UNIVERSIDADE FEDERAL DE SANTA CATARINA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

FERNANDO DE JESUS MOREIRA JUNIOR

SISTEMÁTICA PARA IMPLANTAÇÃO DE TESTES ADAPTATIVOS INFORMATIZADOS BASEADOS NA TEORIA DA RESPOSTA AO ITEM

FLORIANÓPOLIS

2011

FERNANDO DE JESUS MOREIRA JUNIOR

SISTEMÁTICA PARA IMPLANTAÇÃO DE TESTES ADAPTATIVOS INFORMATIZADOS BASEADOS NA TEORIA DA RESPOSTA AO ITEM

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina como requisito parcial para obtenção de grau de Doutor em Engenharia de Produção.
Orientador: Prof. PhD. Dalton Francisco de Andrade
Coorientador: Prof. Antonio Cezar Bornia, Dr.

FLORIANÓPOLIS

2011

Catalogação na fonte pela Biblioteca Universitária
da
Universidade Federal de Santa Catarina

M838s Moreira Junior, Fernando de Jesus
Sistemática para a implantação de testes adaptativos informatizados baseados na teoria da resposta ao item [tese] / Fernando de Jesus Moreira Junior ; orientador, Dalton Francisco de Andrade. - Florianópolis, SC, 2011.
334 p.: il., grafs., tabs.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia de Produção.

Inclui referências

1. Engenharia de produção. 2. Testes adaptativos informatizados. 3. Teoria da Resposta do Ítem. I. Andrade, Dalton Francisco de. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia de Produção. III. Título.

CDU 658.5

Fernando de Jesus Moreira Junior

SISTEMÁTICA PARA IMPLANTAÇÃO DE TESTES ADAPTATIVOS INFORMATIZADOS BASEADOS NA TEORIA DA RESPOSTA AO ITEM

Esta tese foi julgada e aprovada para a obtenção do título de Doutor em Engenharia de Produção no Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina.

Florianópolis, 25 de novembro de 2011.

Prof. Antonio Cezar Bornia, Dr.
Coorientador e Coordenador do PPGEP/UFSC

BANCA EXAMINADORA

Prof. Dalton Francisco de Andrade, PhD. Orientador Universidade Federal de Santa Catarina	Prof. Héliton Ribeiro Tavares, Dr. Examinador Externo Universidade Federal do Pará - UFPA
Profa. Mariana Curi, Dra. Examinadora Externa Universidade de São Paulo - USP	Prof. Francisco Aranha Filho, Dr. Examinador Externo Fundação Getúlio Vargas – SP
Prof. Renato Cislaghi, Dr. Membro Universidade Federal de Santa Catarina	Prof. Roberto Carlos S. Pacheco, Dr. Membro Universidade Federal de Santa Catarina

Dedico esse trabalho
à minha esposa Tatiane,
ao meu filho Daniel,
aos meus pais Fernando e Wasila,
e a Deus.

AGRADECIMENTOS

Gostaria de prestar meus agradecimentos àqueles que me ajudaram e incentivaram na elaboração desse trabalho e na conclusão desse curso.

Ao meu professor orientador Dalton Francisco de Andrade, pelo conhecimento compartilhado da Teoria da Resposta ao Item (TRI) e pelas idéias, discussões e desafios propostos nesse trabalho.

Ao professor coorientador Antonio Cezar Bornia pelo apoio, pelas idéias, sugestões e discussões sobre esse trabalho.

Aos demais professores membros da banca desse trabalho, Héliton Tavares, Mariana Curi, Francisco Aranha, Renato Cislaghi e Roberto Pacheco, pelas correções e contribuições.

Ao colega e amigo Rafael Tezza, pelo companheirismo, amizade, troca de experiências, trabalhos conjuntos e pelo contato com o DETRAN-SC.

Ao DETRAN-SC (Departamento Estadual de Trânsito de Santa Catarina), representado pelo Senhor Diretor Vanderlei Rosso, pela autorização na utilização do banco de dados de 2008 para o estudo de caso nesse estudo.

Ao CIASC-SC (Centro de Informática e Automação do Estado de Santa Catarina), representado pela Sra. Rosmeri Paludo, responsável pela compilação do banco de dados com as informações necessárias para esse estudo.

Às professoras Silvana Bortolotti e Vera do Carmo pela amizade, troca de experiências e trabalhos conjuntos.

Aos demais colegas Priscilla, Igor, Adilson, Luciano, Blênio, Ivan, Juliano, Andréia e Lizandra pela amizade e troca de experiências nos seminários da TRI no PPGEP/UFSC.

Aos professores Bornia, Vera do Carmo, Adriano, Paulo e Pedro Barbetta pela troca de experiências nos seminários da TRI no PPGEP/UFSC.

Ao professor Jean Piton-Gonçalves (DE/UFSCar), ao professor David J. Weiss (*Department of Psychology / University of Minnesota*), aos professores da Espanha, David Aguado, Javier Revuelta, Juan Ramón Barrada, Pedro Hontangas e Julio Olea e a colega Denise Costa pelo fornecimento de material bibliográfico, esclarecimento de dúvidas e troca de informações.

À Universidade Federal de Santa Maria (UFSM) pela oportunidade concedida para a realização deste curso.

À Universidade Federal de Santa Catarina (UFSC) e ao Programa de Pós-Graduação em Engenharia de Produção (PPGEP), representado pela secretária Rosimeri, pela oportunidade e pelos conhecimentos adquiridos durante o curso.

