UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Ailton Fonseca Galvão

Um Modelo Inteligente para Seleção de Itens em Testes Adaptativos Computadorizados

Juiz de Fora

2013

Ficha catalográfica elaborada através do Programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Galvão, Ailton Fonseca.
Um Modelo Inteligente para Seleção de Itens em Testes Adaptativos Computadorizados / Ailton Fonseca Galvão. -- 2013.
79 p.

Orientador: Raul Fonseca Neto
Coorientador: Carlos Cristiano Hasenclever Borges
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, ICE/Engenharia. Programa de Pós-Graduação em Ciência da Computação, 2013.

1. Testes Adaptativos Computadorizados. 2. Seleção de Itens. 3. Inteligência Computacional. I. Fonseca Neto, Raul, orient. II. Borges, Carlos Cristiano Hasenclever, coorient. III. Título.

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Ailton Fonseca Galvão

Um Modelo Inteligente para Seleção de Itens em Testes Adaptativos Computadorizados

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientador: Raul Fonseca Neto

Coorientador: Carlos Cristiano Hasenclever Borges

Juiz de Fora

2013

Ailton Fonseca Galvão

Um Modelo Inteligente para Seleção de Itens em Testes Adaptativos Computadorizados

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Computação da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Mestre.

Aprovada em 06 de setembro de 2013.

BANCA EXAMINADORA

Prof. Dr. Raul Fonseca Neto – Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Carlos Cristiano Hasenclever Borges – Coorientador
Universidade Federal de Juiz de Fora

Prof. Dr. Joaquim José Soares Neto
Universidade de Brasília

Prof. Dr. Alfredo Chaoubah
Universidade Federal de Juiz de Fora

Às três pessoas que mais pensam em mim:
meu pai, minha mãe e minha esposa.

AGRADECIMENTOS

Expresso aqui minha gratidão àqueles que, em maior ou menor grau, dedicaram parte de suas vidas me ensinando que sempre há algo mais que podemos aprender e que, se completo mais uma etapa, é porque tive muito apoio nesse caminho.

Aos meus pais, não há palavras para agradecer cada minuto que vocês dedicam a mim. Obrigado por me mostrarem o valor de ser uma pessoa honesta, de manter a palavra, de se esforçar e valorizar o que se consegue. E por me ensinarem que o tempo dedicado aos estudos não era uma obrigação, mas, um privilégio.

À Natália, por todos os anos de amor e carinho. Obrigado por me apoiar sempre que eu precisei de incentivo e por ser paciente quando não pude te dedicar muito tempo. A cada dia tenho mais certeza que estou com a pessoa certa.

O agradecimento mais que especial aos meus orientadores, professores Raul Fonseca Neto e Carlos Cristiano Borges, que desde o início, ainda na inscrição para o mestrado, apoiaram a minha ideia e dedicaram seu tempo para me ajudar a concretizar este trabalho. Não posso deixar de mencionar meu grande amigo professor Jairo Francisco de Souza que, na criação do curso de mestrado, foi o primeiro a me incentivar a tentar uma vaga. E também o agradecimento a todos os professores que fizeram parte da minha formação, especialmente os da Universidade Federal de Juiz de Fora. Obrigado por me fazerem evoluir.

Aos companheiros de mestrado, além do agradecimento, deixo os meus parabéns por também alcançarem esse objetivo. Meu agradecimento especial ao Roberto Nalon, pelas diversas ideias e opiniões que ajudaram a dar forma a este trabalho. Obrigado também à Gláucia Vargas por estar sempre disposta a nos ajudar na nossa vida acadêmica.

Aos amigos de várias etapas da vida, obrigado por fazerem a minha existência ficar cada vez mais divertida. Aos do CAEd, principalmente os da Coordenação de Medidas, agradeço por sempre darem o incentivo necessário para que eu pudesse chegar até aqui.

À coordenação do CAEd por me dar o tempo necessário para que o mestrado pudesse ser concluído. E meus sinceros agradecimentos a todos que fazem da qualidade da educação uma prioridade.

"Yes there are two paths you can go by

But in the long run

There's still time to change the road you're on"

Led Zeppelin (Stairway To Heaven)

RESUMO

Testes Adaptativos Computadorizados (TAC) são um tipo de avaliação aplicada utilizando-se de computadores que tem como principal característica a adequação do nível das questões do teste ao desempenho de cada indivíduo avaliado. Os dois principais elementos que compõem um TAC são: (i) o banco de itens, que é o conjunto das questões disponíveis para serem utilizadas no teste; (ii) o modelo de seleção, que faz a escolha de quais questões, chamadas aqui de itens, são aplicadas aos indivíduos. O modelo de seleção de itens é o núcleo do TAC, pois é o responsável por identificar o nível de conhecimento dos indivíduos à medida que os itens são aplicados fazendo com que o teste se adapte, selecionando os itens mais adequados para produzir uma medida precisa. Nesta dissertação, é proposto um modelo para seleção de itens baseado em metas para a precisão do teste através da estimativa do erro padrão da proficiência, por meio de um controle específico do mesmo para cada fase do teste. Utilizando simulações de testes, os resultados são comparados aos de outros dois modelos tradicionais de seleção, avaliando o desempenho do modelo proposto em termos da precisão do resultado e do nível de exposição dos itens do banco. Por fim, é feita uma análise específica sobre o cumprimento das metas ao longo dos testes e a possível influência no resultado final, além de considerações sobre o comportamento do modelo em relação às características do banco de itens.

Palavras-chave: Teste Adaptativo Computadorizado. Seleção de Itens. Erro Padrão.

ABSTRACT

Computerized Adaptive Tests (CAT) are a type of assessment tests applied through computers which main feature is the adequacy of the test questions to the performance of each examinee. The two main elements of a CAT are: (i) the item pool, which is the set of available questions for testing; (ii) the selection model, which pick out the questions, named items, applied to the examinees. The item selection model is the core of CAT, and its main task is to identify examinees knowledge level as the items are applied and to adapt the test, selecting the most proper items to produce an accurate measure. This thesis proposes a model for item selection based on goals for the test precision using the estimation of the proficiency standard error. For that, an specific control of the goals for each step of the test is developed. Using simulated tests, the results are compared to two traditional item selection models, evaluating the performance of the proposed model in terms of measure accuracy and the level of exposure of the items. Finally, a specific analysis is performed on the accomplishment of goals over the tests and the possible influence on the final result, in addition to considerations on the behavior of the model in relation to the characteristics of the item pool.

Keywords: Adaptive Computerized Test. Items Selection. Standard Error.

LISTA DE FIGURAS

2.1 Curva característica de um item pelo modelo de três parâmetros 21
2.2 Esquema de representação de um Teste Adaptativo Computadorizado 22
2.3 Distribuição do erro padrão pela escala de proficiência em Língua Portuguesa na 8ª série do SAEB 2003 25
2.4 Exemplo de subgrupo de estratos onde um item é selecionado 28

3.1 Comportamento do erro padrão em simulações de TAC's com seleção de itens por MIF 34
3.2 Diferença média das previsões do erro padrão da proficiência de itens aplicados em simulações de TAC's 38
3.3 Exemplo de distribuição dos itens de acordo com seus parâmetros 39
3.4 Variação da estimativa de proficiência no decorrer da aplicação de um TAC com modelo MIF 40
3.5 Definição do espaço de busca dada a estimativa de proficiência e do erro padrão mais recentes 41

4.1 Distribuição dos itens de acordo com o descritor e parâmetro de dificuldade . . 45
4.2 Distribuição dos itens do banco de acordo com seus parâmetros 47
4.3 Distribuição dos itens do banco pelo parâmetro de discriminação 48
4.4 Distribuição dos itens do banco pelo parâmetro de dificuldade 48
4.5 Distribuição dos itens do banco pelo parâmetro de acerto casual 49
4.6 Comportamento do erro padrão pelo MIF e exemplo de previsão de metas . . 56

5.1 Média do erro padrão a cada aplicação de item 60
5.2 Média do erro padrão pelas estimativas de proficiência obtidas no teste 61
5.3 Distribuição dos itens do banco pelo parâmetro de dificuldade 62
5.4 Média de seleção de itens pela escala de dificuldade 64
5.5 Distribuição das proficiências das simulações dos três modelos 64
5.6 Número de itens selecionados pela escala de dificuldade 65
5.7 Percentual de cumprimento de metas por seleção de itens 68
5.8 Percentual de cumprimento de metas por testes 69

5.9 Distribuição da proficiência dos testes que não cumpriram nenhuma das metas 69
5.10 Média de metas cumpridas de acordo com a proficiência nos testes 70
5.11 Média de metas cumpridas de acordo com a precisão dos testes 71

LISTA DE TABELAS

3.1 Exemplo do cálculo de metas para o erro padrão a cada seleção de itens 37

4.1 Distribuição dos itens pela classificação de descritores 46
4.2 Distribuição dos itens nos subgrupos formados pelos estratos 52

5.1 Teste de Kruskal-Wallis do erro padrão por modelo de seleção de itens 59
5.2 Erro padrão por modelo obtido ao fim das simulações 59
5.3 Índice de precisão atingido e itens utilizados 60
5.4 Proficiência média dos casos com erro padrão acima de 0,3 61
5.5 Itens diferentes selecionados e média de seleção por item 63
5.6 Presença de descritores por teste simulado 66
5.7 Teste de Kruskal-Wallis da distribuição de descritores por modelo de seleção de itens 66
5.8 Teste de Kruskal-Wallis da repetição de descritores nos testes por modelo de seleção de itens 67
5.9 Número de repetições de descritores por testes em cada modelo 67
5.10 Metas cumpridas e erro padrão dados pela combinação entre testes precisos e recuperação de metas 70
5.11 Teste de Kruskal-Wallis do erro padrão pela combinação entre testes precisos e recuperação de metas 71

LISTA DE ABREVIATURAS

CCI	Curva Característica do Item
EAP	Estimador Bayesiano da Média a Posteriori
M3PL	Modelo Logístico de Três Parâmetros
MEV	Modelo de Seleção de Minimização da Variância Esperada
MIF	Modelo de Seleção de Máxima Informação de Fisher
MV	Estimador de Máxima Verossimilhança
SH	Método de Controle de Exposição de Itens Sympson-Hetter
TAC	Teste Adaptativo Computadorizado
TCT	Teoria Clássica dos Testes
TRI	Teoria da Resposta ao Item

SUMÁRIO

1 INTRODUÇÃO 13
1.1 ORGANIZAÇÃO DO TRABALHO 15

2 TESTES ADAPTATIVOS E SELEÇÃO DE ITENS 17
2.1 TESTES E MEDIDAS DE PROFICIÊNCIA 17
2.1.1 Teoria da Resposta ao Item - TRI 18
2.1.2 Modelo Logístico de Três Parâmetros (M3PL) 19
2.2 TESTES ADAPTATIVOS COMPUTADORIZADOS (TAC'S) 21
2.3 SELEÇÃO DE ITENS EM TAC'S 24
2.3.1 Seleção Baseada em Níveis de Dificuldade 26
2.3.2 Seleção Baseada em Medida de Informação 29

3 SELEÇÃO DE ITENS POR METAS DO ERRO PADRÃO - MODELO MEP 32

4 COMPONENTES PARA SIMULAÇÃO DOS MODELOS 44
4.1 COMPOSIÇÃO DO BANCO DE ITENS 44
4.2 MÉTODO DE ESTIMAÇÃO DE PROFICIÊNCIAS 47
4.3 SIMULAÇÃO DAS RESPOSTAS 50
4.4 CRITÉRIOS DE PARADA E PRECISÃO DO TESTE 51
4.5 ESTRATOS DE DIFICULDADE E DISCRIMINAÇÃO - SELEÇÃO POR ESTRATIFICAÇÃO 52
4.6 CONTROLE DE EXPOSIÇÃO DE ITENS - SELEÇÃO POR MÁXIMA INFORMAÇÃO 53
4.7 ESTIMATIVA DAS METAS DO ERRO PADRÃO - SELEÇÃO POR METAS DE ERRO 55

5 SIMULAÇÕES NUMÉRICAS E ANÁLISE DE RESULTADOS 58
5.1 ESTIMATIVAS DO ERRO PADRÃO DA PROFICIÊNCIA 58
5.2 EXPOSIÇÃO DOS ITENS 63
5.3 ANÁLISE DAS METAS 68

6 CONCLUSÕES E CONSIDERAÇÕES FINAIS 72

REFERÊNCIAS 77

1 INTRODUÇÃO

Quando qualquer tipo de teste é desenvolvido, é necessário definir antecipadamente qual o seu propósito. Se, em uma avaliação de conhecimentos, os resultados serão utilizados como parâmetros para estudos do desempenho de uma população, a preocupação principal na formulação do teste é que seu conteúdo seja amplo, cobrindo os diversos tópicos que compõem uma determinada área de conhecimento. Essa característica, aliada a uma participação efetiva da população no teste, é suficiente para uma análise de deficiências e futura formulação de medidas que visam à melhoria da qualidade do ensino para esse público avaliado. Porém, se o objetivo do teste for algum tipo de medida de graduação, qualificação ou classificação de indivíduos, o foco do teste passa a ser o resultado e a precisão com que foi obtido. Todo instrumento utilizado para produzir uma medida possui, em maior ou menor escala, um determinado grau de imprecisão e, por isso, toda medida produzida por este instrumento terá uma margem de erro associada a ela.

A maior parte dos testes já desenvolvidos e aplicados se baseia em características simples, como número de acertos ou percentual de acerto no teste, ou, no máximo, em pontuações baseadas em ponderações dos valores de cada questão do teste. Um conceito básico presente nesse tipo comum de teste é a impossibilidade de separação entre as características dos avaliados e do teste, um só pode ser interpretado no contexto do outro. Assim, torna-se inviável a comparação entre testes que medem diferentes características ou que foram aplicados a populações diferentes. Para esse tipo comum de teste, não se pode estabelecer uma equivalência nos resultados obtidos (LORD, 1980).

Na década de 1960, a partir de trabalhos como os de Rasch (1960) e Lord e Novick (1968), um novo paradigma de testes foi desenvolvido. A Teoria da Resposta ao Item (TRI) permitiu estabelecer, de forma probabilística, uma relação entre a proficiência de um indivíduo em uma determinada área de conhecimento e as questões de um teste, chamadas de itens. Diversos modelos matemáticos foram criados para modelar essa probabilidade dadas as características dos itens que são respondidos. Assim, testes em larga escala se tornaram prática comum em diversos países ao longo das últimas décadas, atingindo um amplo número de áreas do conhecimento e de indivíduos avaliados.

A partir dos modelos desenvolvidos para a TRI, surgiram novas propostas, como a de

Weiss (1973), que apresentaram a ideia de testes personalizados e adaptados às características do indivíduo avaliado. Os testes adaptativos são, em sua maioria, destinados a determinar a qualificação dos indivíduos avaliados, logo, o desenvolvimento de um teste com boa precisão é, além de uma necessidade, um objetivo. O procedimento de construção de um teste adaptativo deve ser capaz de responder dinamicamente ao desempenho do indivíduo ao longo de sua aplicação utilizando essa informação para montar o teste que melhor se ajusta ao real nível de conhecimento do avaliado.

Os Testes Adaptativos Computadorizados (TAC's) surgiram como um novo passo na evolução dos testes em larga escala, permitindo inúmeros avanços do ponto de vista psicométrico. Foi possível utilizar modelos matemáticos e probabilísticos cada vez mais robustos na produção das proficiências, trabalhar com bancos de dados com um número cada vez maior de questões proporcionando maior variedade de conteúdo nos testes, reduzir o tempo de aplicação, produzir resultado imediato ao fim do teste, validar novas questões durante a aplicação dos testes e muitos outros aspectos (WAINER, 2000). Porém, essas novas características trouxeram questionamentos dos mais variados.

Diversas pesquisas sugeriram novas teorias sobre como garantir a melhor composição dos bancos de dados em relação aos itens, qual a quantidade mínima e máxima de itens que devem formar o teste, qual o método de estimação da proficiência é o mais apropriado para um determinado teste e como selecionar os itens que são aplicados no teste, fator que é o objeto de estudo deste trabalho. No processo de evolução dos testes adaptativos houve uma compartimentação dos estudos sobre essas diversas características que compõem esses testes e, muitas vezes, devido a essa compartimentação, a mudança proposta para uma característica afeta outra de forma negativa. No caso específico do procedimento da seleção de itens, os principais efeitos negativos são relativos à seleção incompatível com o nível real da proficiência do indivíduo, causando imprecisão na estimativa do resultado, e à seleção exaustiva de um mesmo grupo de itens, chamada de superexposição de itens.

O estudo dos métodos de seleção de itens tornou-se crucial uma vez que esse procedimento atinge a principal característica do teste: a relação entre os itens aplicados e a estimação da proficiência. Foram propostas as mais variadas soluções para o problema da seleção de itens, baseadas em diversos conceitos como classificação por níveis de dificuldade, maximização de informação, informação global, minimização da variância, informação ponderada, entre outros (LINDEN; PASHLEY, 2000). Alguns desses métodos

controlam melhor a exposição de itens em troca de uma precisão que pode ser considerada inferior enquanto outros visam diretamente a precisão e lançam mão de técnicas independentes para o controle de exposição. Além disso, um dos objetivos dos testes adaptativos é reduzir o tamanho do teste, o que se torna um fator complicador, pois a precisão do teste é proporcional ao seu tamanho, o que acaba fazendo com que o modelo de seleção de itens tenha que compensar esse problema.

Neste trabalho será proposto um novo modelo de seleção de itens em testes adaptativos computadorizados baseado no controle da precisão da estimativa de proficiência, buscando atender, ao mesmo tempo, a minimização da exposição de itens. O modelo apresentado utilizará como referência para avaliação de seu potencial dois métodos de seleção de itens amplamente conhecidos, porém, de estrutura conceitual completamente diferentes: o método da Maximização da Informação de Fisher (LORD, 1980) e o de Estratificação em Faixas de Dificuldade (CHANG et al., 2001). Essa nova proposta não visa, necessariamente, minimizar o número de itens utilizados no teste, nem conseguir uma precisão superior aos modelos existentes. Seu objetivo é conseguir equilibrar os fatores precisão da proficiência, exposição de itens e tamanho do teste.

1.1 ORGANIZAÇÃO DO TRABALHO

No capítulo 2 serão abordados os principais conceitos sobre testes utilizando a Teoria da Resposta ao Item, os fatores que permitiram aos testes tradicionais evoluírem em direção aos testes adaptativos, a composição dos processos que definem o funcionamento dos Testes Adaptativos Computadorizados, a caracterização do problema da seleção de itens em TAC's e a estrutura dos dois modelos tradicionais de seleção que são utilizados nas simulações deste trabalho.

O capítulo 3 apresenta e desenvolve o modelo proposto para a seleção de itens. O modelo baseia-se em uma estratégia de controle diferenciado para o erro padrão da proficiência. Delineiam-se os principais objetivos do modelo bem como as principais proposições para atender a alguns dos fatores que compõem os TAC's.

No capítulo 4 será apresentada a metodologia de definição das características básicas dos TAC's: a composição do banco de itens, o método de estimação da proficiência, a simulação da resposta dos indivíduos aos itens dos testes e os critérios de parada. Também serão apresentados os fatores específicos necessários para que os três modelos de seleção

possam ser simulados: a forma de estratificação do banco de itens, o método utilizado para controle de exposição de itens e os parâmetros para o cálculo das metas do erro padrão da proficiência.

A análise sobre os resultados obtidos nas simulações dos testes utilizando o modelo proposto em comparação com as técnicas mais tradicionais serão abordadas no capítulo 5. No capítulo 6 serão feitas as observações e considerações finais sobre o comportamento do modelo proposto e as possibilidades de uma futura evolução do mesmo.

2 TESTES ADAPTATIVOS E SELEÇÃO DE ITENS

Este capítulo dedica-se a apresentação e discussão das principais características dos Testes Adaptativos Computadorizados, com foco no procedimento de seleção de itens. Serão apresentados também os fundamentos da Teoria da Resposta ao Item, que compõe o conjunto de modelos matemáticos que permitiram aos TAC's evoluírem significativamente a partir dos testes tradicionais. Ao final do capítulo serão abordados os dois modelos de seleção de itens que serviram como parâmetros de comparação nas simulações deste trabalho.

2.1 TESTES E MEDIDAS DE PROFICIÊNCIA

Conhecimento, habilidade ou capacidade de realizar determinadas tarefas são características que as pessoas possuem, mas que, para poderem ser mensuradas, é necessário que seja desenvolvido algum tipo de teste.

Quando é necessário saber quais atletas são capazes de competir em alto nível para participarem dos Jogos Olímpicos, determina-se um índice mínimo de desempenho a ser alcançado em provas do esporte em questão. Se for necessário selecionar entre vários candidatos a uma vaga de emprego, pode-se aplicar uma prova prática e verificar qual deles tem mais habilidade para aquela determinada tarefa.

Da mesma forma, se queremos avaliar o nível de conhecimento de uma pessoa sobre uma determinada área ou assunto é necessário que apliquemos um teste de conhecimentos. Tradicionalmente, nos baseamos na quantidade de questões certas e erradas em um teste para avaliarmos o desempenho do indivíduo. Essa forma tradicional de medida é chamada de Teoria Clássica dos Testes (TCT), sendo de fácil interpretação. Porém, essa característica de simplicidade limita a TCT (BAKER, 2001).

Nem sempre podemos garantir que o desempenho de um aluno que acerte 50% do teste é o mesmo de outro aluno com esse mesmo percentual de acerto. Eles, provavelmente, não acertaram as mesmas questões do teste, portanto, um deles pode ter alguns conhecimentos mais avançados que o outro.

Da mesma forma, se a um aluno são aplicados dois testes distintos, em épocas distintas, com o mesmo obtendo um percentual de acerto 10% superior no segundo teste, não podemos afirmar que seu conhecimento aumentou no período entre os testes, uma vez que os testes podem ter um nível de dificuldade diferente. Para isso, teríamos que determinar se o segundo teste é, realmente, mais difícil do que o primeiro e, também, o quanto mais difícil.

Assim, os resultados dos indivíduos podem variar de teste para teste, dependendo dos conteúdos, fazendo com que seja difícil comparar o desempenho de pessoas aplicando-se testes diferentes e limitando a validade do instrumento de medida. Foi necessário, então, o desenvolvimento de ferramentas que permitissem às avaliações contornar as limitações da TCT e fornecer resultados matematicamente embasados.

2.1.1 TEORIA DA RESPOSTA AO ITEM - TRI

A partir da década de 1960, trabalhos como os de Rasch (1960) e Lord e Novick (1968) impulsionaram o desenvolvimento da Teoria da Resposta ao Item (TRI), trazendo uma nova forma de avaliar o conhecimento e deixando de lado a subjetividade implícita nos métodos clássicos de avaliação. A TRI trabalha com diversos modelos probabilísticos que atendem a uma variedade de testes e avaliações aplicados em todo tipo de área: testes psicológicos, avaliações educacionais, indicadores socioeconômicos, escalas de concordância ou satisfação e várias outras medidas.

Devido ao grande número de modelos existentes, cada área de aplicação deve avaliar quais modelos se adaptam melhor às suas necessidades (BAKER, 2001). Por exemplo, um questionário para um indicador socioeconômico ou uma prova dissertativa utiliza itens politômicos, itens em que há gradações de valores para cada resposta, logo, somente os modelos específicos para esses tipos de itens poderão ser aplicados. Em testes com questões objetivas, com apenas uma resposta correta, são utilizados modelos probabilísticos dicotômicos. Muitas vezes há testes em que são utilizados dois tipos de modelos concomitantemente (KOLEN; BRENNAN, 2004).

