixo:

$$\hat{\theta}_{jt} = \frac{\sum_{q=1}^{Q} X_q L_{jt}(X_q) A(X_q)}{\sum_{q=1}^{Q} L_{jt}(X_q) A(A_q)} \tag{8}$$

Logo, temos:

$$\hat{\theta}_{jt} = \frac{(-5.3874809) \cdot L(-5.3874809) \cdot (2.229394e-13) + \ldots + (5.3874809) \cdot L(5.3874809) \cdot (2.229394e-13)}{L(-5.3874809) \cdot (2.229394e-13) + \ldots + L(5.3874809) \cdot (2.229394e-13)} \tag{9}$$

A função de Verossimilhança é determinada como:

$$\begin{aligned}
L(-5.3874809) &= P(Z_{13} = 1 | \theta = -5.3874809), \textbf{pois a resposta do item } X_{13} = 1 \\
L(-5.3874809) &= \sum_{i=1}^{n} (\log f(x)) \\
L(-5.3874809) &= \sum_{i=1}^{n} \left[\log \left(\frac{\exp\left[\alpha_i \left(z(\theta_{jt} - \delta_i) - \sum_{k=0}^{z} \tau_{ik} \right)\right] + \exp\left[\alpha_i \left((M-z)(\theta_{jt} - \delta_i) - \sum_{K=0}^{z} \tau_{ik} \right)\right]}{\sum_{v=0}^{H} \left[\exp\left(\alpha_i \left[v(\theta_{jt} - \delta_i) - \sum_{k=0}^{v} \tau_{ik} \right] \right) + \exp\left(\alpha_i \left[(M-v)(\theta_{jt} - \delta_i) - \sum_{k=0}^{v} \tau_{ik} \right] \right) \right]} \right) \right]
\end{aligned}$$

Portanto, para o item 13, temos os seguintes parâmetros: $\alpha_{13} = 0.85958055$, $\delta_{13} = 0.8853347$, e $\theta = -5.3874809$ e τ=(-2.616215, -1.730224, -0.9245658, 0, 0.9245658, 1.730224, 2.616215), de modo que $\sum \tau = 0$. O resultado de somente uma das verossimilhanças é apresentado a seguir:

$$\begin{aligned}
L(-5.3874809) &= -4.582771 \\
L(-4.6036824) &= -3.946904 \\
\ldots &= \ldots \\
L(5.3874809) &= -1.094219
\end{aligned}$$

Portanto, o valor de θ_{jt} será dado por:

$$\begin{aligned}
\hat{\theta}_{jt} &= \frac{(-5.3874809) \cdot L(-5.3874809) \cdot (2.229394e-13) + \ldots + (5.3874809) \cdot L(5.3874809) \cdot (2.229394e-13)}{L(-5.3874809) \cdot (2.229394e-13) + \ldots + L(5.3874809) \cdot (2.229394e-13)} \\
\hat{\theta}_{jt} &= \frac{(-5.3874809) \cdot (-4.582771) \cdot (2.229394e-13) + \ldots + (5.3874809) \cdot (-1.094219) \cdot (2.229394e-13)}{(-4.582771) \cdot (2.229394e-13) + \ldots + (-1.094219) \cdot (2.229394e-13)} \\
\hat{\theta}_{jt} &= 1.589208e-16
\end{aligned}$$

Para escolhermos o próximo item, deveremos determinar o valor da Informação de Fisher. O maior valor para a Informação de Fisher, determinará qual o próximo item a ser avaliado. Desta forma, temos:

$$\begin{aligned} I_i(\theta_j) &= -\alpha_i^2 \left[\left(\sum_{z=0}^{D} P(Z_i = z|\theta_j)\, \sigma^2_{Y_i|\theta_j,z} \right) - \sigma^2_{Y_i|\theta_j} \right] \\ I_1(\theta_j) &= 0.5011327 \\ I_2(\theta_j) &= 0.2351315 \\ \ldots &= \ldots \\ I_{25}(\theta_{25}) &= 0.002288616 \end{aligned}$$

O maior valor para a Informação de Fisher foi encontrado para o décimo segundo item ($I_{12} = 1,337644$), sendo o item escolhido para o próximo passo. O décimo segundo item apresenta os seguintes parâmetros: $\alpha = 0.93696036$, $\delta = 0.5295769$ e $\tau_i = (-2.048590, -1.687307, -0.9256431, 0, 0.9256431, 1.687307, 2.048590)$, com $\sum_{n=1}^{n} \tau_i = 0$. Logo, o valor de θ_{jt} será igual a 1.538432e-16.