Na TRI os itens apresentam determinadas características denominadas parâmetros, os quais, em conjunto com a habilidade ou proficiência dos indivíduos, geram uma função de probabilidade de acerto quando os itens são respondidos (BAKER, 2001). Dadas as probabilidades de acerto dos itens que compõem um teste, uma função de verossimilhança

estima qual o valor da proficiência que melhor corresponde ao padrão de respostas, corretas ou incorretas, apresentadas por um indivíduo a esses itens. Para avaliações que utilizam questões objetivas os modelos mais utilizados são os logísticos de um, dois ou três parâmetros e, especificamente para avaliações educacionais, o modelo logístico de três parâmetros tem sido o mais amplamente utilizado (LINDEN; HAMBLETON, 1996).

A principal característica da TRI, e também sua grande vantagem, é o parâmetro de dificuldade do item e a estimativa de proficiência do indivíduo estarem na mesma escala, permitindo obter um posicionamento das probabilidades de acerto dos itens por indivíduo. Assim, pode-se avaliar o nível em que a pessoa avaliada se encontra em relação a todos os itens que estejam naquela escala utilizada e, por consequência, a seus conteúdos. É importante ressaltar a necessidade de que os testes sejam bem construídos, com itens que cubram as diversas regiões da escala de dificuldade. Um banco de itens de baixa qualidade implica em um instrumento de medida mal construído que não consegue estimar corretamente a proficiência dos indivíduos avaliados.

Devido aos itens dos testes passarem a ter características individuais, obtidas através de seus parâmetros, podemos fazer uma análise do comportamento do teste item a item, e não somente do teste como um todo. Uma vez que os itens utilizados estão dispostos em uma mesma escala, passa também a existir uma independência dos resultados dos indivíduos examinados em relação ao teste utilizado. Mesmo que os testes sejam diferentes, sejam aplicados em épocas ou anos de escolaridade diferentes, esses resultados podem ser comparados pois foram gerados dentro de uma mesma escala.

Desde meados dos anos 1980 a TRI vem se tornando a técnica predominante no campo dos testes e avaliações, notadamente em avaliações educacionais. O avanço da informática nesse período permitiu o desenvolvimento de programas computacionais que minimizaram o problema da complexidade dos métodos estatísticos utilizados tornando a TRI mais acessível aos pesquisadores da área de avaliação (PASQUALI; PRIMI, 2003).

2.1.2 MODELO LOGÍSTICO DE TRÊS PARÂMETROS (M3PL)

Entre os diversos modelos matemáticos desenvolvidos dentro da TRI, o modelo logístico de três parâmetros de Birnbaum (1968) é, atualmente, o mais utilizado em avaliações de conhecimento com itens de resposta objetiva e será o modelo utilizado em todos os experimentos deste trabalho. Seu desenvolvimento se baseou nos modelos já existentes à

época, mantendo os dois parâmetros que já eram utilizados, dificuldade e discriminação, e adicionando um componente de probabilidade mínima de acerto.

Conforme visto anteriormente, na TRI os itens são posicionados na mesma escala de habilidades das proficiências dos indivíduos e o parâmetro de dificuldade (parâmetro b) é que determina o ponto dessa escala em que aquele item se encontra. Por exemplo, um item de parâmetro b muito baixo será considerado um item fácil, pois terá uma alta probabilidade de acerto, mesmo por indivíduos que não tenham uma proficiência muito alta.

A discriminação (parâmetro a) determina a capacidade do item de diferenciar os indivíduos que têm maior ou menor probabilidade de responder corretamente a um item, dada a sua dificuldade. Assim, quanto maior for esse parâmetro, maior será a capacidade do item de mensurar o conhecimento dos indivíduos avaliados, obtendo uma estimativa mais precisa das habilidades.

O que diferenciou esse modelo, e que provavelmente o tornou tão atraente para as avaliações educacionais, foi a introdução de uma probabilidade mínima de acerto, mesmo em casos de proficiência muito baixa dos avaliados. Esse terceiro parâmetro (parâmetro c) é comumente chamado de probabilidade de acerto casual, porém ele não é a simples probabilidade de acerto ao acaso de acordo com o número de alternativas de resposta do item, uma vez que os parâmetros influenciam uns nos outros quando estão sendo calculados. Assim a variação da discriminação, ou da dificuldade, também tem influência no valor desse percentual de probabilidade de acerto casual.

De acordo com esse modelo, a probabilidade de um indivíduo de proficiência θ acertar um item é dada por (COSTA, 2009):

$$P(\theta) = c + \frac{1 - c}{1 + e^{-D.a.(\theta - b)}} \tag{2.1}$$

onde:

a é o parâmetro de discriminação do item;

b é o parâmetro de dificuldade do item;

c é o parâmetro de acerto casual;

θ é a estimativa de proficiência do indivíduo;

D é um fator de escala em que se utiliza o valor 1,7 para que a função logística forneça resultados semelhantes aos da função normal.

A chamada curva característica do item (CCI) é a representação da associação entre a estimativa da proficiência e a probabilidade de acerto em um item. A Figura 2.1 mostra a CCI para um item de parâmetro a igual a 0,8, parâmetro b igual a 0,35 e parâmetro c igual a 0,12.

Podemos observar que quanto maior for o parâmetro a, maior será a inclinação da curva, logo, maior será a diferenciação que o item fará entre as probabilidades dos indivíduos acertarem um item de acordo com suas estimativas de habilidade. Da mesma forma, uma variação no parâmetro b faz com que a curva se desloque para a esquerda, se o item for mais fácil, ou para a direita, se for mais difícil. O parâmetro c, no início da curva, mostra a probabilidade mínima de acerto associada ao item sendo que, quanto maior esse valor, pior é a qualidade do item, uma vez que não haveria uma diferenciação muito grande da probabilidade de acerto entre indivíduos com proficiências razoavelmente diferentes.

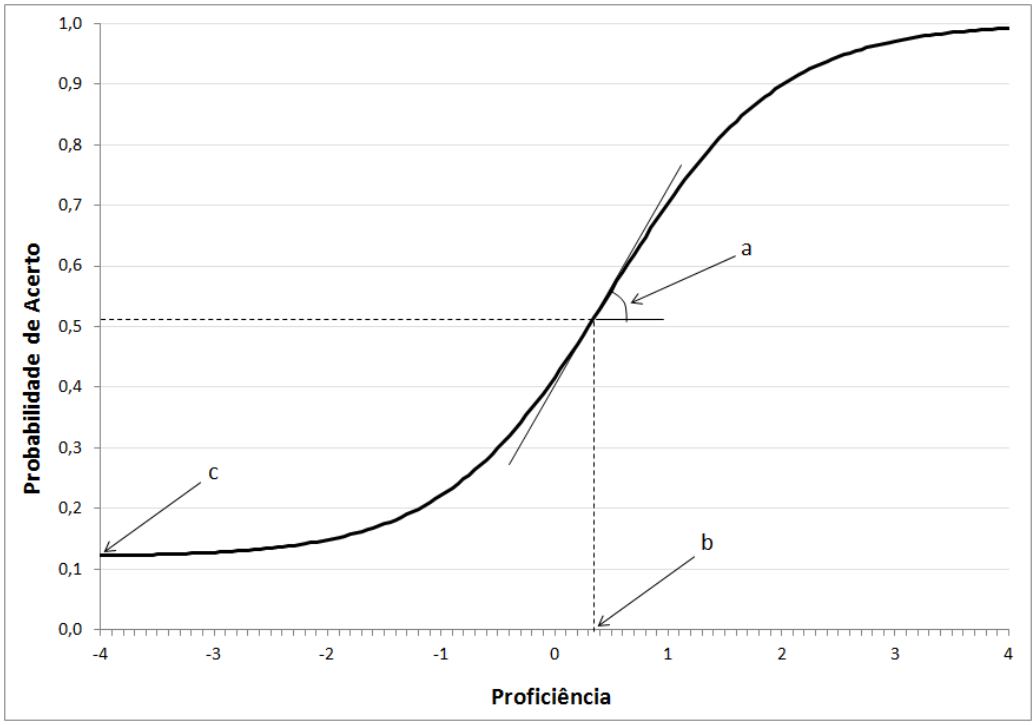

Figura 2.1: Curva característica de um item pelo modelo de três parâmetros

2.2 TESTES ADAPTATIVOS COMPUTADORIZADOS (TAC'S)

A possibilidade da análise dos parâmetros dos itens e da proficiência na mesma escala obtida com a TRI levaram à evolução dos chamados testes adaptativos, que são testes

sequenciais iterativos onde os itens são escolhidos um após o outro se adaptando ao conhecimento/habilidade do respondente (LINDEN; GLAS, 2000). Assim, durante a aplicação do teste, pode-se perceber a região da escala em que a estimativa de proficiência do indivíduo avaliado se encontra, criando um teste específico e adaptado, o que pode garantir uma precisão maior dos resultados.

Em um teste adaptativo, a ideia básica é que a seleção do próximo item depende do resultado do indivíduo até aquele momento. A cada nova seleção, aplicação do item e obtenção da resposta por parte do avaliado, a proficiência é reestimada servindo de referência para a seleção do próximo item a ser utilizado. No início do teste, quando ainda não há nenhuma informação sobre a proficiência, admite-se que o indivíduo pode estar em uma região próxima à média da escala, sendo essa informação utilizada como ponto de partida para a seleção do primeiro item.

Os primeiros estudos sobre Testes Adaptativos Computadorizados (TAC's) tiveram início na década de 1980, mas foi com a popularização da informática no início da década de 1990 que os TAC's se tornaram frequentes nos países em que as avaliações e testes pela TRI já eram prática comum (WAINER, 2000). Assim como os testes adaptativos tradicionais, os TAC's também são testes sequenciais iterativos, porém, os indivíduos avaliados respondem ao teste utilizando um computador. A Figura 2.2 mostra os passos de funcionamento de um sistema de TAC.

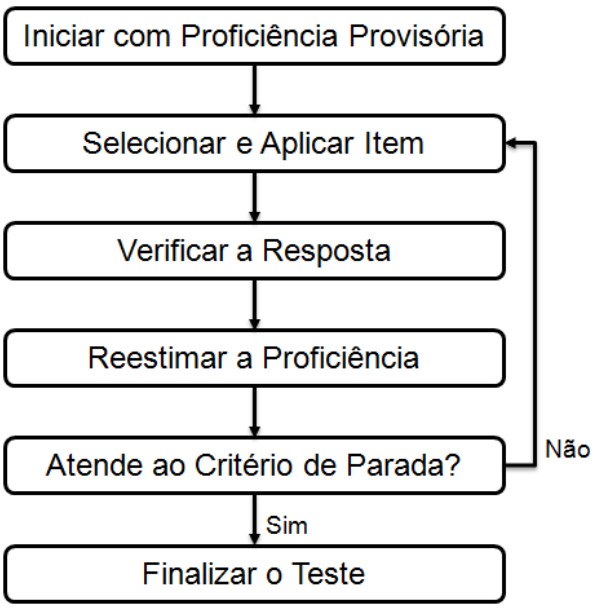

Figura 2.2: Esquema de representação de um Teste Adaptativo Computadorizado

Os testes adaptativos seguem um método iterativo de funcionamento, portanto, é necessário que se estabeleça, pelo menos, um critério de parada. No caso dos TAC's é muito comum que os critérios de parada se baseiem em duas características, uma relacionada à quantidade de itens utilizados e outra à precisão do teste (WEISS; KINGSBURY, 1984).

O primeiro critério a ser estabelecido é o de um número máximo de itens a serem aplicados em um teste. É comum que um teste pré-definido que busca precisão tenha muitos itens, como, por exemplo, o teste do Exame Nacional do Ensino Médio (ENEM) do Ministério da Educação do Brasil (INEP, 2012). Na edição do ano de 2012, o ENEM teve 45 itens aplicados em cada disciplina avaliada, visando abranger os diversos conteúdos e seus níveis de dificuldade. Um dos objetivos de um teste adaptativo é que o número de itens utilizados possa ser reduzido e, assim, é necessário que se estabeleça um limite para o tamanho do teste, limite este geralmente inferior aos utilizados nos testes comuns.

O segundo critério é o da precisão do teste controlada através da estimativa do erro padrão da proficiência. O teste deve ter um valor do erro padrão previamente estabelecido para que seja considerado que a habilidade foi estimada corretamente. Uma vez que a estimativa do erro padrão atinja um valor abaixo de um limite estabelecido, o teste pode ser considerado como finalizado.

Na utilização de um sistema de TAC, esses dois critérios de avaliação para a finalização de um teste estão interligados e devem ser cuidadosamente definidos. Deve-se procurar um equilíbrio entre eles, para que a busca por um resultado otimizado para um dos critérios não cause conflito com o outro.

Por exemplo, é possível que, na tentativa de reduzir o número de itens utilizados, o limite máximo de itens definido para o teste não seja suficiente para atingir a precisão desejada ou, pelo menos, um valor razoável do erro padrão. Da mesma forma, se definirmos uma precisão que seja excessivamente criteriosa, pode ser necessário aumentar muito o número de itens utilizados para atingi-la ou, em alguns casos, pode não ser possível atingi-la através do método de estimação da proficiência que estiver sendo utilizado.

A grande vantagem de um TAC está exatamente em podermos analisar o comportamento do indivíduo avaliado durante o processo e usarmos essa informação para obtermos maior precisão, selecionando os itens mais adequados à proficiência verdadeira.

A precisão na estimativa da proficiência depende diretamente que os itens aplicados no teste avaliem, adequadamente, na região da escala de habilidades em que o indivíduo se

encontra. Se, durante esse processo, a seleção dos itens em um TAC caminha na direção errada, a proficiência estimada estará incorreta e esse método de seleção se mostrará ineficiente.

2.3 SELEÇÃO DE ITENS EM TAC'S

Quando um teste é desenvolvido na forma tradicional, ou seja, um teste com as questões definidas previamente, um ponto importante a se observar é a presença de itens que cubram todos os níveis de dificuldade, de forma a garantir que indivíduos com proficiências em pontos diferentes da escala possam ser avaliados com precisão. Além disso, essa necessidade de cobrir toda a escala de habilidades acarreta o aumento do número de itens que compõem o teste, uma vez que não se sabe, em nenhum momento da aplicação, em que ponto se encontra a proficiência do indivíduo.

No modelo de funcionamento de um TAC, o procedimento de seleção de itens é o responsável por criar o teste de forma adaptativa, tendo a função de buscar a região específica da escala em que o teste melhor se adapta ao indivíduo avaliado. Evidentemente, um dos fatores primordiais para que o modelo de seleção de itens possa ser efetivo é a garantia que o banco de itens utilizado seja compatível com seu objetivo. Isto é, o banco deve ser composto por itens que atinjam todo o intervalo da escala que se deseja avaliar, caso contrário, não será possível selecionar os itens que proporcionarão a melhor indicação sobre a estimativa correta de proficiência.

Normalmente, a distribuição dos itens tende a aproximar-se de uma distribuição normal (PASQUALI; PRIMI, 2003) e esse fator pode levar à imprecisão das estimativas dos indivíduos que se encontram nas extremidades da escala, por ser uma região que conta com menos itens. A Figura 2.3 mostra a variação do erro padrão da proficiência no teste de Língua Portuguesa da 8ª série do ensino fundamental do Sistema de Avaliação da Educação Básica (SAEB) do Ministério da Educação do Brasil em 2003. Podemos observar que nas extremidades da escala o erro padrão chega a ser 50% maior que nas faixas centrais, de maior concentração de itens.

A partir do momento em que os fatores básicos e primordiais de um teste, como tamanho, precisão e garantia de qualidade do banco de itens estão definidos, os esforços se concentram na etapa do processo do TAC que se utiliza desses fatores para, efetivamente, cumprir com o objetivo de criar um teste adaptativo. O modelo de seleção de itens se torna

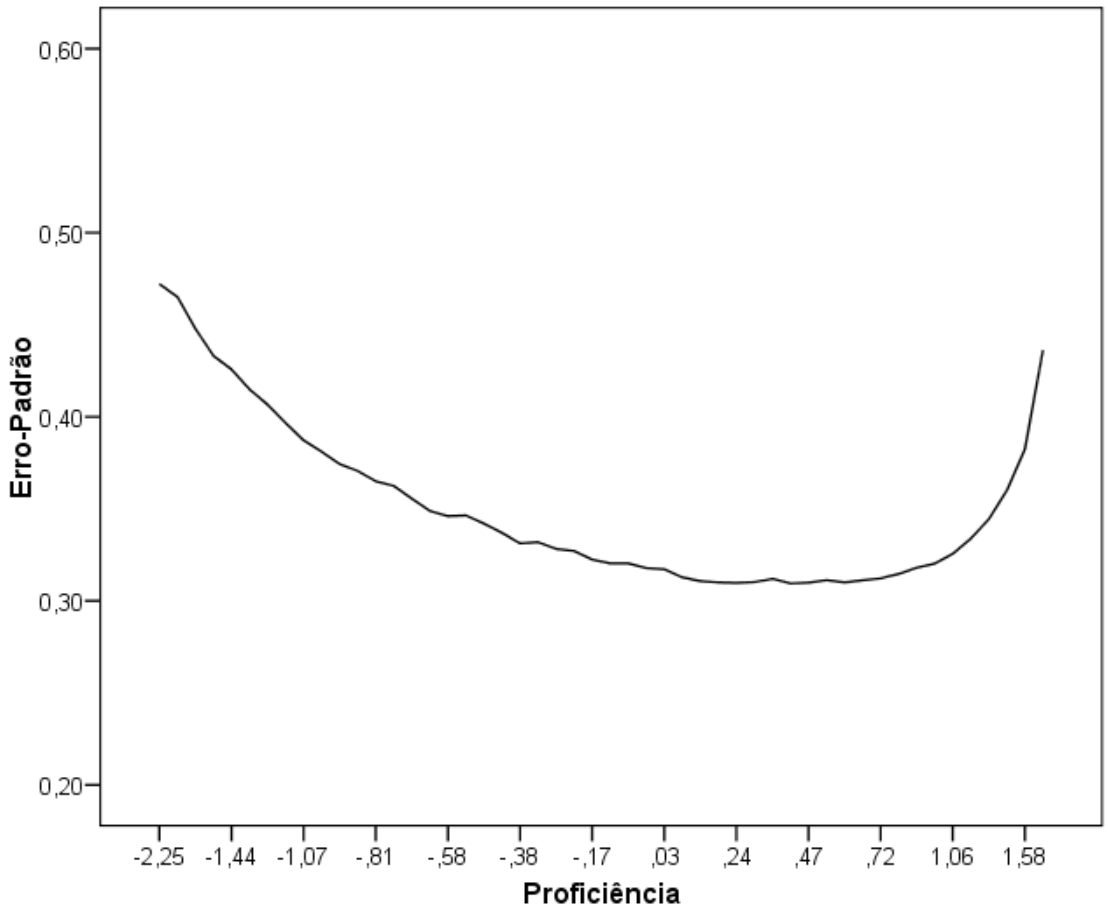

Figura 2.3: Distribuição do erro padrão pela escala de proficiência em Língua Portuguesa na 8ª série do SAEB 2003

o ponto principal para garantir a veracidade da proficiência (LINDEN; GLAS, 2000).

O objetivo de um modelo de seleção é tentar prever, de forma aproximada, a estimativa que será apresentada ao final da aplicação do teste. O foco da seleção de itens está na forma como o modelo reage ao comportamento do indivíduo que está sendo avaliado para fazer essa previsão. Se o modelo não consegue se adaptar às possíveis variações das respostas durante a aplicação do teste, há um sério risco de seleção de itens que não contribuem para a precisão do resultado. Um modelo de seleção de itens pouco eficiente leva a uma estimativa, no mínimo, imprecisa da proficiência do indivíduo. Como um indivíduo de proficiência alta pode ser avaliado corretamente se a ele são aplicados apenas itens fáceis? E como definir em que região da escala se encontra, e mais especificamente, qual o próximo item deve ser selecionado para compor o teste?

Além da produção de uma estimativa precisa para a proficiência, outro ponto a se considerar no comportamento de um modelo de seleção de itens é o gerenciamento da

exposição dos itens. Em TAC's, os mesmos itens ficam disponíveis para utilização por um período relativamente longo. O modelo deve considerar estratégias para tentar equilibrar o número de vezes em que os itens são aplicados nos testes. Esse equilíbrio reduz as chances de que os itens do banco sejam previamente conhecidos pelos indivíduos que responderão ao teste, evitando um viés nos resultados obtidos (BARRADA et al., 2008).

Diversas propostas de modelos para seleção de itens surgiram desde que os testes adaptativos começaram a ser desenvolvidos, porém, esses modelos seguem, geralmente, duas linhas de estratégia: a de classificação de acordo com níveis de dificuldade, a de quantidade de informação agregada ao teste pelos itens. A seguir, apresentam-se os dois modelos tradicionais de seleção de itens utilizados para comparações com o modelo proposto neste trabalho.

2.3.1 SELEÇÃO BASEADA EM NÍVEIS DE DIFICULDADE

Os primeiros testes adaptativos baseavam a seleção de itens nos parâmetros de dificuldade e na resposta do indivíduo ao item anterior. Se a pessoa avaliada acertasse um item, o próximo item selecionado seria mais difícil e, se errasse, seria mais fácil. Posteriormente, para facilitar essa seleção, os itens passaram a ser divididos em diversas faixas de acordo com sua dificuldade e selecionados quando a estimativa de proficiência estivesse dentro de uma das faixas.

Porém, os critérios utilizados para a criação dessas faixas pode influenciar negativamente no processo de funcionamento do TAC e nem sempre os itens que realmente trouxessem algum ganho à precisão da proficiência seriam selecionados para aplicação. Muitas vezes a amplitude da faixa, o número de itens por faixa e uma seleção totalmente aleatória de um item podem influenciar de forma negativa no processo de estimação.

Por exemplo, devido às características de distribuição dos itens na escala, com a concentração maior de itens nas regiões centrais, se a estratificação do banco utilizar faixas de mesma amplitude haverá uma grande diferença entre o número de itens disponíveis nas faixas da extremidade e nas faixas centrais. Da mesma forma, se a amplitude das faixas for grande, a escolha de um item aleatoriamente dentro de uma determinada faixa pode significar a seleção de um item distante da estimativa de proficiência por se situarem em extremos diferentes da faixa.

Algumas variações e adaptações foram feitas aos métodos de estratificação durante os

anos de desenvolvimentos dos TAC's, resultando em modelos mais refinados e eficientes, principalmente na função de administrar a exposição dos itens. Propostas como as de Weiss (1973) e Cronbach (CRONBACH et al., 1972), já na década de 1970, começavam a visualizar o avanço da informática como forma de impulsionar a melhoria dos modelos de seleção estratificada existentes à época.

Um modelo recente de estratificação do banco de itens foi proposto por Chang et al. (2001), a partir de uma fusão entre o método de estratificação de Weiss e o método de controle de exposição e gerenciamento do banco de itens de Chang e Ying (1999), e foi simulado para este trabalho como um dos parâmetros de comparação de resultados. Nesse procedimento, o banco de itens é estratificado com base nos valores dos parâmetros de discriminação e de dificuldade dos itens e o teste adaptativo é dividido em estágios.

O primeiro passo nesse modelo é estabelecer o número de estratos que serão gerados. É importante salientar que o teste será dividido em estágios de acordo com a quantidade de estratos pelo parâmetro a e a divisão pelo parâmetro b deve levar em consideração o número total de itens que o banco possui, de forma que não sejam criados subgrupos com poucos itens.

Os itens são ordenados de forma ascendente pelo parâmetro de discriminação e divididos em poucos estratos. Em seguida, o banco de itens é reordenado pelo parâmetro de dificuldade dos itens e dividido em vários estratos de tamanho menor. Assim, são formados diversos estratos de itens de acordo com suas classificações por parâmetro de dificuldade e, em cada um deles, encontraremos subgrupos de itens separados de acordo com sua discriminação.

O teste deve ser dividido em um número de estágios de acordo com o número de estratos do parâmetro a, sendo que os itens são selecionados no estrato de dificuldade em que a estimativa de proficiência se encontra e no subgrupo de discriminação correspondente ao estágio em que o teste se encontra. Assim, no início do teste, são selecionados os itens de menor parâmetro de discriminação e, à medida que o teste evolui, serão selecionados itens mais discriminativos que pertencem aos diferentes estratos de dificuldade.

A principal justificativa para que o procedimento de estratificação inclua o parâmetro a é que, no início do teste, a imprecisão na estimativa da proficiência ainda é muito alta e, então, a quantidade de informação agregada ao teste pelos itens de maior discriminação não é realmente necessária. Assim, são utilizados primeiro os itens de discriminação mais

baixa, poupando os itens de maior discriminação para os estágios finais do teste (CHANG; YING, 1999).

Esse modelo de seleção também indica que, devido a essa imprecisão inicial, os primeiros cinco itens podem ser selecionados de forma aleatória dentro do subgrupo que será utilizado. Nos itens subsequentes, deve-se definir como critério para seleção a proximidade entre a estimativa atual da proficiência e o parâmetro de dificuldade dos itens, ou seja, dentro do subgrupo em que ocorrerá a seleção deve-se selecionar o item mais próximo da estimativa de proficiência atual.

A Figura 2.4 mostra um exemplo do subgrupo onde um item é selecionado quando o teste se encontra em sua etapa final, isto é, quando são selecionados os itens do estrato de discriminação mais alta. Nesse exemplo, a estimativa atual de proficiência se encaixa no nono estrato de dificuldade e, assim, será selecionado o item cujo parâmetro de dificuldade é o mais próximo dessa estimativa de proficiência dentro desse subgrupo formado pela interseção dos dois estratos.

Estratos *b*	Estratos *a*			Total
	1º	2º	3º	
1º	19	21	18	58
2º	16	20	22	58
3º	22	16	20	58
4º	12	18	28	58
5º	22	21	15	58
6º	22	21	15	58
7º	23	19	16	58
8º	21	19	18	58
9º	21	19	18	58
10º	17	21	24	62
Total	195	195	194	584

Figura 2.4: Exemplo de subgrupo de estratos onde um item é selecionado

Os modelos baseados em estratos são de implementação simples e garantem uma variedade maior dos itens aplicados, numa tentativa de se evitar a superexposição dos mesmos. Porém, são muito dependentes da subjetividade para a definição dos estratos, principalmente os estratos pelo parâmetro de dificuldade, o que afeta a amplitude da divisão das faixas ou a quantidade de itens por estrato, o que pode levar a uma imprecisão maior nas estimativas produzidas. Este modelo é implementado nesse trabalho visando servir

de referência e comparação em relação a outras técnicas.

2.3.2 SELEÇÃO BASEADA EM MEDIDA DE INFORMAÇÃO

Desde o início do desenvolvimento dos TAC's, diversas alternativas surgiram buscando estabelecer modelos de seleção de itens que não dependessem de uma classificação baseada em conceitos subjetivos, como nos modelos de divisão por faixas ou estratos, e que utilizassem cálculos matemáticos como critério fundamental para embasar a seleção. Foram propostos modelos com os mais diversos critérios de seleção de itens, por exemplo, por seleção Bayesiana (OWEN, 1975), pelo critério de Máxima Informação Global (CHANG; YING, 1996), de Máxima Informação Esperada (LINDEN, 1998), da Informação da Verossimilhança Ponderada (VEERKAMP; BERGER, 1997) e vários outros.

Os modelos mais utilizados atualmente se baseiam em métodos estatísticos para medir a quantidade de informação agregada pelo item ao teste, sendo que o modelo de Máxima de Informação de Fisher (MIF) proposto por Lord (1980) tornou-se o mais conhecido e utilizado entre eles (VELDKAMP, 2010). Na TRI, a Informação de Fisher permite analisar quanto um item agrega em termos de informação e, subsequentemente, quanto ele acrescenta em eficiência ao teste na produção da estimativa de habilidade. Essa medida de informação é calculada para cada item individualmente a partir dos seus parâmetros.

Segundo o modelo de MIF, o próximo item a ser selecionado para aplicação será aquele que apresentar a maior quantidade de informação, dada a proficiência estimada no momento da seleção. A justificativa teórica para esse procedimento de seleção de itens é a possibilidade um ganho substancial na eficiência do teste (CHANG; YING, 1999). Para o modelo logístico de três parâmetros, essa medida de informação do item é dada por:

$$I(\theta_s) = D^2.a_i^2.\left[\frac{Q_i(\theta_s)}{P_i(\theta_s)}\right].\left[\frac{P_i(\theta_s) - c_i}{1 - c_i}\right]^2 \tag{2.2}$$

onde:

θ_s é a estimativa atual da habilidade;

a_i é o parâmetro de discriminação do item i;

c_i é o parâmetro de acerto casual do item i;

D é um fator de escala em que se utiliza o valor 1,7 para que a função logística forneça resultados semelhantes aos da função normal;

$P_i(\theta_s)$ é a probabilidade de resposta correta ao item i, pelo modelo logístico de três parâmetros, dada a habilidade θ_s;

$Q_i(\theta_s)$ é a probabilidade de resposta incorreta ao item i, pelo modelo logístico de três parâmetros, dada a habilidade θ_s, sendo $Q_i(\theta_s) = 1 - P_i(\theta_s)$.

A maximização da Informação de Fisher é um procedimento determinístico, uma vez que esse modelo busca selecionar um item específico, com parâmetro de dificuldade próximo à estimativa atual de proficiência do indivíduo avaliado e que tenha a maior discriminação possível. Dessa forma, o método garante apenas que, para a estimativa de proficiência atual, até aquele momento do teste, o próximo item é o melhor a ser aplicado. Porém, a estimativa atual, principalmente nos primeiros itens do teste, pode estar longe da proficiência verdadeira do indivíduo. Assim, a alta informação que esse item agrega ao teste corre o risco de se tornar ineficaz e, à medida que o teste evolui, um item de alta discriminação teria sido utilizado de forma ineficiente (CHEN et al., 2000). Chang e Ying (1996) argumentam que pode até ser mais vantajoso não utilizar a informação do item nos estágios iniciais do teste, de forma a evitar a perda de eficiência devido à imprecisão da estimativa de proficiência baseada em um pequeno número de itens.

Assim como nos modelos de estratificação, esse modelo também pode ser visto apenas como uma solução local para a seleção de itens, isto é, leva em consideração apenas a estimativa atual da proficiência para selecionar o próximo item. Não há nenhuma indicação de estratégia ou, pelo menos, de uma sequência de escolhas para adicionar informação que possa auxiliar na decisão da seleção dos itens subsequentes.

Além disso, para uma determinada região da escala de proficiência existe apenas um item que maximiza a função de informação, ou seja, sempre que a estimativa atingir aquela região da escala o mesmo item será selecionado. É comum que, no início do teste, a estratégia utilizada seja atribuir aos indivíduos uma proficiência provisória, próxima à média da escala, para servir de referência para a seleção do primeiro item e, assim, segundo o modelo de MIF, todos os testes começariam com o mesmo item.

Um modelo de seleção de itens baseado em medida de informação pode levar a taxas irregulares de exposição de itens, isto é, alguns itens podem ser frequentemente selecionados em um TAC enquanto outros talvez nunca sejam usados (CHANG; YING, 1996). O problema da superexposição de itens deve ser controlado, garantindo que o número de vezes em que os itens são aplicados seja equilibrado, melhorando a segurança e confiabi-

lidade do teste. O método de controle de exposição deve avaliar a possível substituição de um item selecionado por outro que tenha uma frequência menor de exposição. No entanto, é importante ressaltar que, quando se impede a seleção de um item para evitar a superexposição, há interferência direta no modelo de Máxima Informação de Fisher, uma vez que o item de maior informação deixa de ser utilizado.

Diferentes métodos para implementação do controle de exposição dos itens já foram propostos, como os apresentados por Stocking e Lewis (1995), Davey e Parshall (April 1995), Chang e Ying (1996), Linden (2003), entre outros. Segundo Linden (2003), atualmente, o mais popular entre os métodos de controle de exposição de itens em TAC's é o proposto por Hetter e Sympson (1997) e, por isso, foi o escolhido para ser implementado e utilizado como referência neste trabalho, sendo apresentado e discutido na seção 4.6.

3 SELEÇÃO DE ITENS POR METAS DO ERRO PADRÃO - MODELO MEP

Este capítulo se dedica à apresentação da proposta de um modelo de seleção de itens baseado em metas definidas para o erro padrão da proficiência resultante da aplicação do teste. O modelo proposto foi desenvolvido para tentar controlar o erro local, com metas que devem ser cumpridas a cada seleção de itens visando atingir um objetivo global, isto é, a meta definida para a precisão da estimativa de proficiência ao final do teste. A estratégia para a seleção de itens é baseada na previsão da variância *a posteriori* a ser obtida caso um item seja selecionado.

Esse modelo adapta a ideia de meta global na montagem de cadernos de testes tradicionais, apresentada por Verschoor (2007), para atender a um modelo de seleção de itens baseado na variância *a posteriori*, similar à proposta apresentada por Linden e Pashley (2000).

Verschoor (2007) apresentou um modelo de escolha de itens para a montagem de diferentes cadernos de testes, selecionando itens pela quantidade de informação agregada ao teste utilizando-se da técnica de algoritmos genéticos (GOLDBERG, 1989). Esse tipo de técnica busca encontrar soluções em problemas de otimização através de princípios inspirados na biologia evolutiva, em que uma população inicial de possíveis soluções resulta, através de cruzamentos e mutações, em soluções mais adequadas. Essas, por sua vez, passam pelos mesmos procedimentos até que se consiga a solução mais adequada ao problema.

Normalmente os algoritmos genéticos têm como objetivo se aproximar da solução ótima para um problema, porém, o modelo de Verschoor (2007) propõe que o resultado atinja um valor suficiente de informação total no teste, e não o maior valor possível. Dessa forma, os diferentes tipos de cadernos produzidos apresentam um equilíbrio entre a quantidade de informação agregada em cada um, encontrando várias boas soluções e não apenas uma solução ótima. Esse modelo foi especificamente desenvolvido para a solução de um problema estático, que é a montagem de cadernos de testes pré-definidos, e, da forma como se encontra, não se aplica a um problema dinâmico como a seleção de itens para TAC's.

No caso do trabalho de Linden e Pashley (2000) o modelo propõe a seleção de itens para TAC's utilizando como critério a minimização da variância *a posteriori*, isto é, o modelo faz uma previsão de qual será a variância obtida após a aplicação dos itens restantes no banco e seleciona o que apresentar a menor das previsões. Assim como no modelo MIF, o item selecionado estará em uma região próxima da atual estimativa de proficiência do teste e o modelo indica que sejam testados os itens nessa região para a seleção daquele que produzir a menor variância. É importante salientar que essa estimativa da variância assume valores diferentes caso o indivíduo responda o item corretamente ou não. A solução do modelo é selecionar o item que produza o menor valor na soma dos dois casos que podem ser obtidos, de acordo com as possíveis respostas.

Uma característica que deve ser salientada em relação a esse modelo é a interação entre o próximo item candidato à seleção e os itens anteriores, pois, para que a variância seja calculada, o método de estimação de proficiências utiliza todos os itens já apresentados ao indivíduo, acrescido do item que está sendo testado para a possível seleção. Esse comportamento é contrário ao do MIF que baseia a seleção apenas na medida de informação obtida com a estimativa de proficiência atual, que pode não ter a precisão adequada, e não em uma medida que envolve todos os itens aplicados até aquele momento do teste, como é o caso da variância.

Porém, Segall (2004), que se refere a esse modelo de minimização da variância como MEV, salienta que, da mesma forma que o MIF, a solução obtida também é determinística, pois seleciona o item que, naquele momento do teste, minimiza a variância. Dada uma mesma proficiência provisória inicial, um determinado item será sempre selecionado para iniciar o teste por ser o que apresenta a menor variância *a posteriori*. Essa característica conduzirá o processo de seleção de itens a ter somente duas opções a cada passo seguinte, dependendo apenas da resposta do indivíduo ao item selecionado, com esse comportamento se repetindo para todos os itens subsequentes. Consequentemente, mesmo entre itens com parâmetros muito parecidos a taxa de exposição pode variar muito e o modelo MEV passa, também, a necessitar de um método de controle de exposição de itens.

A partir desses dois trabalhos foi possível conceber um modelo que pudesse fundir a ideia de uma meta suficiente, e não a melhor, defendida por Verschoor (2007), com um modelo similar ao MEV proposto por Linden e Pashley (2000). Esse modelo aqui proposto, chamado de modelo de metas do erro padrão (MEP), substitui a variância *a*

posteriori pelo valor do erro padrão da proficiência e busca definir um critério para atingir uma meta do erro ao fim da aplicação do teste, obtendo um bom nível de precisão que não seja, necessariamente, o menor possível. Como citado anteriormente, a seleção de itens em TAC's é um problema dinâmico devido à própria natureza do teste adaptativo e uma solução que projeta apenas a meta final do erro acabaria por trazer uma forma estática à solução ou torná-la determinística como o MEV.

Para que possamos estabelecer um critério sobre como o modelo define as metas a serem cumpridas no decorrer do teste, é necessário avaliar o comportamento do decaimento do erro padrão em um TAC. A Figura 3.1 mostra o decréscimo da média do erro padrão em simulações de TAC's em que os itens foram selecionados pelo modelo MIF. Podemos observar que a queda do valor segue um comportamento próximo ao exponencial, com uma variação maior na aplicação dos primeiros itens, demonstrando as observações feitas anteriormente sobre a imprecisão inicial do teste.

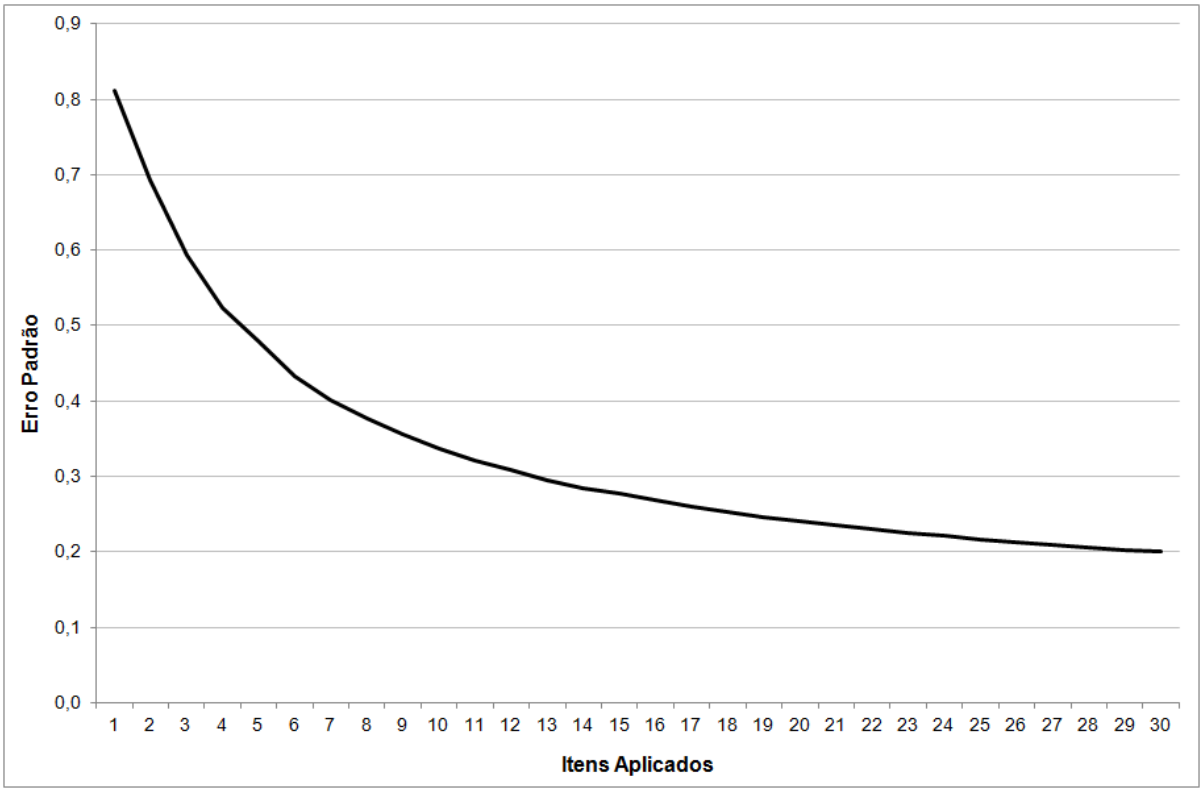

Figura 3.1: Comportamento do erro padrão em simulações de TAC's com seleção de itens por MIF

A primeira etapa de funcionamento do modelo envolve a definição de metas para o erro padrão a cada seleção de item através de um modelo com comportamento similar ao

exponencial, que prioriza um maior esforço para reduzir esse erro no início do teste. Para isso o modelo se baseia em uma progressão geométrica (PG). Justifica-se a utilização de uma PG para controle do erro padrão por item devido ao seu caráter 'discreto', onde cada termo estará associado a um item do teste e onde há a possibilidade de ajuste da razão da PG para melhor atender ao decaimento das metas. No processo de ajuste de uma PG ao padrão de erros de itens de um teste, o primeiro passo é adotar o termo inicial como a unidade. Além disto, adota-se uma relação inversa dos termos da PG com os itens do teste, ou seja, os primeiros itens do teste estão associados aos últimos e maiores termos da PG, enquanto os últimos itens do teste estão associados aos valores iniciais da PG. Essa associação define, de forma proporcional ao termo da PG, quanto o erro padrão deve cair na seleção do k-ésimo item do teste, fazendo com que o esforço para atingir a meta do erro padrão nos itens iniciais do teste sejam superiores às dos itens finais, em que o erro padrão cai vagarosamente.

Tomando o decaimento total como sendo a diferença entre o erro padrão inicial e a meta de erro estabelecida para a precisão ao fim do teste, o modelo faz a divisão proporcional das metas para cobrir este decaimento adotando um comportamento similar ao de divisão de lucros entre acionistas de uma empresa. Em uma divisão dessas, o lucro é dividido pelo número total de ações, obtendo um valor por ação, e depois paga-se esse valor proporcionalmente ao número de ações que cada acionista possui. No caso da divisão das metas do erro padrão, divide-se o decaimento total pela soma dos termos da PG e, em seguida, esse decaimento mínimo é multiplicado por cada termo da PG, que funciona como a quantidade de 'ações' que cada item possui. Assim, os itens iniciais, que estão associados aos maiores termos da PG, são os maiores 'acionistas' e, por isso, recebem as maiores partes do decaimento a serem cumpridas. O último item sempre estará associado ao termo inicial da PG, que é 1, e receberá o decaimento mínimo para cumprir, como se fosse um acionista que possui apenas uma ação da empresa. Deve-se ressaltar que o modelo de seleção de itens MEP pode ser formalizado através de um problema de otimização, onde se busca o argumento (item) que otimiza o desvio do decaimento do erro padrão real para o previsto.

Para definir o decaimento total a ser feito pela seleção de itens devemos obter primeiramente o erro padrão inicial. Para isso um primeiro item é selecionado aleatoriamente em uma região próxima à média da escala, assumindo provisoriamente, assim como em

outros modelos de seleção de itens, que a proficiência do indivíduo se encontra em torno da média. Logo, em um teste de tamanho n, todas as metas serão traçadas considerando-se apenas os $n-1$ itens restantes, uma vez que o primeiro item já foi utilizado para a obtenção do erro padrão inicial. A equação 3.1 identifica a soma dos termos da PG de razão q, também em função de $n-1$ itens.

$$s_t = \frac{q^{n-1} - 1}{q - 1} \tag{3.1}$$

Dado a soma dos termos da PG (s_t), o erro padrão inicial ($erro_1$) obtido após a aplicação do primeiro item e o erro padrão da proficiência esperado ao final do teste ($erro_n$), podemos determinar o decaimento mínimo δ_{min} como:

$$\delta_{min} = \frac{erro_1 - erro_n}{s_t} \tag{3.2}$$

O decaimento mínimo será multiplicado pelos termos da PG para obtermos o esforço δ_k necessário para redução do erro padrão na seleção do k-ésimo item, ou seja, quanto a seleção desse item deve fazer diminuir a meta para o erro padrão (eq. 3.3).

$$\delta_k = \delta_{min} \times q^{n-k} \tag{3.3}$$

Assim, a meta do erro padrão que serve de critério para a seleção do k-ésimo item é dada pela meta anterior menos o esforço calculado para o item k:

$$erro_k = erro_{k-1} - \delta_k \tag{3.4}$$

A Tabela 3.1 mostra um exemplo da definição de metas para um teste com o total de 30 itens, partindo de um erro padrão inicial de 0,85, com uma meta para o erro padrão final em 0,2 e utilizando uma PG de razão de 1,15. Dividindo-se o esforço total de 0,65 pela soma dos termos dessa PG ($\delta_t = 377,17$) obtém-se o decaimento mínimo $\delta_{min} = 0,00172$. Esse valor é multiplicado por cada termo da PG relacionado a cada item, exceto o primeiro, e assim obtemos o esforço δ_k e, consequentemente, a meta do erro padrão para cada seleção de item.

Uma vez determinadas as metas, o modelo deve, na seleção do k-ésimo item, verificar quais itens conseguiriam reduzir o erro padrão para atender à meta do $erro_k$ caso fossem

Tabela 3.1: Exemplo do cálculo de metas para o erro padrão a cada seleção de itens

Item k	Termo PG	Esforço δ	Erro Padrão
1			0,850
2	50,0656	0,08628	0,764
3	43,5353	0,07503	0,689
4	37,8568	0,06524	0,623
5	32,9190	0,05673	0,567
6	28,6252	0,04933	0,517
7	24,8915	0,04290	0,474
8	21,6447	0,03730	0,437
9	18,8215	0,03244	0,405
10	16,3665	0,02821	0,377
11	14,2318	0,02453	0,352
12	12,3755	0,02133	0,331
13	10,7613	0,01855	0,312
14	9,3576	0,01613	0,296
15	8,1371	0,01402	0,282
16	7,0757	0,01219	0,270
17	6,1528	0,01060	0,259
18	5,3503	0,00922	0,250
19	4,6524	0,00802	0,242
20	4,0456	0,00697	0,235
21	3,5179	0,00606	0,229
22	3,0590	0,00527	0,224
23	2,6600	0,00458	0,219
24	2,3131	0,00399	0,215
25	2,0114	0,00347	0,212
26	1,7490	0,00301	0,209
27	1,5209	0,00262	0,206
28	1,3225	0,00228	0,204
29	1,1500	0,00198	0,202
30	1,0000	0,00172	0,200

aplicados no teste. Porém, o erro padrão varia de acordo com a resposta do indivíduo ao item, isto é, o valor da estimativa do erro quando o item é respondido corretamente é diferente da estimativa de quando o item é respondido incorretamente. Esse comportamento tende a ser minimizado à medida que mais itens são respondidos, mas, como mais um dos fatores influenciados pela imprecisão do início do teste, nos primeiros itens essa diferença é significativa. A Figura 3.2 mostra a média das diferenças do erro padrão da proficiência previsto para a resposta correta ou incorreta dos itens aplicados em simulação de TAC's.