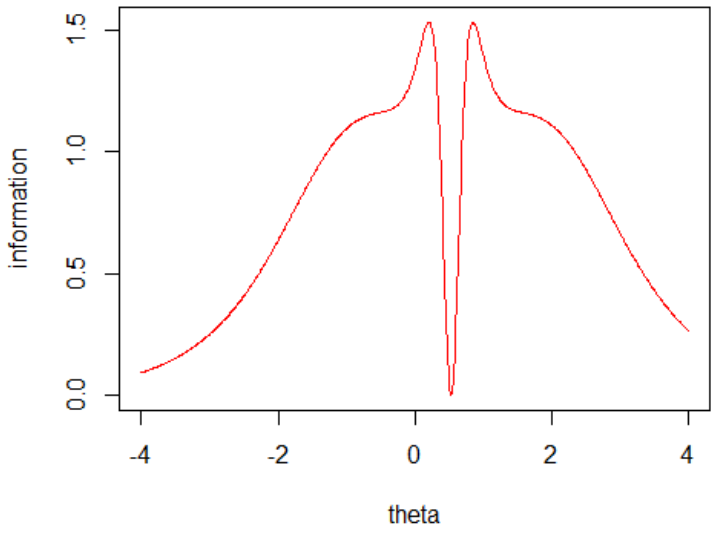

Figura 4. Curva de Informação de Fisher para o Item 12

Pelo critério da Informação de Fisher, o próximo item escolhido foi o item 21. Os parâmetros apresentados para este item são: $\alpha_{21} = 0.78910663$, $\delta_{21} = -1.0281821$ e $\tau_{21} = (-2.204685, -1.488901, -0.7411177, 0, 0.7411177, 1.488901, 2.204685)$ e $\theta_{jt} = 1.371074e-16$. Para finalizarmos o método, necessitamos de um critério de parada. Neste trabalho, adotaremos um critério de parada relacionado ao número de itens escolhidos, ou seja, o Teste Adaptativo Computadorizado será interrompido para um total de 15 itens escolhidos. Para efeito de simulação, o valor obtido para $\theta_{13} = 0.3879431$, ou seja, é possível observar no gráfico acima uma tendência para o valor simulado.

Passo	Item	θ_{jt}
1	13	1,589208e-16
2	12	1.538432e-16
3	21	1.311907e-16
4	5	1.649399e-16
5	19	1.475241e-16
...	...	...
15	1	1.259331e-16