Devido a esse comportamento do erro padrão, a seleção dos itens é feita pela média dos valores previstos para o erro padrão, isto é, se a média entre os erros possíveis em caso de reposta correta ou incorreta ao item estiver abaixo da meta proposta para a k-ésima seleção, esse será um dos itens candidatos à seleção. A equação 3.5 define o conjunto de itens que atendem a esse critério como

$$C_i = \left\{ \forall i \mid \frac{erro_i^{rc} + erro_i^{ri}}{2} < erro_k \right\} \tag{3.5}$$

onde:

$erro_i^{rc}$ é a previsão de erro padrão em caso de resposta correta ao item;

$erro_i^{ri}$ é a previsão de erro padrão em caso de resposta incorreta ao item;

$erro_k$ é a meta do erro padrão da proficiência a ser cumprida após a seleção do $k-$ésimo item.

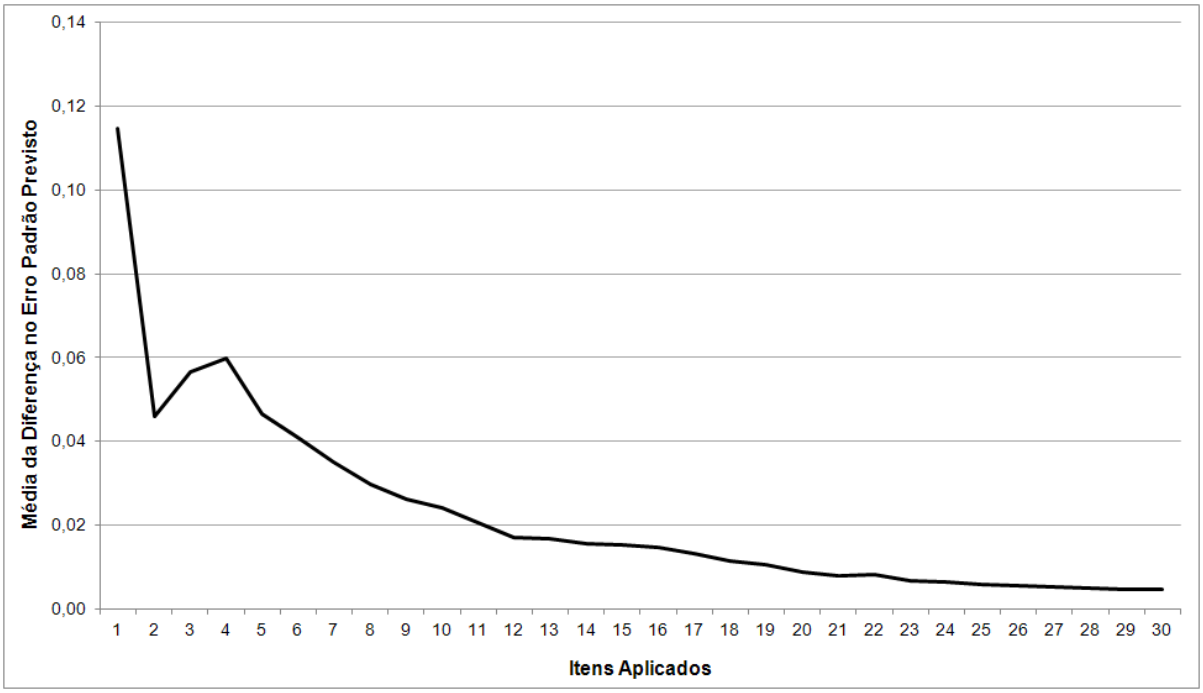

Figura 3.2: Diferença média das previsões do erro padrão da proficiência de itens aplicados em simulações de TAC's

Dentro do conjunto C_i de itens que atendem ao critério da meta, será selecionado aquele de menor distância euclidiana em relação ao item mais recentemente aplicado no teste. Podemos definir a distribuição dos itens, dados seus três parâmetros, como pontos em um espaço delimitado, conforme a Figura 3.3.

Pensando-se os itens existentes na base de dados como pontos definidos em três dimensões, de acordo com seus parâmetros a, b e c, abre-se uma nova perspectiva na busca de próximos itens por meio de procedimentos baseados em distância. De certa forma, pretende-se, neste método, adotar a escolha de itens de acordo com a estratégia utilizada em Aprendizagem de Máquina, tanto para aprendizagem supervisionada quanto não-supervisionada, conhecida como $K - NearestNeighbors$ (KNN)(MITCHELL, 1997).

O método KNN infere sobre uma nova instância de acordo com o padrão das K instâncias mais próximas. A adaptação à escolha de itens, da mesma forma, fará a inferência do próximo item baseando-se na distância dos K itens mais próximos ao mais recente item

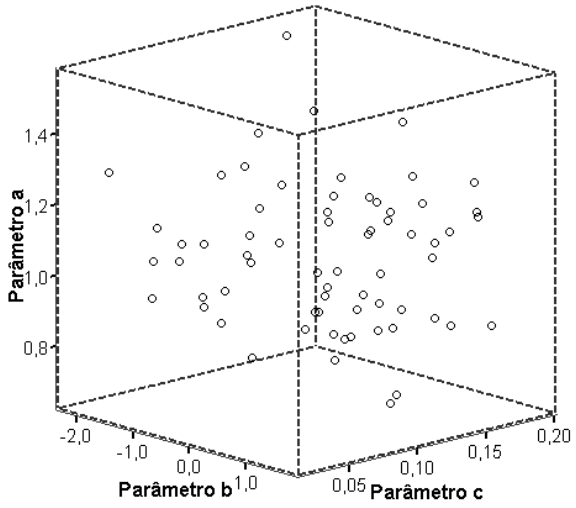

Figura 3.3: Exemplo de distribuição dos itens de acordo com seus parâmetros

aplicado. Utilizando-se $K = 1$, a escolha será determinada pelo vizinho mais próximo do item anterior. Logicamente, restrições serão necessárias e adotadas visando à obtenção de uma seleção mais efetiva, no que tange ao decaimento do erro padrão por metas.

Para definir a distância $d_{k,k-1}$ entre o mais recente item aplicado $(k-1)$ e os itens avaliados para a próxima meta (k), adota-se

$$d_{k,k-1} = \sqrt{(a_k - a_{k-1})^2 + (b_k - b_{k-1})^2 + (c_k - c_{k-1})^2} \tag{3.6}$$

onde, deve-se calcular a distância do item anterior $k-1$ em relação a todos os outros itens da base de dados para definição dos mais próximos.

O critério da distância objetiva a minimização da imprecisão inicial do teste sobre a seleção dos itens. Na Figura 3.4 podemos observar a variação da estimativa de proficiência na aplicação de um TAC com seleção de itens por MIF, atentando, principalmente, para o modo como a estimativa de proficiência do indivíduo sofre uma variação brusca nos primeiros itens do teste até que a região correta de sua proficiência seja encontrada. Como citado anteriormente, a seleção dos itens tende a acontecer em uma região próxima ao valor da estimativa da proficiência, e, selecionando o item mais próximo ao anterior, impede-se que a seleção de itens seja influenciada por essa variação acentuada do início do teste. Porém, o item escolhido não necessariamente deve ser o vizinho mais próximo. Além de manter uma estimativa de proficiência próxima ao item anterior, o item escolhido deve satisfazer os objetivos das metas do erro padrão da melhor forma possível. A estratégia

adotada para que as metas sejam cumpridas da forma mais adequada se dá através de restrições do espaço de busca. Assim, nem sempre o item escolhido será o vizinho mais próximo do item anterior. A seguir, descrevem-se as restrições.

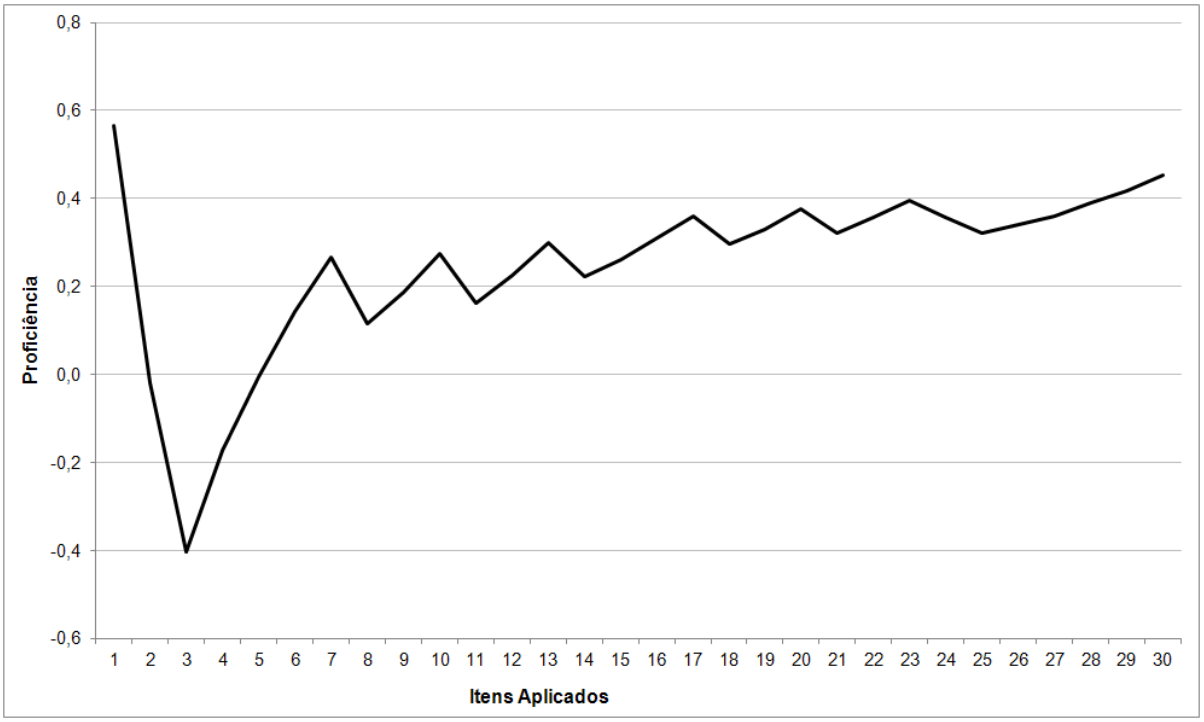

Figura 3.4: Variação da estimativa de proficiência no decorrer da aplicação de um TAC com modelo MIF

A primeira restrição estabelecida na escolha dos itens está relacionada à prioridade que deve ser dada aos itens que cumpram a meta estabelecida e não extrapolem a meta da próxima seleção, ou seja, a média da previsão do erro padrão da proficiência deve estar entre a meta k e a meta $k+1$. Assim, o modelo impede a seleção de itens de assumir um comportamento similar à dos modelos já existentes que buscam a solução ótima. Caso não sejam encontrados itens que atendam a esse critério, mas que consigam atender ao critério extrapolando a meta $k+1$, então estes itens serão aceitos como candidatos e o de menor distância do item anteriormente aplicado será o selecionado. Se, em último caso, não forem encontrados itens que consigam cumprir a meta, será selecionado o item que sua previsão média do erro padrão mais se aproximar da meta, desconsiderando a distância do item anterior.

A segunda restrição estabelecida na busca dos itens é relativa à faixa em que os itens de provável seleção são procurados, restringindo, novamente, o espaço de busca. Estabelece-se que as buscas por itens que atendam as metas sejam feitas, inicialmente, para itens com

parâmetro b dentro de uma faixa com limites inferior e superior definidos pelas estimativas mais recentes da proficiência e do erro padrão. Dessa forma, a busca pelos itens candidatos começam com aqueles cujo parâmetro b seja superior à estimativa de proficiência menos o erro padrão e inferior à estimativa mais o erro padrão. A Figura 3.5 mostra o exemplo da definição deste espaço de busca, dado que a estimativa de proficiência e do erro padrão após a aplicação do item $k-1$ são, respectivamente, 2 e 0,7. O modelo buscará dentro deste espaço de busca os itens que atendem à previsão da meta e selecionará o de menor distância do item $k-1$, considerando seus três parâmetros.

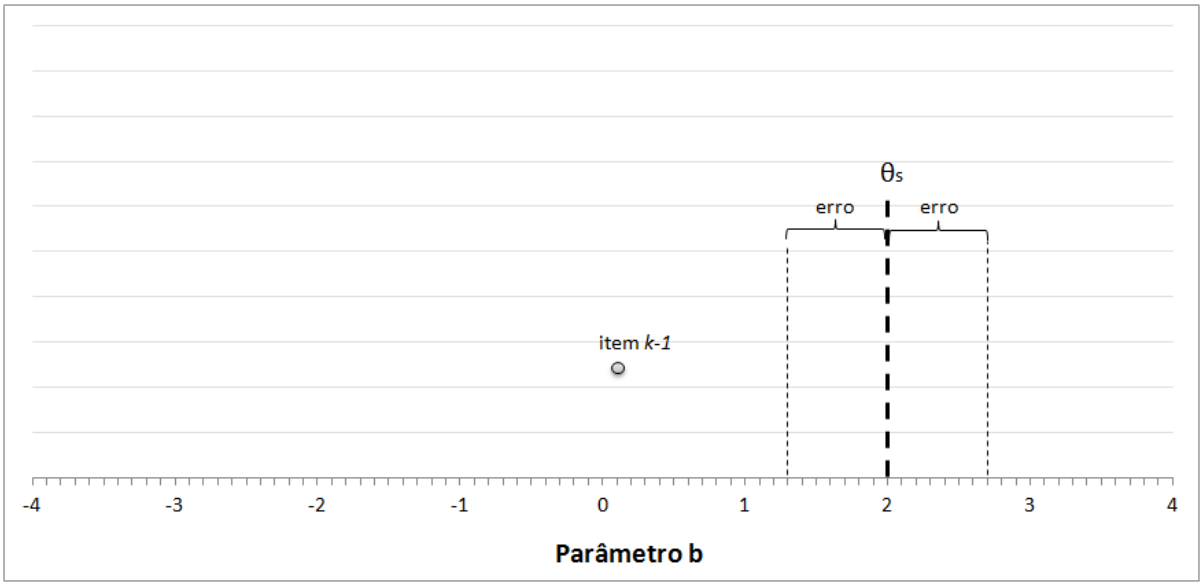

Figura 3.5: Definição do espaço de busca dada a estimativa de proficiência e do erro padrão mais recentes

Caso não sejam encontrados itens dentro dessa faixa de erro padrão ao redor da estimativa de proficiência o modelo aplicará as buscas a todos os itens seguindo os critérios de metas definidos anteriormente. Essa restrição visa proteger a seleção de itens da possibilidade de um item ser selecionado em um ponto muito afastado da estimativa de proficiência. Assim, a seleção de itens utiliza a menor distância entre itens para limitar a ação da imprecisão e a faixa de busca, por sua vez, não deixa o critério da distância agir de forma contrária à indicação da região onde se encontra a proficiência do indivíduo, obtida durante o teste.

Evidentemente, é muito provável que não sejam encontrados itens que atendam a essas duas restrições simultaneamente. À medida que o teste avançar e mais itens forem aplicados, o erro padrão diminuirá e a faixa de busca se tornará cada vez menor, reduzindo

a possibilidade de encontrar itens que atendam as metas. Porém, é importante ressaltar que essa restrição de região de busca é colocada para tentar contornar o problema da imprecisão inicial do teste. Após a aplicação de vários itens esse problema é superado e a aplicação dessa restrição se torna indiferente ao modelo, uma vez que a probabilidade de ter que buscar itens fora desse limite será alta.

Podemos definir a implementação do modelo MEP da seguinte forma:

1. Definir o valor para a meta final do erro padrão da proficiência, o número de itens com que se quer atingir essa meta e a razão da PG que vai determinar o decaimento das metas intermediárias.

2. Definir os valores de proficiência e erro padrão para determinar o espaço de busca do primeiro item a ser aplicado. É prática comum nos TAC's que a seleção do primeiro item seja na região média da escala de dificuldade.

3. Selecionar, aleatoriamente, o primeiro item a ser aplicado. Após a aplicação do mesmo, estimar a nova proficiência e o erro padrão inicial, que será utilizado no cálculo das metas do modelo.

4. Dado o erro padrão inicial, calcular o decaimento total do erro necessário para atingir a meta final. Em seguida, calcular o decaimento mínimo, dada a soma dos termos da PG. Através do valor do decaimento mínimo, calcular a meta do erro padrão da proficiência para cada item com base nos termos da PG associados a eles.

5. Determinar o espaço de busca dos itens candidatos de acordo com a estimativa atual da proficiência e do erro padrão. Buscar, dentro desse espaço, os itens cujas previsões médias do erro padrão atendam à meta para a k-ésima seleção. Selecionar, preferencialmente entre os itens que não extrapolem a meta para a seleção $k+1$, o mais próximo do item aplicado na seleção $k-1$.

6. Caso não sejam encontrados itens candidatos dentro do espaço de busca determinado, expandir essa busca a todos os itens do banco. Utilizar os mesmos critérios em relação ao cumprimento das metas para selecionar o item mais próximo do item da seleção $k-1$.

7. Se, mesmo com a expansão da busca, não forem encontrados itens que cumpram a meta do erro padrão, selecionar o item cuja previsão média do erro padrão mais se aproxime da meta.

8. Aplicar o item selecionado e reestimar a proficiência e o erro padrão. Caso os critérios de parada ainda não tenham sido atendidos, retornar ao processo de busca de itens candidatos para a próxima seleção (Item 5).

Na prática, o modelo MEP tem por objetivo controlar a precisão e o comportamento do teste através, respectivamente, das metas estabelecidas para o erro padrão e da seleção de itens próximos. Assim, espera-se que o teste possa atingir uma boa precisão sem que haja a necessidade de utilizar somente os itens de maior discriminação do banco.

Como já discutido na seção 2.3.2 o modelo MIF é o mais utilizado para seleção de itens em TAC's, sendo considerado o mais eficiente e tendo seus resultados servindo de comparação para os modelos posteriores. Neste trabalho não há a intenção de superar os resultados do MIF, mas tomá-los como exemplo de como deve se comportar um modelo de seleção em relação à precisão da estimativa de proficiência e, ao mesmo tempo, tentar equilibrar esse fator com a exposição dos itens e o tamanho do teste.

4 COMPONENTES PARA SIMULAÇÃO DOS MODELOS

Nesse capítulo são apresentadas as informações sobre os componentes básicos necessários às simulações dos modelos de seleção de itens utilizados neste trabalho. Esses componentes serão apresentados em duas etapas:

1. componentes necessários às simulações de todos os modelos (seções 4.1, 4.2, 4.3 e 4.4);

2. componentes específicos a cada modelo simulado neste trabalho (seções 4.5, 4.6 e 4.7).

Para cada modelo foram simuladas as aplicações de mil testes de língua portuguesa, com tamanho máximo de trinta itens cada.

4.1 COMPOSIÇÃO DO BANCO DE ITENS

Em um teste adaptativo, a composição de um bom banco de itens é um fator primordial para o sucesso do algoritmo de seleção de itens. Segundo Flaugher (2000), três fatores são essenciais na construção desse banco: um número suficiente de itens nas várias regiões da escala de dificuldade, uma revisão pedagógica da qualidade dos itens e um pré-teste com análise psicométrica dos itens.

Reckase (2010) ressalta que as características do teste adaptativo resultam em diferentes requisitos em relação ao banco de itens e, na realidade, não há uma resposta exata para a questão 'Que tamanho deve ter um banco de itens de um TAC?'. Stocking (1994) indica que o tamanho de um banco de itens deve ser, pelo menos, seis vezes maior que o tamanho de um teste no formato tradicional, para atender adequadamente a testes adaptativos que tenham até a metade desse tamanho. Para a aplicação de um TAC com 30 itens seria necessário então um banco com, no mínimo, seis vezes o tamanho de um teste tradicional de 60 itens, totalizando 360 itens no banco.

O banco de itens utilizado nesse estudo conta com quinhentos e oitenta e quatro itens de testes da disciplina de Língua Portuguesa pertencentes ao banco de itens do Centro

de Políticas Públicas e Avaliação da Educação da Universidade Federal de Juiz de Fora (CAEd), com parâmetros gerados através do modelo logístico de três parâmetros. Todos esses itens foram utilizados em avaliações administradas pelo CAEd com seus parâmetros calculados em uma mesma escala e cobrem áreas de conhecimento desde o ensino fundamental até o ensino médio. Por esse motivo, podemos encontrar no banco toda uma variedade de conteúdos que esses itens abrangem, desde as habilidades mais simples até as mais complexas. Essas habilidades foram previamente definidas e os itens desse banco estão separados em 27 dessas classificações, chamadas de descritores.

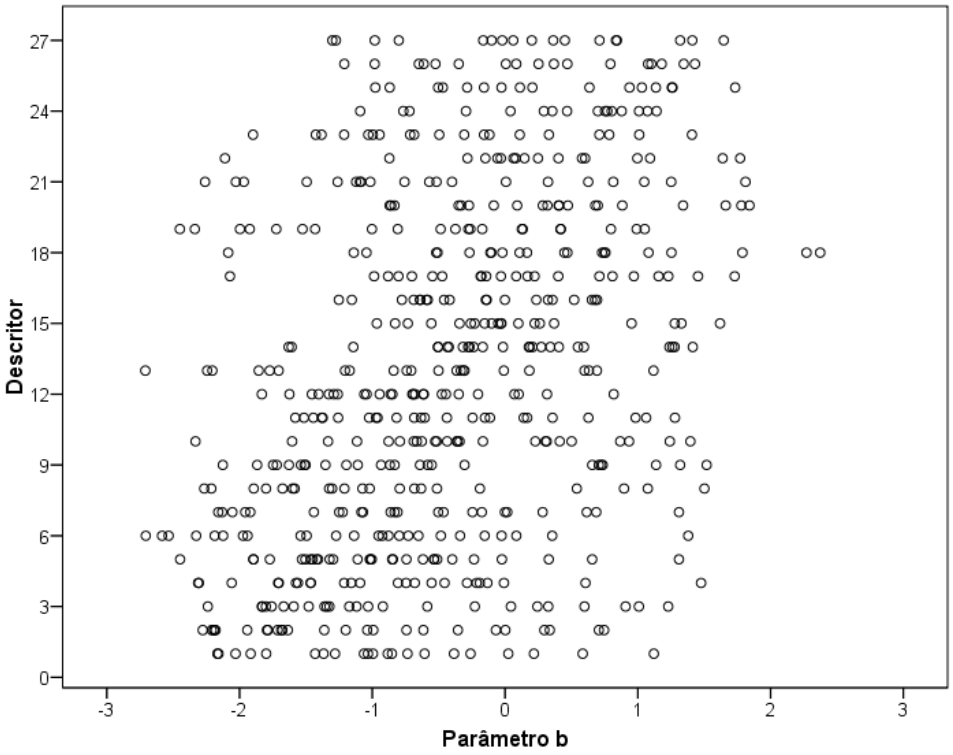

Figura 4.1: Distribuição dos itens de acordo com o descritor e parâmetro de dificuldade

A Tabela 4.1 mostra o número de itens e o percentual de acordo com a classificação por descritores, e a Figura 4.1 exibe a distribuição dos itens de cada descritor pela escala de habilidade. É importante notar que, mesmo que alguns conteúdos da área de conhecimento avaliada sejam considerados mais fáceis ou mais difíceis, há itens de diversos conteúdos por toda a escala de habilidades. O parâmetro de dificuldade do item é independente de seu descritor, ele depende apenas da forma como o item foi construído e como se comporta em um teste. Por exemplo, se um item de uma habilidade considerada mais complexa for construído com um enunciado que deixa sua resposta muito óbvia, esse item terá um

comportamento de item fácil, pois não exigirá muito conhecimento para ser respondido.

Tabela 4.1: Distribuição dos itens pela classificação de descritores

Descritor	Nº de Itens	Percentual
1	21	3,6
2	25	4,3
3	24	4,1
4	24	4,1
5	29	5,0
6	25	4,3
7	23	3,9
8	21	3,6
9	25	4,3
10	25	4,3
11	24	4,1
12	25	4,3
13	22	3,8
14	25	4,3
15	20	3,4
16	20	3,4
17	22	3,8
18	21	3,6
19	21	3,6
20	20	3,4
21	20	3,4
22	17	2,9
23	19	3,3
24	16	2,7
25	16	2,7
26	17	2,9
27	17	2,9
Total	584	100

Podemos verificar, também, pela Figura 4.2, que os parâmetros dos itens utilizados se combinam de diferentes formas, isto é, temos itens de maior ou menor discriminação e acerto casual por toda a escala de dificuldade e, não necessariamente, os itens mais difíceis são os que melhor discriminam os respondentes e agregam mais informação ao teste.

Esses itens, por já terem sido aplicados antes, passaram por análises estatísticas e pedagógicas, permitindo atestar sua qualidade para a produção de medidas. Também, para garantir a capacidade dos itens de produzir uma estimativa de proficiência confiável, foram selecionados apenas aqueles itens que apresentassem um valor mínimo do parâmetro de discriminação (parâmetro a) em torno de 0,5 e um valor máximo para o parâmetro de acerto ao acaso (parâmetro c) de, aproximadamente, 0,2 (FLAUGHER, 2000). As Figuras 4.3, 4.4 e 4.5 mostram as distribuições dos itens segundos os parâmetros a, b e

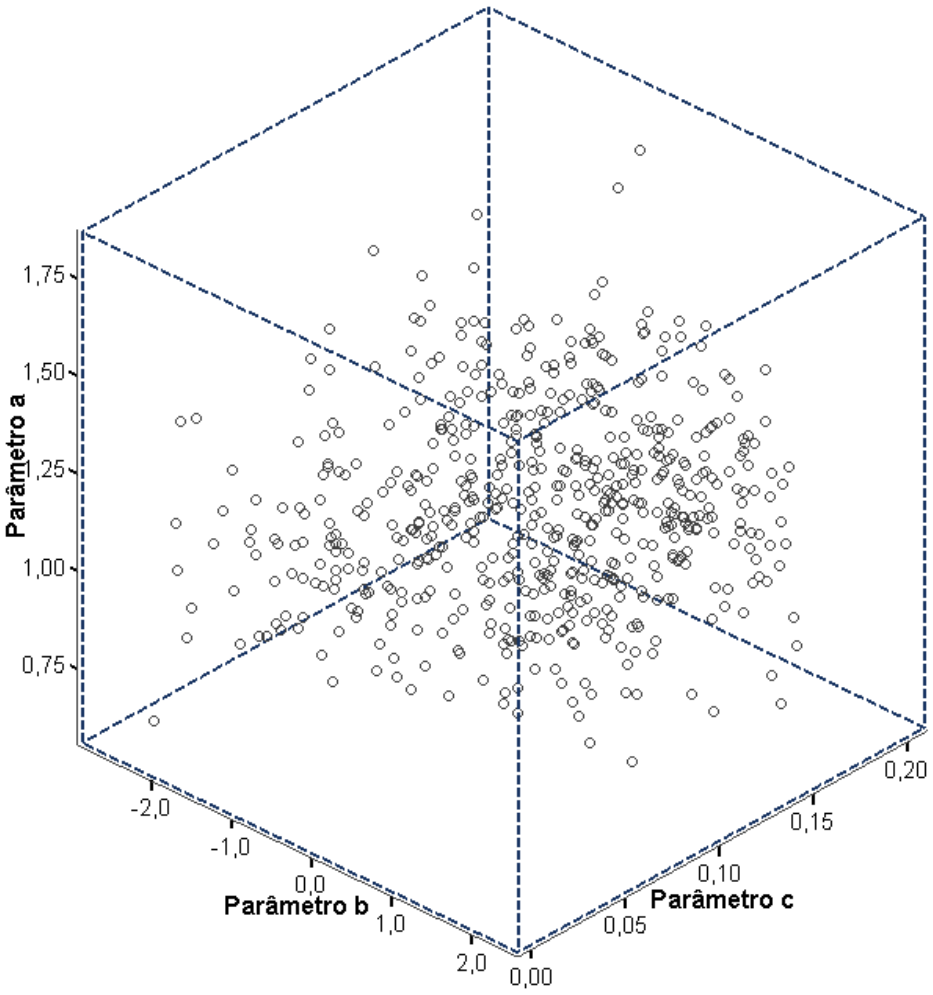

Figura 4.2: Distribuição dos itens do banco de acordo com seus parâmetros

c respectivamente.

4.2 MÉTODO DE ESTIMAÇÃO DE PROFICIÊNCIAS

Nos testes adaptativos, a cada novo item apresentado e respondido, é necessário que a proficiência do indivíduo avaliado seja reestimada. Inicialmente, o estimador de Máxima Verossimilhança (MV) foi o mais empregado nos TAC's devido, principalmente, à facilidade para a implementação do mesmo (MISLEVY, 1986). Porém, esse estimador apresenta algumas limitações, uma vez que nem sempre existe um único máximo da função de verossimilhança para alguns modelos da TRI, inclusive o M3PL. Além disso, esse máximo pode não existir para alguns padrões de resposta, como quando o indivíduo acerta ou erra todos os itens. Uma alternativa às limitações da MV são os métodos bayesianos e, nas simulações apresentadas neste trabalho, foi utilizado o método bayesiano de Média

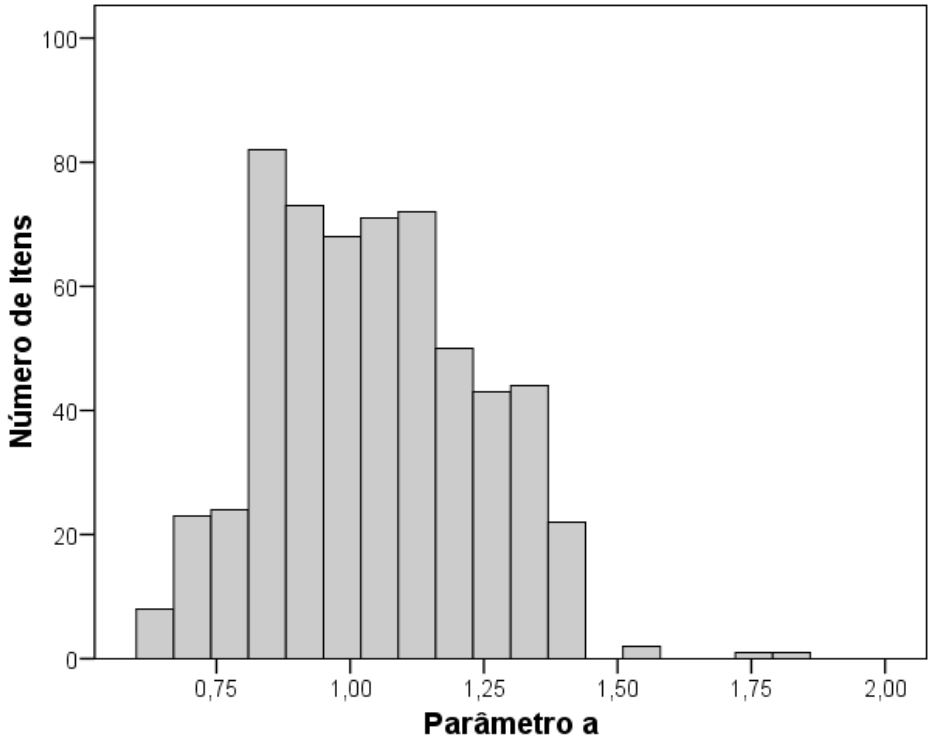

Figura 4.3: Distribuição dos itens do banco pelo parâmetro de discriminação

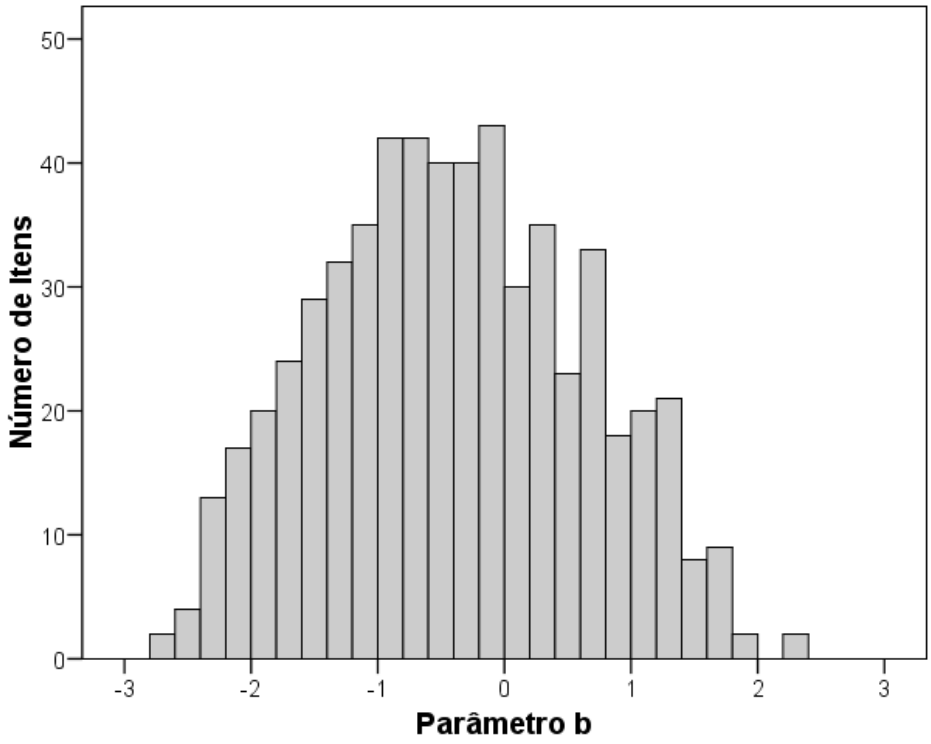

Figura 4.4: Distribuição dos itens do banco pelo parâmetro de dificuldade

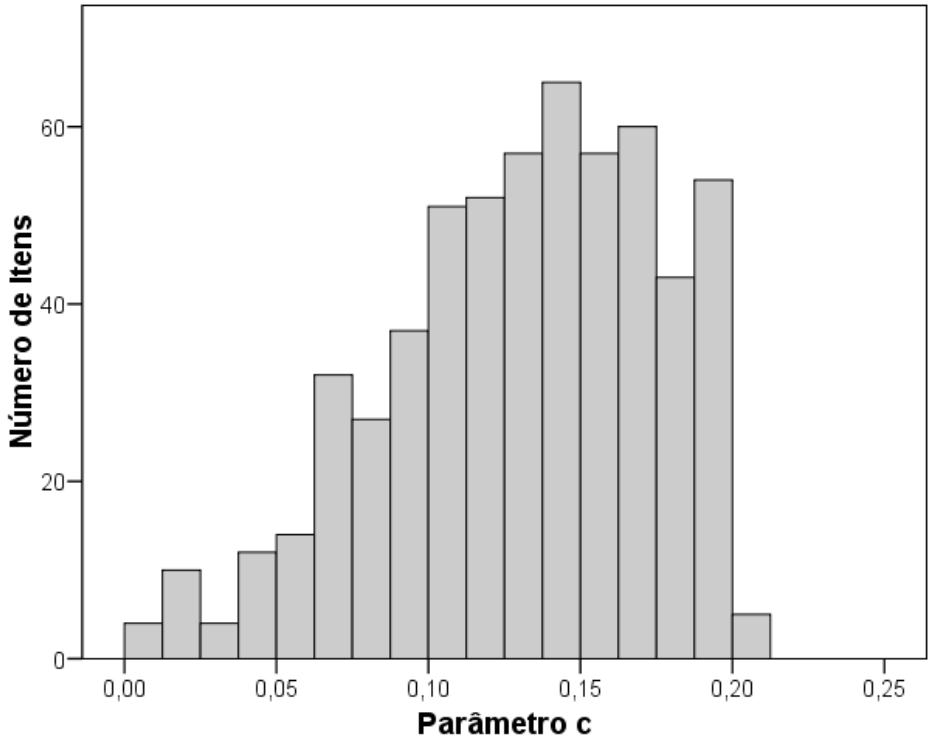

Figura 4.5: Distribuição dos itens do banco pelo parâmetro de acerto casual

a Posteriori (EAP) para a estimação das proficiências dos indivíduos.

Os métodos Bayesianos combinam uma função de verossimilhança com uma distribuição *a priori*, frequentemente modelada por uma distribuição normal (LINDEN; PASHLEY, 2000) . Dada a proficiência θ, a função de verossimilhança associada às respostas aos primeiros k - 1 itens é

$$L(\theta_j; u_1, ..., u_{k-1}) = \prod_{i=1}^{k-1} P_{ji}(\theta_j)^{u_i}[1-P_{ji}(\theta_j)]^{1-u_i} \tag{4.1}$$

onde u_i tem valor 1 ou 0, caso o indivíduo responda o item i corretamente ou não, e $P_{ji}(\theta_j)$ é a probabilidade de resposta correta ao item i, pelo modelo logístico de três parâmetros, dada a habilidade θ do indivíduo.

Combinando a função de verossimilhança com a distribuição *a priori*, temos a distribuição *a posteriori* da habilidade dada por:

$$g(\theta_j|u_1, ..., u_{k-1}) = \frac{L(\theta_j; u_1, ..., u_{k-1})g(\theta_j)}{\int L(\theta_j; u_1, ..., u_{k-1})g(\theta_j)d\theta} \tag{4.2}$$

Pelo método EAP utilizam-se pontos de quadratura da distribuição *a priori* para apro-

ximar as estimativas das habilidades dos indivíduos dessa distribuição. O procedimento de quadratura se baseia em encontrar a soma das áreas de um número finito de retângulos para obtermos aproximadamente a área sob a curva (BAKER; KIM, 2004). Os pontos médios desses retângulos são chamados de pontos de quadratura e neste trabalho são utilizados oitenta pontos de quadratura, em intervalos iguais, de -4 a 4 desvios-padrão em uma distribuição *a priori* modelada por uma normal.

Utilizando a metodologia de pontos de quadratura, podemos redefinir o estimador do EAP como

$$\theta^{EAP}_{j_{u_1,...,u_{k-1}}} = \frac{\int\limits_{\mathbb{R}} \theta_j L(\theta_j|u_1,...,u_{k-1})g(\theta_j)d\theta_j}{\int\limits_{\mathbb{R}} L(\theta_j|u_1,...,u_{k-1})g(\theta_j)d\theta_j} \approx \frac{\sum\limits_{t=1}^{q} X_t L(X_t|u_1,...,u_{k-1})A_t\Delta_t^{-1}}{\sum\limits_{t=1}^{q} L(X_t|u_1,...,u_{k-1})A_t\Delta_t^{-1}} \tag{4.3}$$

onde X_t representa os pontos de quadratura, A_t é o valor da altura da função da distribuição no ponto X_t, que nesse caso é igual a probabilidade da *priori* neste ponto, e Δ_t é o comprimento do intervalo do retângulo correspondente.

O procedimento de estimação da proficiência pelo método de EAP é computacionalmente vantajoso, uma vez que a utilização dos pontos de quadratura permite que este não seja um método iterativo, além de garantir a estimação da proficiência independentemente do padrão de respostas do indivíduo, o que era uma limitação da MV. O erro padrão associado à estimativa de proficiência é obtido pela raiz quadrada da variância da distribuição a posteriori de θ (LINDEN; PASHLEY, 2000), que podemos definir por

$$\begin{aligned} Var(\theta_j|u_1,...,u_{k-1}) &= \frac{\int\limits_{\mathbb{R}} (\theta_j - \theta^{EAP}_{j_{u_1,...,u_{k-1}}})^2 L(\theta_j|u_1,...,u_{k-1})g(\theta_j)d\theta_j}{\int\limits_{\mathbb{R}} L(\theta_j|u_1,...,u_{k-1})g(\theta_j)d\theta_j} \\ &\approx \frac{\sum\limits_{t=1}^{q} (X_t - \theta^{EAP}_{j_{u_1,...,u_{k-1}}})^2 L(X_t|u_1,...,u_{k-1})A_t\Delta_t^{-1}}{\sum\limits_{t=1}^{q} L(X_t|u_1,...,u_{k-1})A_t\Delta_t^{-1}} \end{aligned} \tag{4.4}$$

4.3 SIMULAÇÃO DAS RESPOSTAS

Pelos modelos da TRI, quando um item é selecionado para a aplicação ele está associado a uma probabilidade de acerto. Portanto, quando simulamos as respostas de um indivíduo, não podemos simplesmente admitir um valor automático de resposta correta ou incorreta

sem levar em consideração as probabilidades associadas aos parâmetros do item.

Assim, o algoritmo de simulação das respostas parte do cálculo da probabilidade de acerto ao item selecionado, dada a estimativa atual da proficiência, para gerar um valor aleatório de acordo com a distribuição uniforme de intervalo [0,100] que satisfaça ou não a essa probabilidade, computando então a resposta como certa ou errada.

Por exemplo, dada uma estimativa de proficiência, a probabilidade de acerto em um novo item selecionado é de 68%. O simulador gera, então, um número real aleatório entre 0 e 100 que será comparado com essa probabilidade de acerto. Caso esse número seja menor ou igual à probabilidade será computado um acerto no item e, caso seja maior, será computado um erro.

Esse procedimento é simples e eficiente para a simulação, uma vez que a probabilidade teórica de acertar o item é a mesma de que seja gerado um número aleatório menor que essa probabilidade. Mesmo com a reestimação da proficiência após a aplicação do novo item, as respostas que já foram simuladas não são alteradas, pois são fruto da probabilidade referente à proficiência estimada no momento da seleção do item.

4.4 CRITÉRIOS DE PARADA E PRECISÃO DO TESTE

Os critérios de parada nos TAC's foram discutidos na seção 2.2. Nas simulações feitas para este trabalho foi definido um critério de 30 itens como tamanho máximo do teste, uma vez que não há uma regra fixa estabelecida para o número de itens no teste. Esse valor foi tomado a partir de diversos experimentos já apresentados anteriormente, como os de Chang e Ying (1999), Chang et al. (2001), Linden (2003), Barrada et al. (2010), Eggen e Straetmans (2000), entre outros, que simulam testes com tamanho entre 25 e 40 itens.

Da mesma forma, esses e outros trabalhos também apontam análises da precisão da estimativa de proficiência considerando-se como objetivo a estimativa do erro padrão final em diferentes valores, variando desde 0,2 a 0,4. Neste trabalho as análises são feitas utilizando-se três valores para o erro padrão como parâmetro de comparação: 0,3, 0,25 e 0,2.

4.5 ESTRATOS DE DIFICULDADE E DISCRIMINAÇÃO - SELEÇÃO POR ESTRATIFICAÇÃO

Para a simulação do modelo de seleção por estratificação foram criados dois tipos de estratos conforme a proposta de Chang et al. (2001), discutida na seção 2.3.1. Inicialmente os itens foram ordenados e divididos em três grupos, contendo aproximadamente o mesmo número de itens, de acordo com o parâmetro de discriminação (parâmetro a). Depois foram ordenados e divididos em dez grupos, também com quantidade de itens aproximadamente igual, de acordo com o parâmetro de dificuldade (parâmetro b). Dessa forma foram criados trinta subgrupos de onde são selecionados os itens de acordo com a estimativa de proficiência e com a etapa do teste. Como discutido na seção 4.1, independentemente do parâmetro de dificuldade temos itens com maior ou menor grau de discriminação, portanto, mesmo que em alguns estratos tenhamos valores próximos, não é possível garantir que os subgrupos sejam formados com o mesmo número de itens.

A Tabela 4.2 mostra o número de itens por estrato do parâmetro a em cada estrato do parâmetro b.

Tabela 4.2: Distribuição dos itens nos subgrupos formados pelos estratos

Estratos b	Estratos a			Total
	1º	2º	3º	
1º	19	21	18	58
2º	16	20	22	58
3º	22	16	20	58
4º	12	18	28	58
5º	22	21	15	58
6º	22	21	15	58
7º	23	19	16	58
8º	21	19	18	58
9º	21	19	18	58
10º	17	21	24	62
Total	195	195	194	584

4.6 CONTROLE DE EXPOSIÇÃO DE ITENS - SELEÇÃO POR MÁXIMA INFORMAÇÃO

Como discutido na seção 2.3.2, pelo modelo de seleção de itens por MIF, com a proficiência provisória inicial sendo a mesma para qualquer indivíduo, o item de maior informação seria selecionado, o segundo item seria selecionado entre duas opções dependendo do desempenho no item anterior, o mesmo acontecendo com todos os itens subsequentes. Assim, a sequência de itens se torna previsível e os itens que agregam maiores valores de informação ao teste são frequentemente selecionados. Esse comportamento é chamado de superexposição dos itens e resulta no risco de que, com o passar do tempo, um grande número de pessoas avaliadas tenha conhecimento prévio de alguns itens que possam aparecer na aplicação do teste (CHANG; ANSLEY, 2003).

Para tentar contornar esse comportamento há diversas propostas de métodos de controle da exposição dos itens. Neste trabalho foi utilizado o método Sympson-Hetter (SH) (HETTER; SYMPSON, 1997) que propõe a aplicação de um parâmetro de controle de exposição E_i para cada item do banco. Na prática, o método consiste na criação de uma taxa de probabilidade do item ser aplicado uma vez que seja selecionado para o teste. Itens que produzem maior informação, que frequentemente seriam aplicados, possuem um valor baixo para esse parâmetro de exposição, evitando a superexposição. Por outro lado, itens de menor informação possuem um parâmetro de exposição alto, permitindo que esses itens tenham uma probabilidade alta de que sejam aplicados quando forem selecionados. Dessa forma, um item selecionado só será aplicado após o resultado positivo em um teste de probabilidade de acordo com sua taxa de exposição.

O parâmetro de controle de exibição para cada item é obtido através de um procedimento iterativo. A simulação dos testes considera o comportamento dos itens dada uma amostra de casos criada para representar uma distribuição de proficiência de forma similar a uma população real. Os passos desse procedimento podem ser descritos da seguinte forma:

1. Definir a taxa máxima esperada t de exposição de itens para o teste. Hetter e Sympson (1997) citam um valor de $t = 1/3$ em seu experimento e Linden e Glas (2000) indicam que esse valor não deve ser menor que n/I, sendo n o tamanho do teste aplicado e I o total de itens no banco, sendo comum utilizar um valor entre

0,20 e 0,30 para t.

2. Inicializar o parâmetro E_i de controle de exposição de todos os itens do banco com valor 1. Assim, inicialmente, todos os itens que forem selecionados serão aplicados.

3. Simular um teste adaptativo para todos os indivíduos da amostra criada selecionando os itens pelo modelo MIF dada a proficiência $\hat{\theta}$ desse indivíduo. A cada seleção de um item, gerar um número aleatório x de acordo com uma distribuição uniforme de intervalo [0,1] e, caso esse valor x seja menor ou igual ao parâmetro E_i, aplicar esse item no teste. Independentemente de um item ser aplicado ou não, caso ele seja selecionado uma vez, não deverá ser selecionado novamente durante a aplicação do teste para o mesmo indivíduo, ou seja, seleção sem reposição.