Tabela 3. Passos do Teste Adaptativo

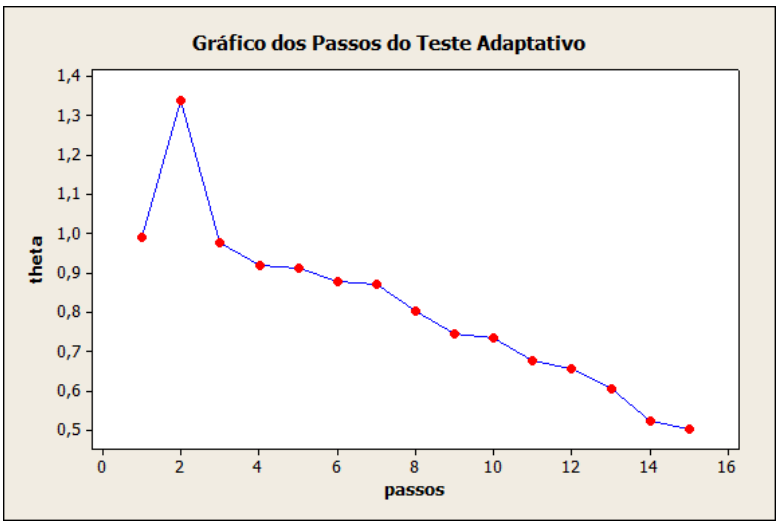

Figura 5. Passos para o CAT com 25 itens

O critério de parada ocorre quando um número determinado de itens for apresentado. Neste caso, escolhemos arbitrariamente que o CAT iria chegar ao fim com um total de 15 itens. No entanto, o gráfico acima não mostra que houve uma convergência para o verdadeiro valor do parâmetro $\theta = -2.31887093$, mostrando apenas uma tendência para zero. Uma das possíveis causas é o cálculo da verossimilhança com valores de pesos de quadratura na ordem de 16 casas decimais negativas. O algoritmo CAT deve ser compilado mais uma vez para outro item e o critério de convergência deve ser observado. A figura **??**, mostra um fluxograma com todos os passos envolvidos neste algoritmo CAT/GGUM.

4. CONCLUSÃO

Este texto está trata-se de uma pesquisa em desenvolvimento. O seu objetivo é a criação de banco de itens para ser modelado pelos modelos de desdobramentos graduados generalizados, estes modelos possuem a vantagens de serem aplicados tanto para dados dicotômicos, quanto para dados categóricos. Após obtermos um banco de itens considerado ajustado, passaremos para a criação do teste adaptativo computadorizado. Neste estudo inicial, o banco de itens simulado, passou a mostrar um sinal de convergência para cerca de 12 itens.

REFERÊNCIAS

Baker, F., e Kim, S.-H. 1992. *Item response theory: Parameter estimation techniques*, vol. 176. CRC Press.

Costa, D. R. 2009. *Métodos estatísticos em testes adaptativos informatizados*. Dissertação de Mestrado , Universidade Federal do Rio de Janeiro, Rio de Janeiro.
URL `http://www.pg.im.ufrj.br/teses/Estatistica/Mestrado/121.pdf`

Hambleton, R. K., e Cook, L. L. 1977. Latent trait models and their use in the analysis of educational test data1, 2, 3. *Journal of educational measurement*, *14*(2), 75–96.
URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.1977.tb00030.x/abstract`

Masters, G. N. 1982. A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.
URL `http://link.springer.com/article/10.1007/BF02296272`

Moreira, F. 2011. *Sistemática para a implantação de testes adaptativos informatizados baseados na teoria da resposta ao item*. Tese de Doutorado, Universidade Federal de Santa Catarina, Florianópolis.
URL `http://repositorio.ufsc.br/handle/123456789/95506`

Muraki, E. 1992. A generalized partial credit model: Application of an em algorithm. *Applied psychological measurement*, *16*(2), 159–176.
URL `http://apm.sagepub.com/content/16/2/159.short`

Roberts, J. S., Donoghue, J. R., e Laughlin, J. E. 2000. A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*(1), 3–32.
URL `http://apm.sagepub.com/content/24/1/3.short`

Smyth, G., Hu, Y., Dunn, P., Phipson, B., e Chen, Y. 2014. *statmod: Statistical Modeling*. R package version 1.4.20.
URL `http://CRAN.R-project.org/package=statmod`

Team, R. C. 2014. R: A language and environment for statistical computing.
URL `http://www.R-project.org`

Wainer, H., Dorans, N. J., Green, B. F., Steinberg, L., Flaugher, R., Mislevy, R. J., e Thissen, D. 1990.

Computerized adaptive testing: A primer.. Lawrence Erlbaum Associates, Inc.
URL http://www.springerlink.com/index/M58R14228306306V.pdf

5. SOBRE RESPONSABILIDADE AUTORAL E O USO DE IDIOMA PORTUGUÊS OU ESPANHOL

Os trabalhos escritos em português ou espanhol devem incluir (após direitos autorais) título, os nomes dos autores e afiliações, o resumo e as palavras chave, traduzidos para o inglês e a declaração a seguir, devidamente adaptada para o número de autores.

RESPONSABILIDADE AUTORAL

"O(s) autor(es) é(são) o(s) único(s) responsável(is) pelo conteúdo deste trabalho".