4. Registrar o número de vezes em que cada item foi selecionado (NS) e o número de vezes em que foi aplicado (NA) em todos os testes simulados. Ao fim da simulação de todos os testes calcular a probabilidade de um item ser selecionado, *P(S)*, e a probabilidade de ser aplicado, *P(A)*, dado o número do total de pessoas examinadas (NE):

$$\begin{aligned} P(S) &= NS/NE \\ P(A) &= NA/NE \end{aligned} \tag{4.5}$$

5. Calcular o novo valor para E_i, de acordo com o valor de t definido anteriormente e o com o valor de *P(S)*:

$$\begin{aligned} E_i &= t/P(S), \text{ se } P(S) > t \\ E_i &= 1, \text{ se } P(S) \leq t \end{aligned} \tag{4.6}$$

6. Para testes de tamanho n, se não houver, pelo menos, n itens com o novo E_i igual a 1, transforme os n itens de maior E_i para esse valor, de forma a garantir a aplicação de testes com esse tamanho a todos os avaliados antes de esgotar o banco de itens.

7. Após o cálculo dos novos valores de E_i, retomar o procedimento a partir do passo 3 até que o maior valor obtido para $P(A)$ entre todos os itens seja um pouco superior a t. Quando esse patamar for atingido por algumas simulações consecutivas, o valor de E_i obtido é o valor final para a taxa de exposição de cada item em futuros testes.

Na prática, o processo de obtenção dos parâmetros de exposição de itens pelo método SH consome tempo, sendo comum que sejam feitas de 100 a 150 rodadas de simulação

de testes antes que se obtenha os parâmetros finais a serem usados em um TAC. Além disso, caso haja alguma mudança no banco de itens, como a adição ou retirada de itens, o procedimento SH deve ser todo refeito (LINDEN, 2003).

4.7 ESTIMATIVA DAS METAS DO ERRO PADRÃO - SELEÇÃO POR METAS DE ERRO

Para o modelo proposto neste trabalho, a estimativa das metas de erro padrão da proficiência utilizadas nas seleções dos itens é definida a partir de quatro parâmetros:

1. meta para o erro padrão final;
2. número de itens para cálculo das metas;
3. razão da PG;
4. amplitude da faixa de seleção do primeiro item do teste.

Os testes foram simulados com o intuito de alcançar o valor de 0,2 para o erro padrão final, e, a partir desses resultados, são feitas análises para os valores superiores a esse citados anteriormente (seção 4.4). Essa meta foi definida prevendo-se a utilização de 26 itens para que fosse alcançada. Dado que o teste tem o limite de 30 itens, há ainda quatro itens restantes para possível aplicação no teste caso o erro padrão ainda não tenha atingido a meta. Nos casos em que haja necessidade de aplicação desses itens além do previsto, será mantida a meta final para o erro padrão. Essa quantidade de itens baseou-se nos resultados dos testes aplicados com o modelo MIF onde foram necessários, pelo menos, 26 itens nos casos em que a estimativa do erro padrão conseguiu atingir o valor de 0,2.

Assim como para a meta final do erro padrão, a amplitude da faixa inicial de seleção aleatória do primeiro item também teve como base um erro padrão de 0,2. Dessa forma, o modelo escolheu aleatoriamente um item com parâmetro b entre -0,2 e 0,2 para aplicação inicial em cada teste, dentre um universo de 73 itens do banco que atendem a esse critério. Por último, o valor de 1,2 para a razão da PG também foi tomado com base no comportamento do erro padrão nas simulações pelo modelo MIF nos 30 itens aplicados por teste. A Figura 4.6 mostra o comportamento do erro padrão nas simulações conduzidas pelo modelo MIF e um exemplo da previsão das metas utilizando os parâmetros definidos

acima, considerando o mesmo erro padrão inicial do MIF. Esse valor para a razão da PG determina um comportamento inicial das metas similar ao exibido pelo modelo MIF.

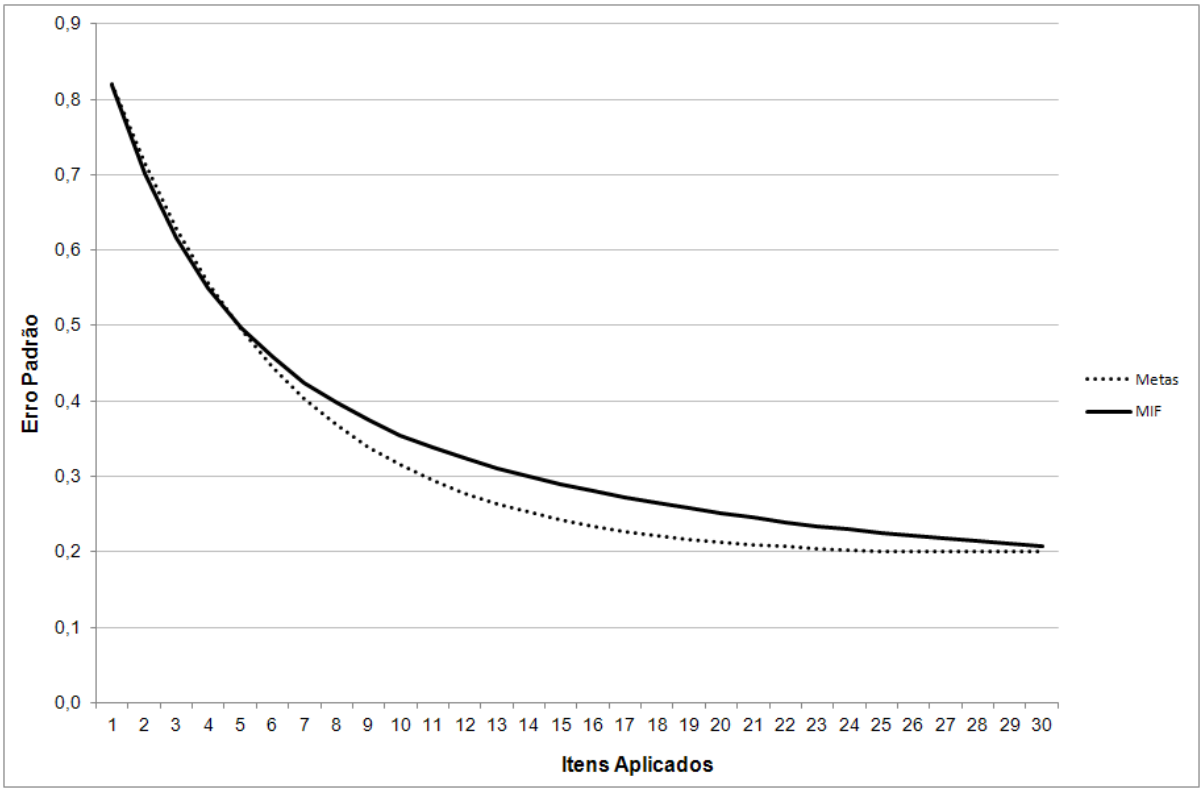

Figura 4.6: Comportamento do erro padrão pelo MIF e exemplo de previsão de metas

Podemos observar que a variação do erro padrão na parte final dos testes pelo MIF é pequena, com redução média dos valores em torno de 0,004 na aplicação dos últimos dez itens selecionados. Admitindo-se que o cumprimento de metas torna-se mais difícil no fim do teste devido a esse comportamento, garantir as metas na primeira parte do teste, mais exatamente no primeiro terço do teste, se torna essencial para a redução do valor do erro padrão. O primeiro experimento para o modelo proposto por Chang e Ying (1996) se limitava à seleção de apenas 14 itens pois avaliava o comportamento do teste apenas nas seleções iniciais, partindo do princípio que esse estágio do teste possa definir a precisão da estimativa do erro padrão da proficiência ao final. Um valor razoável para a razão da PG também garante que as metas definidas para o segundo terço do teste sejam suficientemente exigentes de forma a poder compensar possíveis metas não cumpridas na primeira parte do teste e colaborar para uma melhor precisão final.

A seguir, as simulações numéricas utilizando os dois modelos de referência e o modelo baseado em metas proposto nesse trabalho são apresentadas. Uma análise criteriosa dos

resultados é desenvolvida visando avaliar as características e potencial de cada uma das estratégias.

5 SIMULAÇÕES NUMÉRICAS E ANÁLISE DE RESULTADOS

Neste capítulo serão apresentados e analisados os resultados das simulações dos TAC's buscando estabelecer comparações entre os modelos tradicionais de estratificação e máxima informação e o de metas do erro padrão, de forma a avaliar este modelo proposto não só em termos de desempenho, mas, também, em como ele reage à dinâmica dos TAC's. A primeira parte da análise é focada na estimativa do erro padrão, avaliando o nível de sucesso obtido pelos testes dados os três níveis de precisão, apresentados seção 4.4, estabelecidos como parâmetros para comparação, o comportamento do erro padrão ao longo do teste e em relação às estimativas de proficiência obtidas.

A segunda parte da análise é referente ao nível de exposição dos itens nos testes, observando a variedade de itens utilizados nos testes e a taxa de exposição a que foram submetidos. É feita também uma análise específica para o modelo de metas proposto neste trabalho, avaliando o relacionamento entre as metas e os resultados obtidos, o cumprimento das metas e as possibilidades de recuperação do modelo em caso de metas não cumpridas.

Antes da avaliação dos resultados obtidos, é importante ressaltar que as simulações para o modelo de metas do erro padrão (MEP), aqui apresentado, foram feitas visando alcançar a menor meta esperada para o erro padrão (0,2). Assim, o comportamento do modelo em relação aos valores maiores usados para avaliação (0,3 e 0,25) pode ser afetado por não ter sido simulado com uma meta específica para esses valores. Também devemos, novamente, salientar que, para o desenvolvimento do modelo MEP, o desempenho do modelo MIF foi uma referência, atestando a importância deste modelo em seleção de itens para TAC's.

5.1 ESTIMATIVAS DO ERRO PADRÃO DA PROFICIÊNCIA

As simulações realizadas para esse trabalho apontaram uma diferença significativa na precisão final dos testes de cada modelo, como podemos observar pelo *Mean Rank* dos modelos e pelo *p-value* resultante do teste de Kruskal-Wallis exposto na Tabela 5.1. Esse

é um teste não paramétrico, utilizado para definir se as distribuições de mais de dois grupos de dados podem ser consideradas iguais, sendo utilizado em substituição à análise de variância quando as distribuições a serem comparadas não seguem uma distribuição normal, como acontece com as estimativas do erro padrão aqui analisadas. Os resultados apresentados na Tabela 5.2 apresentam os valores médios, mínimos e máximos obtidos para o erro padrão, bem como o desvio padrão. Os modelos de MIF e de MEP atingiram valores mais precisos que o de estratificação, tanto na média como nos menores valores obtidos, porém com uma variância maior, principalmente do modelo de metas.

Tabela 5.1: Teste de Kruskal-Wallis do erro padrão por modelo de seleção de itens

Classificação

	Modelo	N	Média
erro padrão	Estratificação	1000	2232,023
	Máx. Informação	1000	840,874
	Metas	1000	1428,603
	Total	3000	

Estatísticas (a,b)

	erro padrão
Qui-quadrado	1300,102
G. Liberdade	2
p-valor	0,000

a. Teste de Kruskal Wallis

b. Variável de grupo: Modelo

Tabela 5.2: Erro padrão por modelo obtido ao fim das simulações

Modelo	Erro Padrão			
	Mínimo	Máximo	Média	DP
Estratificação	0,219	0,391	0,246	0,016
Máx. Informação	0,186	0,393	0,207	0,023
Metas	0,189	0,432	0,232	0,051

Analisando os dados expostos na Tabela 5.3 observamos novamente o comportamento diferenciado dos modelos. Para o parâmetro de erro padrão mais alto (0,3) o modelo de estratificação tem um percentual de sucesso maior, porém, utilizando em torno de 8 itens a mais que os outros modelos para conseguir atingir essa precisão. Se analisarmos a velocidade com que o modelo MIF e o de metas atingiram essa precisão, em alguns casos com 11 itens apenas, vemos que a possibilidade de redução no tamanho de um TAC depende apenas do objetivo da avaliação e, consequentemente, do limite aceitável para a precisão.

Tabela 5.3: Índice de precisão atingido e itens utilizados

Precisão	Modelo	% Testes	Itens			
			Mínimo	Máximo	Média	DP
0,30	Estratificação	98,7	21	30	23,3	1,5
	Máx. Informação	98,3	11	30	14,3	2,6
	Metas	90,8	11	30	15,8	3,3
0,25	Estratificação	76,9	26	30	28,8	1,0
	Máx. Informação	95,8	16	30	19,8	2,4
	Metas	79,8	17	30	21,4	3,1
0,20	Estratificação	0				
	Máx. Informação	43,6	26	30	28,7	1,0
	Metas	22,4	26	30	29,2	0,9

Na medida em que a precisão se torna mais rigorosa o MIF se destaca dos outros modelos pelo percentual de sucesso com que atinge o valor proposto e, mesmo não conseguindo um valor alto de sucesso no patamar mais rígido do erro padrão (0,2), ainda tem um desempenho bastante superior aos outros modelos. Mesmo que o percentual de sucesso em atingir a precisão proposta seja razoavelmente mais baixo, um ponto a se destacar em relação ao modelo de metas do erro padrão é o seu comportamento parecido com o do MIF entre os casos que cumprem esse objetivo.

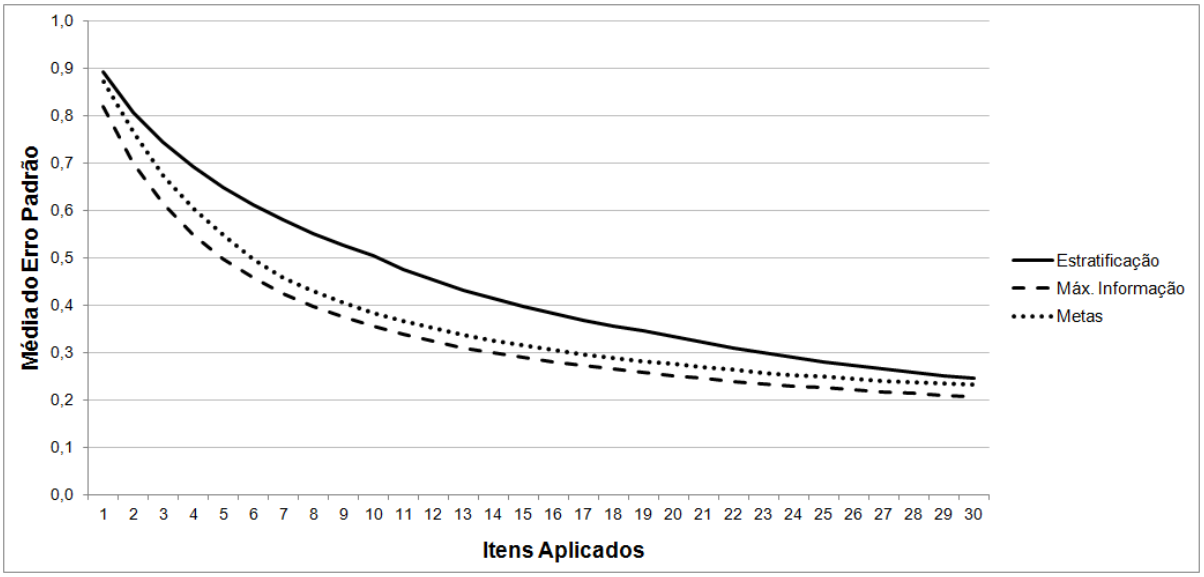

Figura 5.1: Média do erro padrão a cada aplicação de item

A Figura 5.1 apresenta a comparação, para cada aplicação de item nos testes, da média do erro padrão por modelo permitindo confirmar, dessa vez considerando-se todos

os testes, o comportamento parecido do modelo de metas em relação ao MIF. Sendo assim, é necessário buscar uma explicação para o fato de que, mesmo com um comportamento parecido, o modelo de metas não consegue estabelecer percentuais de sucesso em relação aos parâmetros de precisão parecido com os do MIF.

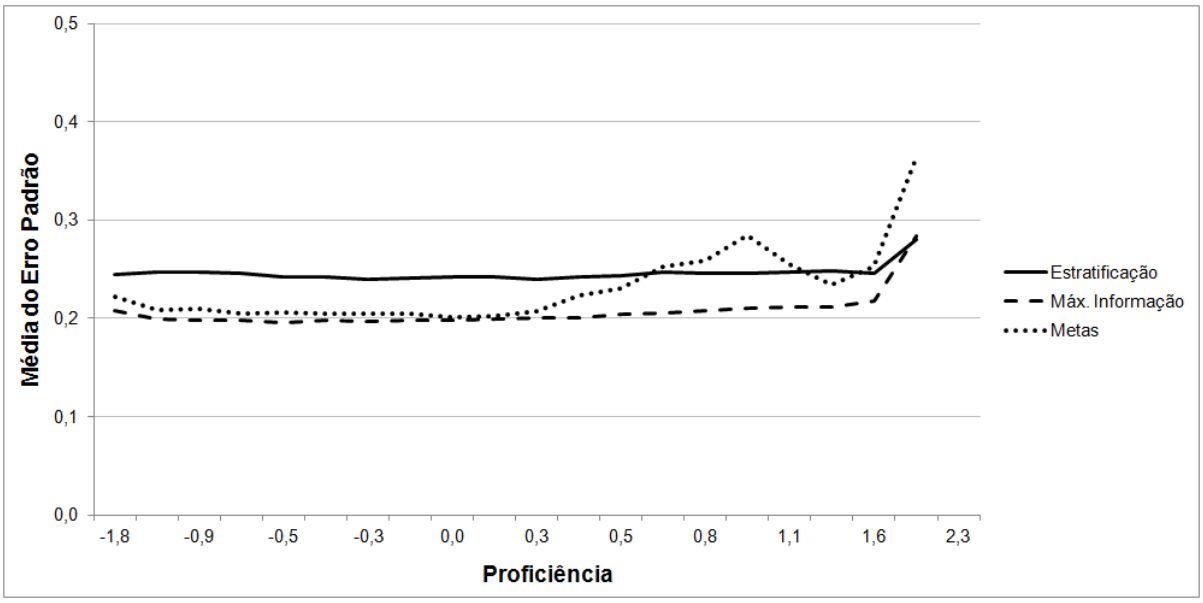

Figura 5.2: Média do erro padrão pelas estimativas de proficiência obtidas no teste

Na Figura 5.2 são comparadas as médias do erro padrão em relação às estimativas de proficiência obtidas nas simulações e, acompanhando essas médias, podemos observar que o erro padrão é afetado no extremo superior da escala em todos os modelos. A partir do que foi discutido na seção 2.3 sobre a relação entre presença de itens por toda a escala e precisão nos testes, torna-se necessário avaliar mais detalhadamente a distribuição dos itens por seu parâmetro b.

Tabela 5.4: Proficiência média dos casos com erro padrão acima de 0,3

Modelo	Proficiência Média	Número de Testes
Estratificação	-2,88	1
	-1,46	1
	2,70	11
Máx. Informação	2,66	17
Metas	-3,01	2
	-1,66	1
	1,13	43
	2,43	46

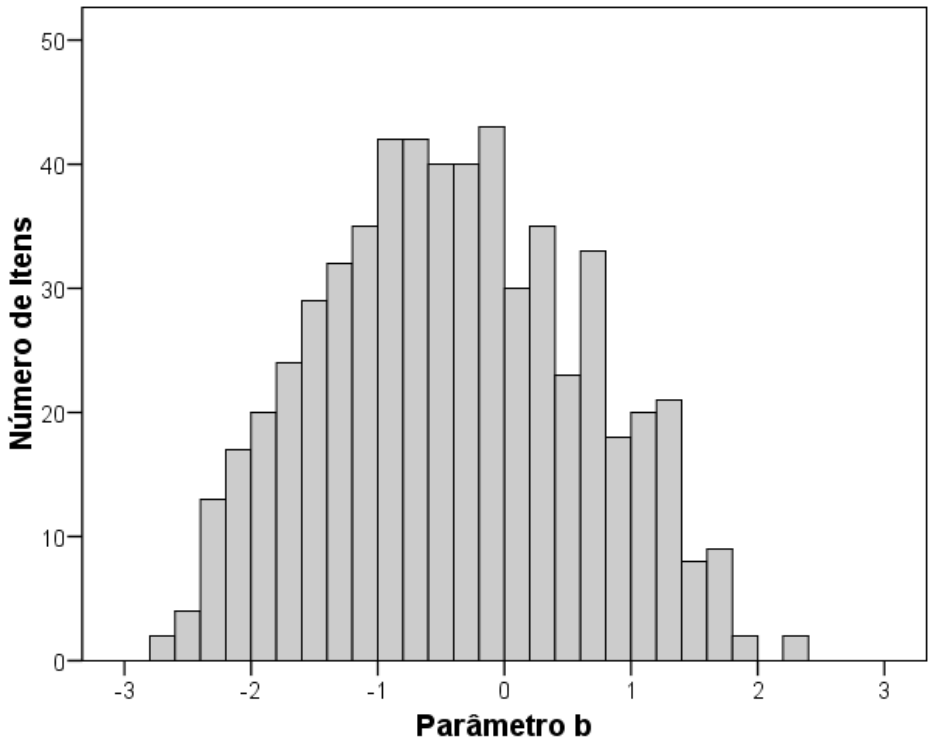

Figura 5.3: Distribuição dos itens do banco pelo parâmetro de dificuldade

Como podemos observar pela Figura 5.3, a distribuição é assimétrica, com cerca de 65% dos itens com parâmetro b abaixo de zero, e, além da pequena presença de itens no extremo superior da escala, principalmente se comparado ao extremo inferior, há uma redução brusca de itens na região de dificuldade igual a 1. Essa redução afeta diretamente o modelo de metas do erro padrão, fazendo com que o mesmo apresente uma irregularidade na média do erro padrão nessa região da escala, demonstrando uma maior sensibilidade do modelo às imperfeições do banco de itens. Se avaliarmos somente os casos que falharam em atingir o parâmetro mais alto de precisão, isto é, testes com erro padrão acima de 0,3, essa irregularidade fica mais evidente.

Na Tabela 5.4 temos uma análise por região da escala com a média da estimativa de proficiência desses casos e o número de vezes em que essas falhas ocorreram para os três modelos de seleção de itens. Essas regiões foram definidas em seis intervalos dentro da escala: [-4,-2], (-2,-1], (-1,0], (0,1], (1,2], (2,4]. Exceto por poucos casos na região inferior da escala, a falta de itens no extremo superior foi um fator constante de interferência para todos. No caso do modelo MIF, só houve falhas no extremo superior da escala. Porém, no caso do modelo de metas do erro padrão, a irregularidade do banco na região da escala

com valor em torno de 1 teve um efeito negativo na estimativa de precisão quase tão grande quanto o do extremo superior.

5.2 EXPOSIÇÃO DOS ITENS

A análise da exposição de itens está relacionada à análise do comportamento do teste como um todo. Quando encontramos uma disparidade entre resultados de modelos diferentes, é necessário avaliar quais os fatores característicos dos modelos de seleção e quais os característicos dos testes, independente de modelo, influenciam nesses resultados. Podemos tomar como ponto de partida os resultados gerais apresentados na Tabela 5.5, relativos às simulações realizadas, para estabelecer uma ligação com as características básicas dos modelos.

Tabela 5.5: Itens diferentes selecionados e média de seleção por item

Modelo	Itens Selecionados	Média por Item
Estratificação	577	52,0
Máx. Informação	235	127,7
Metas de Erro	314	95,5

O resultado apresentado pela seleção por estratos mostra a diferença conseguida por um modelo que tem como um de seus principais objetivos o aumento, mesmo que de forma controlada, na aleatoriedade da seleção. Esse modelo usou quase todos os itens do banco e apresentou uma média de 52 seleções por item, o que significa que um item era selecionado, em média, a cada 19 testes. Conforme esperado, o modelo MIF apresenta o menor valor para o número de itens selecionados, 235 em um universo de 584 itens do banco, e consequentemente, tem a maior média de seleções por item. Mesmo com a implementação de um método de controle de exposição, o determinismo ligado ao modelo MIF ainda é preponderante para a seleção de itens. No caso do modelo de metas do erro padrão, os resultados se aproximam mais do determinismo do modelo MIF do que da aleatoriedade da estratificação. Apesar de ter um número de itens selecionados 33% maior que o MIF, o modelo utilizou apenas 54% dos itens do banco e teve uma média aproximada de um item aplicado a cada 10 testes.

A concepção da forma como os modelos trabalham na seleção de itens influenciam diretamente na exposição dos mesmos, porém, parte desse comportamento pode ser atribuído

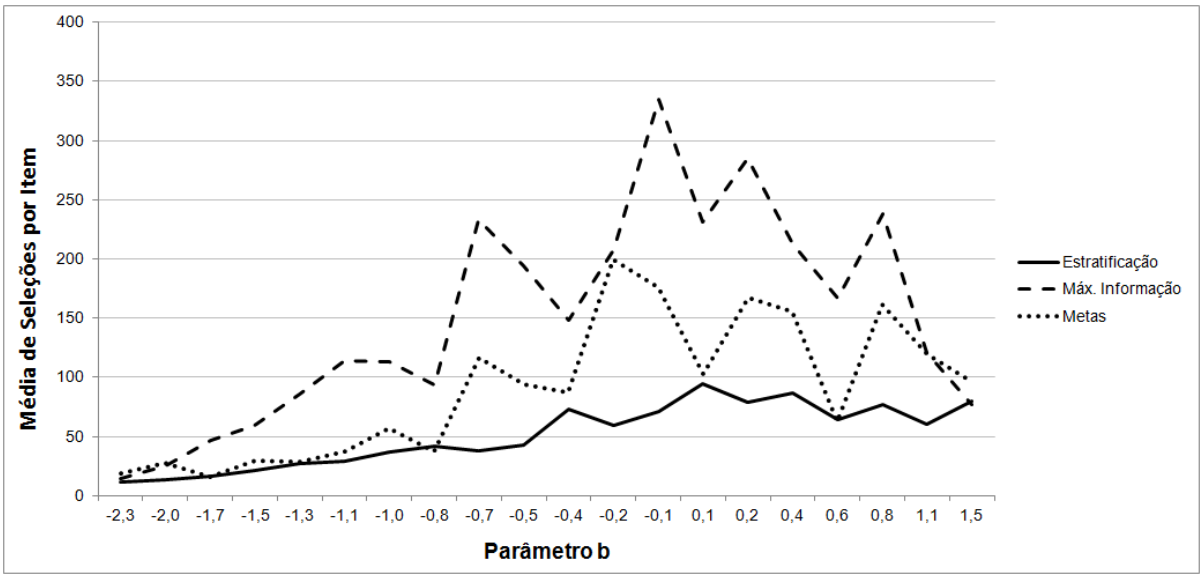

Figura 5.4: Média de seleção de itens pela escala de dificuldade

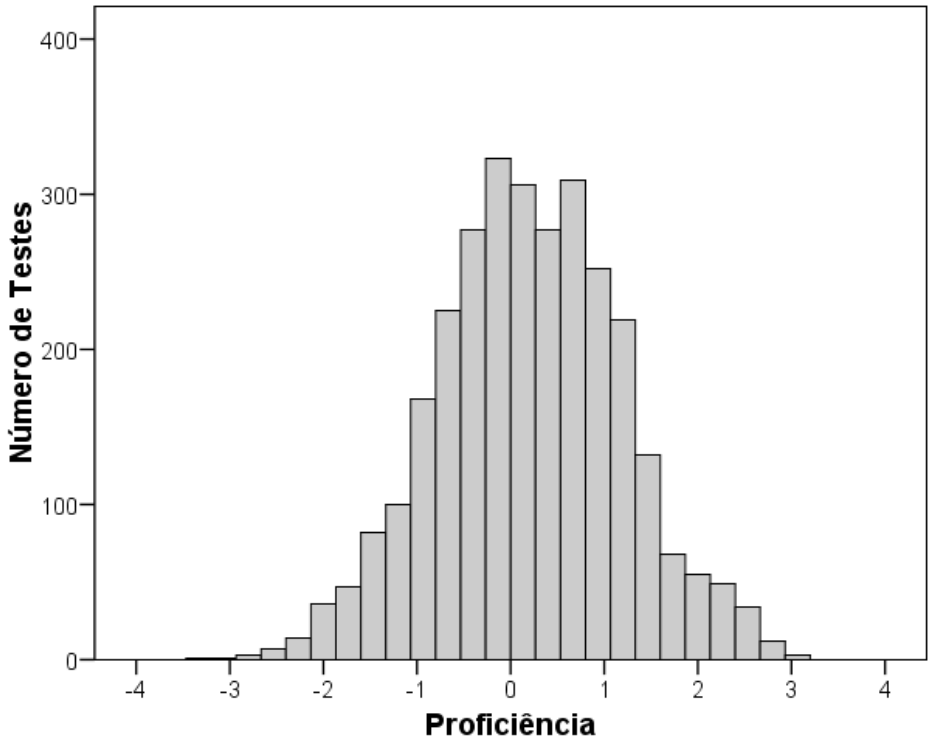

Figura 5.5: Distribuição das proficiências das simulações dos três modelos

às características dos testes como um todo. A Figura 5.4 mostra a média de seleções dos itens de acordo com a escala de dificuldade e, confirmando as proporções dos resultados gerais, o modelo de metas do erro padrão se comporta de madeira parecida com o MIF. Por essas médias podemos observar o aumento no índice de seleções na região próxima à

média da escala (zero). Esse fator é explicado pela própria natureza das estimativas de proficiência resultante dos testes, uma vez que a maioria dos indivíduos avaliados tende a se localizar próxima à média da escala. Essa tendência se confirma pela Figura 5.5, que mostra a distribuição de todas as estimativas de proficiência obtidas pelas simulações dos três modelos.

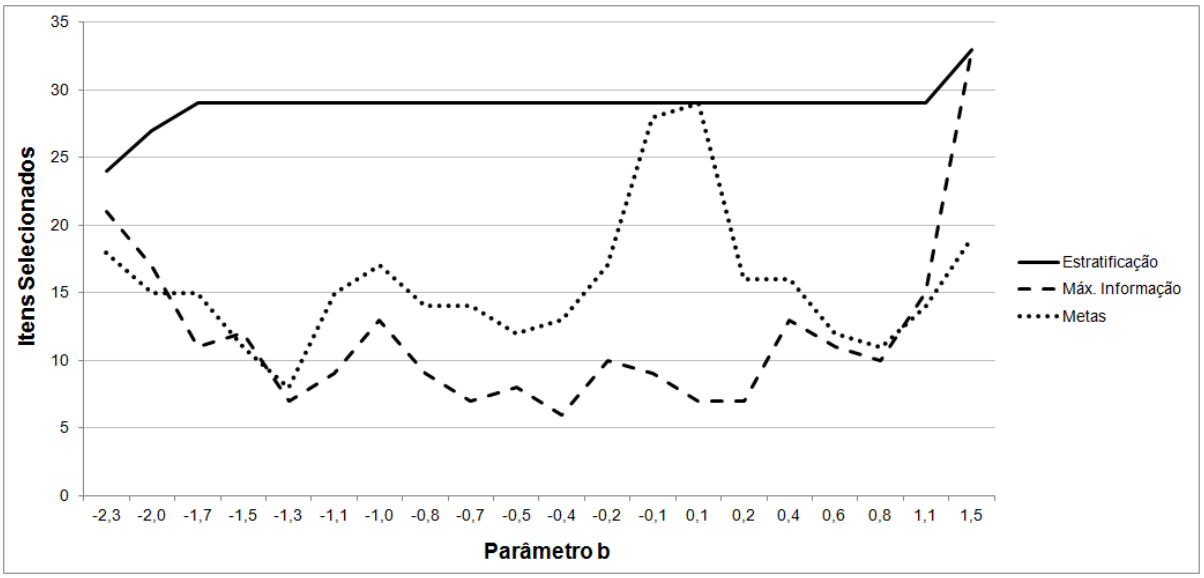

Figura 5.6: Número de itens selecionados pela escala de dificuldade

Um fato que merece atenção é o comportamento do modelo de metas do erro padrão na região média da escala onde, assim como os outros modelos, ele tem a maior média de exposição dos itens. Porém, ao contrário das outras áreas da escala, nesse ponto o modelo de metas teve a maior variedade de itens selecionados, apresentando um comportamento próximo ao modelo de estratificação e não ao de MIF. Na Figura 5.6 é apresentado o número de itens diferentes selecionados por região da escala de dificuldade e nela observamos esse comportamento diferenciado do modelo de metas. Se considerarmos a área em que ocorre essa variação diferenciada do número de itens selecionados como sendo entre -0,3 e 0,3, encontramos, pelas estimativas de proficiência finais, cerca de 25% dos indivíduos avaliados pelo modelo. Assim, uma vez que houve maior variação dos itens na região de maior concentração de indivíduos e onde, evidentemente, ocorreu a maior parte das seleções, esse fator pode ter causado uma pequena compensação considerando que a média de exposição dos itens não foi muito mais baixa que a do modelo MIF.

Um outro fator que muitas vezes é deixado de lado quando se avalia a exposição de itens é o equilíbrio dos conteúdos exibidos nos testes. Evidentemente, essa característica dos

testes depende muito da constituição do banco de itens, discutida na seção 4.1. No banco utilizado neste trabalho existem itens de diversos conteúdos, abrangendo toda a escala de dificuldade, mesmo que, naturalmente, existam conteúdos mais fáceis e outros mais difíceis. Existem diversos estudos específicos sobre a composição dos bancos e, inclusive, propostas como a de Kingsbury e Zara (1989) sobre a inclusão do balanceamento de conteúdo como um dos critérios nos modelos de seleção de itens.

Tabela 5.6: Presença de descritores por teste simulado

	Mínimo	Máximo	Média	DP
Estratificação	17	25	21,8	1,3
Máx. Informação	18	25	21,7	1,2
Metas	18	25	21,7	1,0

Tabela 5.7: Teste de Kruskal-Wallis da distribuição de descritores por modelo de seleção de itens

Classificação

	Modelo	N	Média
Descritores	Estratificação	1000	1542,20
	Máx. Informação	1000	1491,94
	Metas	1000	1467,35
	Total	3000	

Estatísticas (a,b)

	Descritores
Qui-quadrado	4,162
G. Liberdade	2
p-valor	0,125

a. Teste de Kruskal Wallis

b. Variável de grupo: Modelo

Para avaliar a distribuição de conteúdos pelos testes podemos nos basear em duas informações: o número de descritores diferentes presentes nos testes simulados e o número de repetições de descritores em um mesmo teste. As Tabelas 5.6 e 5.7 mostram, respectivamente, a análise geral do número de descritores diferentes utilizados por teste e o resultado do teste não paramétrico de Kruskal-Wallis que atesta a similaridade da distribuição de descritores obtida pelos modelos de seleção (p-*value* = 0,125). Dificilmente um único teste seria composto de itens de todos os descritores diferentes no banco, logo, podemos considerar que o conteúdo dos testes teve um bom balanceamento por ter obtido uma média próxima a 22 descritores por teste. Considerando a presença de itens de 27 descritores no banco, mesmo os testes com menor variação de descritores conseguiram utilizar em torno de 65% das possibilidades.

Tabela 5.8: Teste de Kruskal-Wallis da repetição de descritores nos testes por modelo de seleção de itens

Classificação

	Modelo	N	Média
Descritores	Estratificação	1000	1554,19
	Máx. Informação	1000	1466,04
	Metas	1000	1481,27
	Total	3000	

Estatísticas (a,b)

	Repetições
Qui-quadrado	7,231
G. Liberdade	2
p-valor	0,027

a. Teste de Kruskal Wallis

b. Variável de grupo: Modelo

Tabela 5.9: Número de repetições de descritores por testes em cada modelo

Repetições	Estratificação	Máx. Informação	Metas de Erro
2	5	1	
3	347	352	338
4	454	540	560
5	149	95	90
6	36	12	12
7	7		
8	2		

Quanto ao número de repetições de descritores por teste, considerando um nível de significância de 0,05, o teste de Kruskal-Wallis indica que não é possível considerar que os resultados sejam similares (Tab. 5.8). O *Mean Rank* do teste indica que há uma pequena disparidade do modelo de estratificação em relação aos outros, logo, ao analisarmos as frequências com que essas repetições acontecem, os modelos MIF e de metas devem ter um comportamento parecido. Isso pode ser constatado pela Tabela 5.9 que apresenta o número de repetições de descritores por testes em cada modelo. Os modelos MIF e de metas apresentam números quase idênticos, enquanto os resultados do modelo de estratificação podem ser considerados piores por apresentarem um número maior de testes com mais repetições de descritores, principalmente 5 e 6 repetições por teste. Assim, encontramos um comportamento inesperado dos modelos MIF e de metas do erro padrão, pois apresentam menos repetições de conteúdo por teste apesar de utilizarem muito menos itens no total das simulações.

5.3 ANÁLISE DAS METAS

O comportamento do modelo de seleção de itens por metas do erro padrão em relação ao cumprimento, ou não, das metas estabelecidas se baseia em três pontos: até quais itens as metas de decaimento do erro padrão estão sendo cumpridas, quantas metas são cumpridas e se há recuperação no cumprimento de metas após alguma falha. Além disso, é necessário avaliar se há uma relação entre o cumprimento das metas e a precisão dos resultados obtidos nos testes.

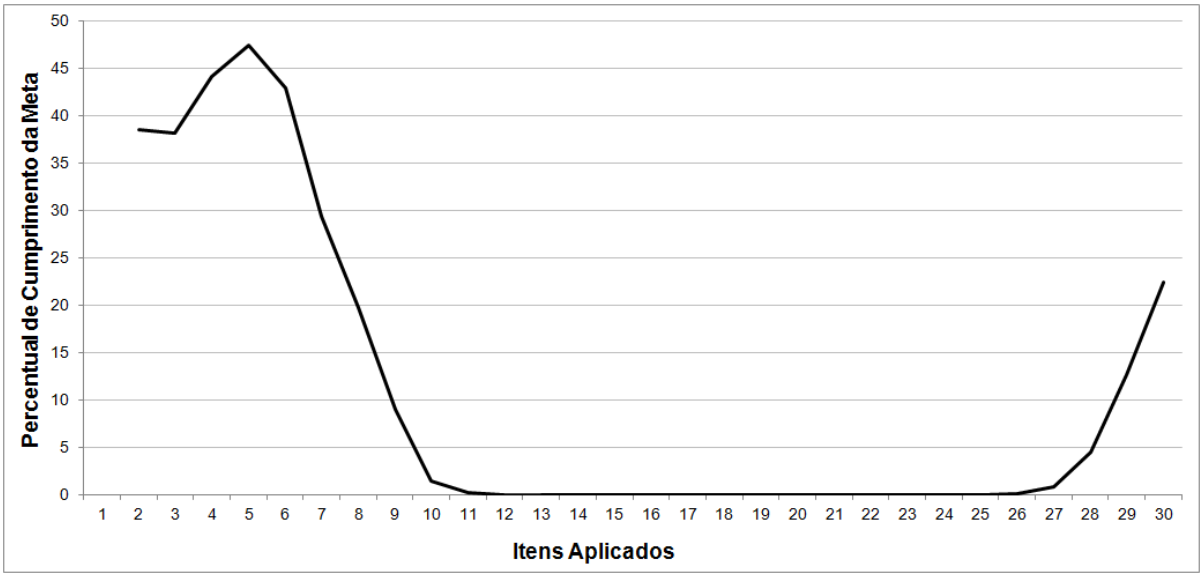

Figura 5.7: Percentual de cumprimento de metas por seleção de itens

A Figura 5.7 mostra o percentual de metas cumpridas a cada seleção de itens. Como o primeiro item é selecionado para se obter o erro padrão inicial e calcular a metas, a análise começa a partir da seleção do segundo item. Podemos observar que as metas são alcançadas, em uma proporção cada vez menor, até o décimo item selecionado e só voltam a ser cumpridas pelos testes que conseguem atingir a meta final. Se avaliarmos o percentual de cumprimento das metas por testes na Figura 5.8, observamos que aproximadamente 25% dos testes não cumpriram nenhuma meta. Entre esses casos podemos destacar que a maioria deles acontece nas mesmas regiões da escala em que o modelo apresentou-se mais irregular, em torno de 1 e acima de 2 (Fig. 5.9).

Se analisarmos apenas os casos que cumpriram pelo menos uma das metas, obtemos uma média de 4 metas cumpridas por teste e observamos que o número de metas cumpridas cai para os testes que se encontram nas regiões da escala que foram mais problemáticas

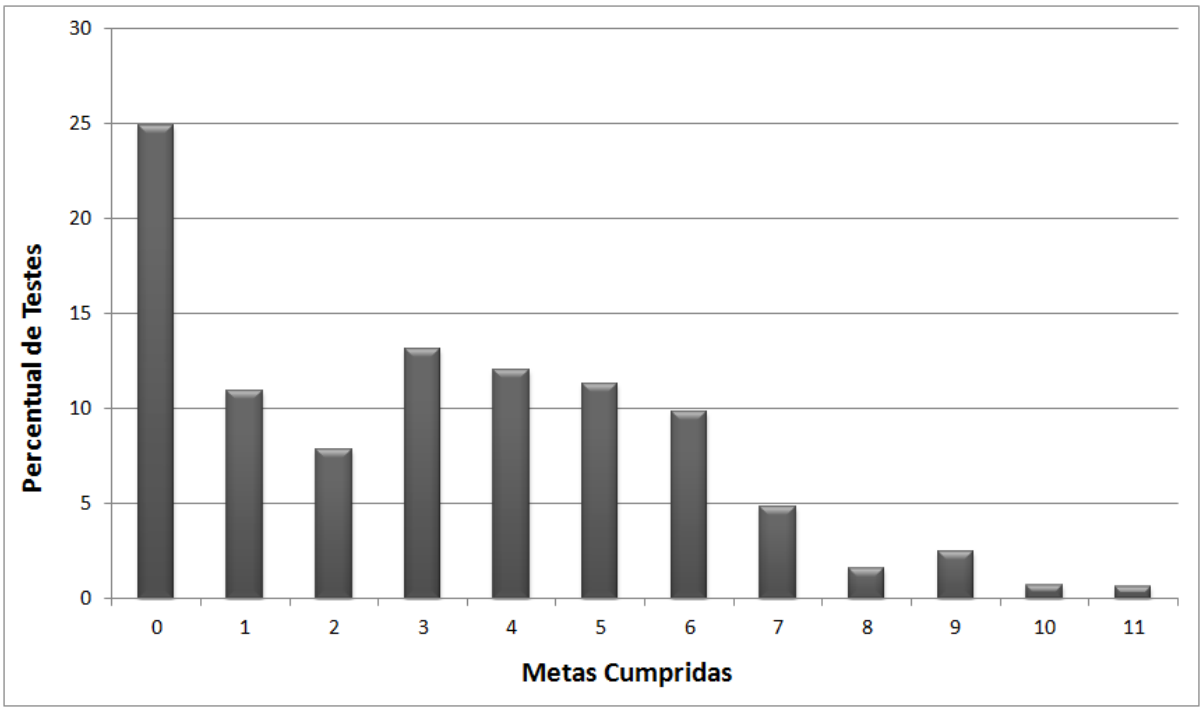

Figura 5.8: Percentual de cumprimento de metas por testes

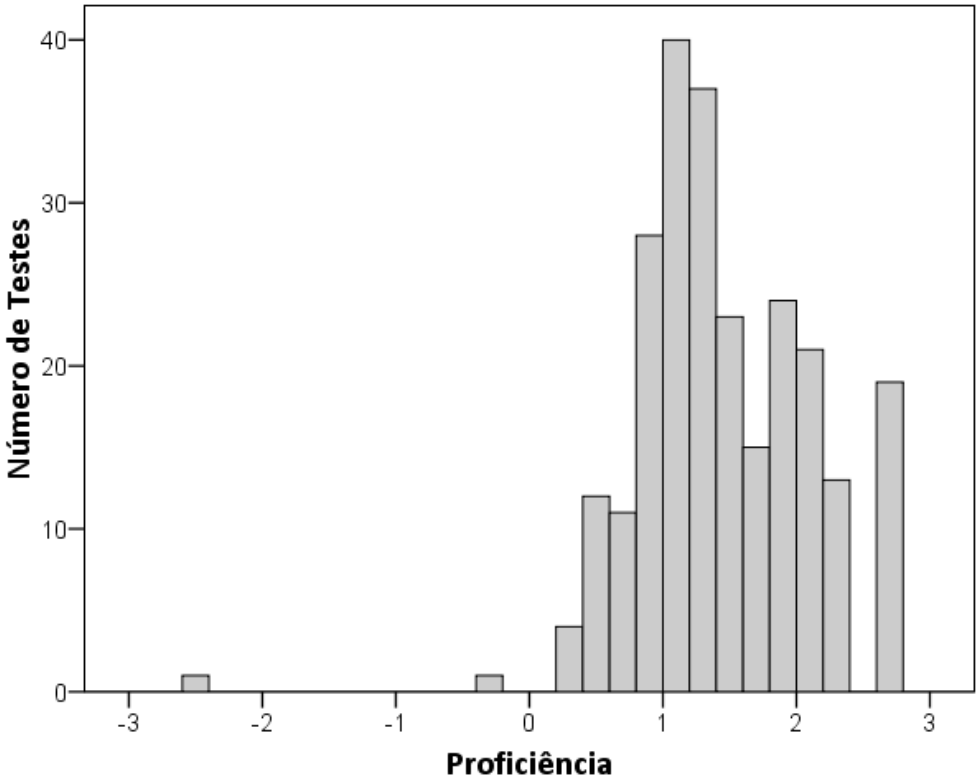

Figura 5.9: Distribuição da proficiência dos testes que não cumpriram nenhuma das metas

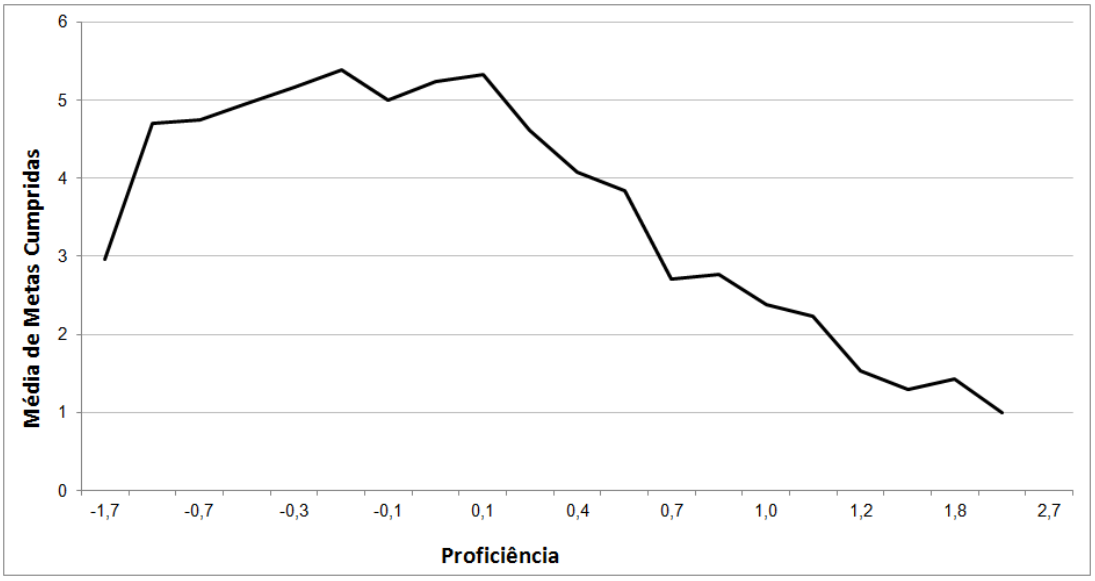

Figura 5.10: Média de metas cumpridas de acordo com a proficiência nos testes

para o modelo (Fig. 5.10) e que o resultado é mais preciso para os testes que cumpriram um maior número de metas (Fig. 5.11). É importante ressaltar que, entre todos os que cumpriram pelo menos uma meta, somente um teste atingiu apenas a meta final garantindo a precisão sem cumprir nenhuma outra meta anterior. Nesse caso o erro padrão acompanhou o decaimento das metas e, mesmo não as cumprindo ao longo do teste, se manteve próximo o suficiente para conseguir atingir a meta final. Os 224 testes que conseguiram atingir a precisão de 0,2 para o erro padrão tiveram uma média de 6 metas cumpridas por teste, sendo 4 delas apenas no primeiro terço do teste.

Tabela 5.10: Metas cumpridas e erro padrão dados pela combinação entre testes precisos e recuperação de metas

Precisão 0,2 Atingida	Recuperação de Metas	Número de Testes	Metas Cumpridas	Erro Padrão
Não	Não	400	1,25	0,25
	Sim	376	3,39	0,23
Sim	Sim	224	6,00	0,20

A última análise a ser feita sobre o modelo de metas do erro padrão é sobre a capacidade de recuperação das metas, isto é, se quando um teste deixa de cumprir uma meta ele é capaz de compensar essa perda nas próximas seleções de itens. De todos os testes aplicados, em 60% deles houve recuperação de metas e, mesmo entre aqueles em que a precisão final não foi atingida, seu erro padrão é significativamente menor do que nos testes

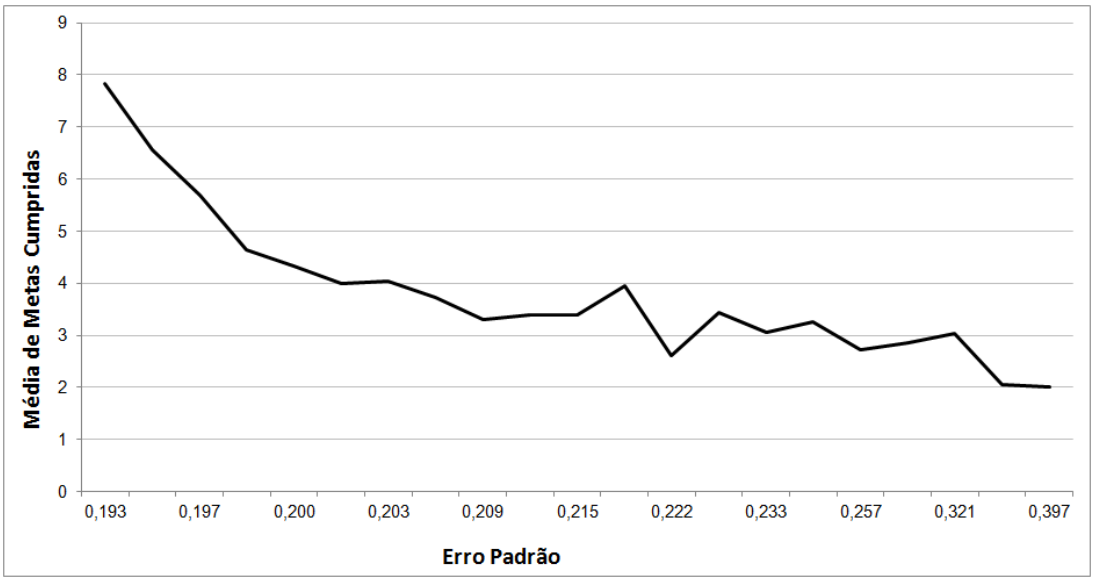

Figura 5.11: Média de metas cumpridas de acordo com a precisão dos testes

Tabela 5.11: Teste de Kruskal-Wallis do erro padrão pela combinação entre testes precisos e recuperação de metas

Classificação

		N	Média
Combinação	Impreciso Sem Recuperação	400	668,00
	Impreciso Com Recuperação	376	553,46
	Preciso Com Recuperação	224	112,50
	Total	1000	

Estatísticas (a,b)

	Erro Padrão
Qui-quadrado	551,432
G. Liberdade	2
p-valor	0,000

a. Teste de Kruskal Wallis

b. Variável de grupo: Modelo

em que não houve nenhuma recuperação durante a aplicação do teste. A Tabela 5.10 apresenta as médias de metas cumpridas e do erro padrão dados pela combinação entre testes que atingiram a precisão final e os que conseguiram recuperação de metas durante o processo. O teste de Kruskal-Wallis para o erro padrão dadas essas combinações atesta a diferença significativa entre os resultados (Tab. 5.11). A recuperação de metas durante o estágio inicial da aplicação do teste, mesmo que aconteça em apenas alguns itens, pode ajudar a manter o erro padrão em um comportamento próximo das metas e garantir o cumprimento da meta final ou, pelo menos, uma precisão próxima à meta.

6 CONCLUSÕES E CONSIDERAÇÕES FINAIS

Nas últimas décadas, a evolução das técnicas utilizadas nas avaliações educacionais, aliadas ao avanço e disseminação da informática, possibilitaram novas opções para determinar o grau de domínio de indivíduos nas mais diversas áreas do conhecimento. Os Testes Adaptativos Computadorizados (TAC's) trouxeram não apenas as características dinâmicas da tecnologia, mas, também, a personalização e, consequentemente, a precisão necessárias às avaliações individuais. Durante esse processo evolutivo, experimentos dos tipos mais variados surgiram buscando a melhoria das diversas características que compõem um TAC. Porém, existem duas características básicas que são essenciais para a própria existência do TAC: o banco de itens e o modelo de seleção de itens. Um modelo de seleção visa, principalmente, garantir a precisão da estimativa de proficiência obtida no teste, porém, deve equilibrar esse objetivo com as limitações do banco e com a proteção dos itens à superexposição.

O modelo de metas do erro padrão, proposto neste trabalho, foi desenvolvido com o intuito de tentar controlar o comportamento do teste, mantendo uma estimativa razoavelmente precisa ao longo de sua aplicação visando cumprir uma meta de erro padrão final. Além disso, busca trabalhar com um grau menor de determinismo, procurando por soluções que atendam às metas independentemente de serem as soluções ótimas. Essa característica tem como objetivo reduzir a exposição dos itens de maior discriminação, que são os que produzem os resultados ótimos em termos de precisão e, por isso, correm maior risco de serem selecionados para aplicação. Como parâmetro de comparação para os resultados, foram simulados dois conhecidos modelos de seleção de itens com estruturas completamente diferentes: o modelo de estratificação e o de máxima informação de Fisher (MIF).

O modelo MIF é uma referência em testes adaptativos desde o seu desenvolvimento no início dos anos 1980, obtendo os melhores resultados em termos de precisão. Porém, sua característica predominantemente determinística, como a maioria dos modelos existentes, faz com que seja necessário que se aplique algum método de controle de exposição dos itens. O modelo de estratificação foi proposto há pouco mais de uma década, mas conseguiu destaque por tratar a exposição de itens como um fator central ao modelo, abrindo mão

de resultados mais precisos em favor do equilíbrio entre precisão e controle de exposição.

O comportamento de qualquer modelo de seleção de itens é diretamente afetado por, pelo menos, três fatores: o objetivo do teste, a composição do banco de itens e o tamanho do teste. Assim, a comparação de desempenho dos modelos não pode considerar exclusivamente a precisão obtida nos testes. Se, por exemplo, utilizarmos um banco de itens pequeno, o modelo MIF continuará obtendo resultados mais precisos, mas causará uma superexposição dos itens de maior discriminação mesmo com o uso de um método de controle. Em caso de testes em que o objetivo é avaliar um sistema de ensino, é possível exigir uma precisão menor por indivíduo e utilizar um teste com menos itens, protegendo o sigilo do banco. O modelo de metas do erro padrão, assim como o MIF, conseguiu atingir uma precisão de 0,3 com testes que chegaram a utilizar apenas 11 itens. Comparados ao exemplo dado da avaliação de Língua Portuguesa da 8ª série do SAEB 2003 (seção 2.3), todos os modelos simulados nesse trabalho tiveram desempenho superior.

Considerando os diferentes aspectos que envolvem a aplicação de um TAC, o modelo proposto conseguiu, em termos gerais, um desempenho razoável na tentativa de equilibrar precisão no teste com exposição menor de itens. Conseguir manter as médias das estimativas do erro padrão em um comportamento próximo ao do modelo MIF utilizando uma variação de itens 33% maior foi, provavelmente, o resultado mais expressivo do modelo. Na realidade, a grande vantagem do modelo de metas está no método como ele lida com os itens. O banco pode ser modificado, adicionando-se ou retirando-se itens, sem que seja necessário qualquer tipo de procedimento antes da utilização do modelo.

Existem estudos específicos para o aprimoramento do banco de itens, desde a forma como devem ser compostos até a possibilidade de revezamento de grupos de itens similares a fim de evitar a superexposição. No caso dos modelos tradicionais, qualquer mudança no banco de itens acarreta algum tipo de procedimento de adequação ao modelo, desde o processo simples de estratificação do banco até a complexa, e geralmente lenta, reestimação das taxas de exposição dos itens. Reestimar as taxas de exposição pode se tornar inviável em sistemas que contam com muitos itens no banco ou que tenham um grande número de indivíduos examinados em pouco tempo, uma vez que as taxas são recalculadas com base nas aplicações dos testes.

Porém, apesar da facilidade para gerenciamento do banco de itens proporcionada pelo modelo de metas, este aspecto dos testes é um dos pontos que comprometem o desempenho

do modelo proposto. As irregularidades no banco de itens o afetaram de forma mais acentuada do que os tradicionais, uma vez que ele parte do princípio de que deve existir um conjunto de opções viáveis para a seleção. Pela própria definição dos testes adaptativos, qualquer modelo de seleção é afetado quando há uma região da escala em que não se encontram itens, como acontece, por exemplo, na falta de itens cobrindo as extremidades da escala. Nas simulações para o modelo de metas, a redução do número de itens em uma faixa central da escala já foi suficiente para interferir no seu comportamento e, consequentemente, no seu desempenho.

Outro ponto que pode interferir diretamente no desempenho do modelo é a diferença razoável nos valores para o erro padrão em caso de resposta correta ou incorreta aos itens na fase inicial do teste. Devemos destacar que o modelo de metas do erro padrão se baseia em objetivos intermediários definidos a partir de uma meta global e esse elemento de aleatoriedade pode ser decisivo no não cumprimento de metas mais rigorosas. Evidentemente, essa característica foi devidamente considerada no desenvolvimento do modelo e optou-se pela utilização da média entre esses valores como fator de viabilidade dos itens para cumprir as metas. Porém, uma sequência inicial de itens que não cumprem as metas torna-se condição adversa o suficiente para afetar todo o teste e eliminar a possibilidade de recuperação da precisão.

Uma diferença importante entre os modelos de estratificação e os baseados em medidas de informação ou na estimativa do erro padrão é a complexidade dos algoritmos de seleção e, consequentemente, o tempo necessário para que o algoritmo produza um resultado. Por utilizar o algoritmo mais complexo entre os três modelos simulados, o modelo de seleção por metas do erro padrão demanda mais tempo na seleção de itens e, nas simulações deste trabalho, necessitava, aproximadamente, do dobro do tempo do modelo MIF para simular um teste. Naturalmente, esse é um fator que deve ser considerado no caso da sua implementação e utilização em testes reais. Nesse caso, a infraestrutura computacional disponível será determinante para um bom desempenho do modelo, uma vez que há a possibilidade de centenas, talvez milhares, de pessoas utilizando o sistema ao mesmo tempo.

Há, pelo menos, duas alternativas plausíveis para possíveis melhorias desse aspecto do modelo: a exigência que as duas previsões de erro padrão obedeçam à meta estipulada ou uma nova projeção das metas a cada seleção de item. Exigir que ambas as previsões do

erro padrão possam cumprir a meta, independentemente da resposta correta ou incorreta ao item, torna a seleção mais determinística, pois reduz o número de itens que atendem ao critério da meta e aumenta a taxa de exposição dos itens. Além disso, cria mais possibilidades de que o erro padrão estimado após a seleção de um item não só cumpra sua meta, mas, também, ultrapasse a meta seguinte. Esse comportamento seria totalmente contrário à ideia do modelo de conseguir seleções que produzam resultados suficientemente bons, sem a necessidade de serem os melhores possíveis.

A possibilidade de refazer a projeção de metas não entra em atrito com a ideia básica do modelo, pelo contrário, é uma opção de ajuste durante o andamento do teste, condizente com a própria natureza dinâmica de um TAC. Esse procedimento pode aumentar a probabilidade de recuperação das metas e aumentar a precisão do teste. Porém, o modelo de metas é conceitualmente mais complexo e, consequentemente, mais exigente em termos computacionais do que os outros modelos simulados neste trabalho. Seria necessário um teste para avaliar a viabilidade de refazer suas metas a cada seleção sem afetar o tempo de resposta aos indivíduos em casos de testes reais.

Uma possibilidade mais simples tendo por objetivo uma melhor precisão final do teste seria um esforço para se obter melhores resultados nas seleções dos primeiros itens. Dadas as análises do capítulo 5 deste trabalho, os testes de melhor precisão do modelo foram aqueles que apresentaram as maiores médias de cumprimento de metas na fase inicial do teste, mais especificamente nos primeiros dez itens selecionados. Talvez, um melhor ajuste em alguns dos parâmetros necessários ao modelo de metas possa ajudar a cumprir de forma mais efetiva essas metas iniciais. Seriam necessários outros testes para encontrar novas configurações, por exemplo, da razão da PG e do número de itens previsto que permitam um melhor desempenho no início do teste.

Uma outra proposta bastante diferenciada seria a possibilidade de adaptar o modelo de metas para a seleção de itens em pequenos blocos, e não item a item como nos TAC's tradicionais. Esse modelo de aplicação por blocos adaptados já é utilizado em alguns testes como, por exemplo, no Graduate Record Examination (GRE). Este é um dos mais populares testes de admissão a cursos de graduação nos Estados Unidos, país onde a cultura de testes adaptativos está em um estágio bem mais avançado e difundido que no Brasil. Nesse caso, uma possibilidade de mudança no modelo seria o uso de algoritmos genéticos para obtenção de blocos de itens que atendessem a um determinado nível de

precisão, dependendo do estágio do teste. Assim, o modelo daria uma capacidade dinâmica a uma característica estática, que são os blocos de itens.

A área de testes adaptativos ainda possui diversos pontos a serem aperfeiçoados e muitas soluções a serem experimentadas. Os estudos em TAC's se tornaram complexos a ponto de criarem uma separação de áreas, com pesquisadores se dedicando a apenas uma característica de cada vez. A proposta deste trabalho não teve como objetivo primordial se tornar uma referência ou produzir resultados melhores que os modelos já existentes. Seu objetivo é apresentar um modelo que funcione de forma diferente e que, talvez, possa servir de base para futuros projetos nessa área dinâmica que ainda é pouco estudada em nosso país.

REFERÊNCIAS

BAKER, F. B. **The Basics of Item Response Theory**. 2nd. ed. Wisconsin, USA: ERIC Clearinghouse on Assessment and Evaluation, 2001.

BAKER, F. B.; KIM, S. H. (Ed.). **Item Response Theory: Parameter Estimation Techniques**. 2nd. ed. New York, USA: CRC Press, 2004.

BARRADA, J. R.; OLEA, J.; ; ABAD, F. J. Rotating item banks versus restriction of maximum exposure rates in computerized adaptive testing. **The Spanish Journal of Psychology**, v. 11, n. 2, p. 618 – 625, 2008.

BARRADA, J. R.; OLEA, J.; PONSODA, V.; ABAD, F. J. A method for the comparison of item selection rules in computerized adaptive testing. **Applied Psychological Measurement**, v. 34, n. 6, p. 438 – 452, 2010.

BIRNBAUM, A. Some latent trait models and their models and their use in inferring an examineeś ability. In: LORD, F. M.; NOVICK, M. R. (Ed.). **Statistical Theories of Mental Test Scores**. Reading, USA: Addison-Wesley, 1968.

CHANG, H. H.; QIAN, J.; YING, Z. a-stratified multistage computerized adaptive testing with b blocking. **Applied Psychological Measurement**, v. 25, n. 4, p. 333 – 341, 2001.

CHANG, H. H.; YING, Z. A global information approach to computerized adaptive testing. **Applied Psychological Measurement**, v. 20, n. 3, p. 213 – 229, 1996.

CHANG, H. H.; YING, Z. a-stratified multistage computerized adaptive testing. **Applied Psychological Measurement**, v. 23, n. 3, p. 211 – 222, 1999.

CHANG, S. W.; ANSLEY, T. N. A comparative study of item exposure control methods in computerized adaptive testing. **Journal of Educational Measurement**, v. 40, n. 1, p. 71 – 103, 2003.

CHEN, S. Y.; ANKENMANN, R. D.; CHANG, H. H. A comparison of item selection rules at the early stages of computerized adaptive testing. **Applied Psychological Measurement**, v. 24, n. 3, p. 241 – 255, 2000.

COSTA, D. R. **Métodos Estatísticos em Testes Adaptativos Informatizados**. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, 2009. Disponível em: <http://www.pg.im.ufrj.br/teses/Estatistica/Mestrado/121.pdf>.

CRONBACH, L. J.; GLESER, G. C.; NANDA, H.; RAJARATNAM, N. **The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles**. New York, USA: John Wiley & Sons, 1972.

DAVEY, T.; PARSHALL, C. G. New algorithms for item selection and exposure control with computerized adaptive testing. In: **Paper presented at the annual meeting of the American Educational Research Association**, April 1995. Disponível em: <http://files.eric.ed.gov/fulltext/ED421525.pdf>.

EGGEN, T. J. H. M.; STRAETMANS, G. J. J. M. Computerized adaptive testing for classifying examinees into three categories. **Educational and Psychological Measurement**, v. 60, n. 5, p. 713 – 734, 2000.

FLAUGHER, R. Item pools. In: WAINER, H. (Ed.). **Computerized Adaptive Testing: A Primer**. Mahwah, USA: Lawrence Erlbaum Associates, 2000.

GOLDBERG, D. **Genetic Algorithms in Search, Optimization and Machine Learning**. Reading, USA: Addison-Wesley, 1989.

HETTER, R. D.; SYMPSON, J. B. Item exposure control in cat-asvab. In: SANDS, W. A.; WATERS, B. K.; MCBRIDE, J. R. (Ed.). **Computerized Adaptive Testing: From Inquiry to Operation**. Washington - D.C., USA: APA Books, 1997. cap. 4.

INEP. **ENEM - Dúvidas Frequentes**. 2012. Acesso em: 24 nov. 2012. Disponível em: <http://enem.inep.gov.br/duvidas-frequentes.html>.

KINGSBURY, G. G.; ZARA, A. R. Procedures for selecting items for computerized adaptive tests. **Applied Measurement in Education**, v. 2, n. 4, p. 359 – 375, 1989.

KOLEN, M. J.; BRENNAN, R. L. **Test Equating, Scaling, and Linking: Methods and Practices**. 2nd. ed. New York, USA: Springer, 2004.

LINDEN, W. J. van der. Bayesian item selection criteria for adaptive testing. **Psychometrika**, v. 63, n. 2, p. 201 – 216, 1998.

LINDEN, W. J. van der. Some alternatives to sympson hetter item exposure control in computerized adaptive testing. **Journal of Educational and Behavioral Statistics**, v. 28, n. 3, p. 249 – 265, 2003.

LINDEN, W. J. van der; GLAS, C. A. W. (Ed.). **Computerized Adaptive Testing: Theory and Practice**. Netherlands: Kluwer Academic, 2000.

LINDEN, W. J. van der; HAMBLETON, R. K. (Ed.). **Handbook of Modern Item Response Theory**. New York, USA: Springer, 1996.

LINDEN, W. J. van der; PASHLEY, P. J. Item selection and ability estimation in adaptive testing. In: LINDEN, W. J. van der; GLAS, C. A. W. (Ed.). **Computerized Adaptive Testing: Theory and Practice**. Netherlands: Kluwer Academic, 2000.

LORD, F. M. **Applications of Item Response Theory To Practical Testing Problems**. New York, USA: Routledge, 1980.

LORD, F. M.; NOVICK, M. R. (Ed.). **Statistical Theories of Mental Test Scores**. Reading, USA: Addison-Wesley, 1968.

MISLEVY, R. J. Bayes modal estimation in item response models. **Psychometrika**, v. 51, n. 2, p. 177 – 195, 1986.

MITCHELL, T. M. **Machine Learning**. New York, USA: McGraw-Hill, 1997.

OWEN, R. J. A bayesian sequential procedure for quantal response in the context of adaptive mental testing. **Journal of the American Statistical Association**, v. 70, n. 350, p. 351 – 356, 1975.

PASQUALI, L.; PRIMI, R. Fundamentos da teoria da resposta ao item. **Avaliação Psicológica**, v. 2, n. 2, p. 99 – 110, 2003.

RASCH, G. **Probabilistic Models for Some Intelligence and Attainment Tests**. Copenhagen, Denmark: Danish Institute for Educational Research, 1960. Expanded Edition - Reprint 1980. Chicago, USA: The University of Chicago Press.

RECKASE, M. D. Designing item pools to optimize the functioning of a computerized adaptive test. **Psychological Test and Assessment Modeling**, v. 52, n. 2, p. 127 – 141, 2010.

SEGALL, D. O. A sharing item response theory model for computerized adaptive testing. **Journal of Educational and Behavioral Statistics**, v. 29, n. 4, p. 439 – 460, 2004.

STOCKING, M. L. **Three Practical Issues for Modern Adaptive Testing Item Pools**. Reports - Evaluative/Feasibility, Educational Testing Service, Princeton, USA, Feb. 1994.

STOCKING, M. L.; LEWIS, C. **A New Method of Controlling Item Exposure in Computerized Adaptive Testing**. Research Report 95-25, Educational Testing Service, Princeton, USA, Aug. 1995.

VEERKAMP, W. J. J.; BERGER, M. P. F. Some new item selection criteria for adaptive testing. **Journal of Educational and Behavioral Statistics**, v. 22, n. 2, p. 203 – 226, 1997.

VELDKAMP, B. P. Bayesian item selection in constrained adaptive testing using shadow tests. **Psicológica**, v. 31, n. 1, p. 149 – 169, 2010.

VERSCHOOR, A. J. **Genetic Algorithms for Automated Test Assembly**. Tese (Doutorado) — University of Twente, Enschede, Netherlands, 2007. Disponível em: <http://doc.utwente.nl/60710/>.

WAINER, H. (Ed.). **Computerized Adaptive Testing: A Primer**. Mahwah, USA: Lawrence Erlbaum Associates, 2000.

WEISS, D. J. **The Stratified Adaptive Computerized Ability Test**. Research Report 73-3, University of Minnesota, Department of Psychology, Psychometric Methods Program, Minneapolis, USA, Sep. 1973.

WEISS, D. J.; KINGSBURY, G. G. Application of computerized adaptive testing to educational problems. **Journal of Educational Measurement**, v. 21, n. 4, p. 361 – 375, 1984.