À minha esposa Tatiane Rocha Cardoso Moreira, pelo apoio e dedicação nas atividades caseiras e familiares, pelo cuidado com o nosso filho Daniel e pela compreensão da minha ausência nas várias horas de estudo.

Aos meus pais Fernando e Wasila que sempre me incentivavam a estudar desde quando eu era pequeno.

A Deus acima de tudo e por todas as coisas.

(...) A maior genialidade
não é aquela que vem da carga genética
nem a que é produzida pela cultura acadêmica,
mas a que é construída nos vales dos medos,
no deserto das dificuldades,
nos invernos da existência,
no mercado dos desafios. (...)
Todos deveríamos em algum momento da existência
questionar nossas vidas e
analisar pelo que estamos lutando.
Quem não consegue fazer este questionamento
será servo do sistema,
viverá para trabalhar,
cumprir obrigações profissionais
e apenas sobreviver.
Augusto Cury

(...) a IRT parece que veio para ficar
e substituir grande parte
da teoria clássica da psicometria —
isto é um fato que já ocorre
no Primeiro Mundo (...).
Luiz Pasquali

Nem olhos viram
nem ouvidos ouviram,
nem jamais penetrou em coração humano
o que Deus tem preparado para aqueles que o amam.
I Co. 2:9

RESUMO

Essa tese apresenta uma sistemática para a implantação de Testes Adaptativos Informatizados baseados na Teoria da Resposta ao Item (SITAI), com a finalidade de servir como um suporte para a orientação nas etapas do desenvolvimento e elaboração de um TAI. A literatura existente sobre Testes Adaptativos Informatizados (TAI) geralmente está focada no que diz respeito ao algoritmo do TAI. Dessa forma, existe uma carência em encontrar um método para a implantação de TAIs. Nessa tese, foi desenvolvido um método, constituído por diretrizes, que sistematiza etapas para a elaboração de um TAI qualquer, independente das características particulares de cada um, onde o pesquisador poderá utilizá-lo como um guia para o desenvolvimento do TAI que deseja elaborar. A sistemática desenvolvida nessa tese pretende servir como um guia, principalmente ao usuário leigo que deseja desenvolver e implantar um TAI, orientando-o quanto à criação das etapas do teste, salientando os cuidados necessários na elaboração do mesmo, e indicando os métodos e critérios adequados para o teste a ser construído, de acordo com as características e especificações do teste. Nesse trabalho, também foi feito um levantamento bibliográfico exaustivo sobre TAIs, um levantamento bibliográfico sucinto sobre a TRI voltado para a aplicação de TAIs, foram definidas as etapas necessárias para a implantação de um TAI, segundo as características específicas do teste, foram identificados os mais diferentes métodos e critérios utilizados para a elaboração de um TAI e foi realizado um estudo de caso na avaliação teórica para a obtenção da carteira de habilitação de motorista realizada pelo Departamento de Trânsito do Estado de Santa Catarina (DETRAN-SC) a fim de aplicar e validar a sistemática desenvolvida. Embora a SITAI não tenha sido aplicada em sua plenitude, nas etapas em que ela foi aplicada, pode-se constatar que ela proporciona uma boa referência para a implantação de TAIs. Não houve grandes dificuldades para seguir a SITAI em relação à definição do teste, à calibração do banco de itens, à elaboração do algoritmo e à análise da precisão e da validade. Os resultados mostraram as deficiências e as vantagens potenciais da implantação de um TAI no DETRAN-SC.

Palavras-chave: Testes Adaptativos Informatizados, Teoria da Resposta ao Item, avaliação de proficiência.

ABSTRACT

This thesis presents a scheme to the implementation of Computerized Adaptive Testing based on Item Response Theory (SICAT), in order to serve as a support for guidance on the stages of development and preparation of a CAT. The literature on Computerized Adaptive Testing (CAT) is usually focused on what concerns the CAT algorithm. Thus, there is a need to find a method to implement of CAT. In this thesis, a method has been developed, consisting of guidelines that systematize the steps to developement a any CAT, independent of the particular characteristics of each one, where the researcher can use it as a guide for the development of CAT wishes to develop. The scheme developed in this thesis is intended as a guide, especially to the lay user who wishes to develop and deploy a CAT, guiding him to the creation of test steps, highlighting the necessary care in compiling the report, and stating the appropriate methods and criteria for the test to be built, according to the features and specifications of the test. In this work, was also made an extensive literature review on the CAT, a short literature review on the IRT facing the implementation of CAT, defined the steps necessary to deploy an CAT, according to the specific characteristics of the test, we identified the most different methods and criteria used for the preparation of an CAT and we performed a case study in the theoretical evaluation to obtain a driver's license held by the Departamento de Trânsito do Estado de Santa Catarina (DETRAN-SC) to apply and validate the developed scheme. Although SICAT has not been applied in its fullness, the stages in which it was applied, it can be seen that it provides a good reference for the implementation of such. There were no major difficulties to follow the SICAT regarding the definition of the test, calibration of the item bank, the development of the algorithm and analysis of the accuracy and validity. The results showed the weaknesses and the potential benefits of deploying a CAT in DETRAN-SC.

Key words: Computerized Adaptive Testing, Item Response Theory, assessment of proficiency.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ADEPT – *Adaptive English Proficiency Test for the Web*
AERA – *American Educational Research Association*
AMT – *Adaptive Mastery Testing*
ASVAB – *Armed Services Vocational Aptitude Battery*
BI – Banco de Itens
BIB – Bloco Incompleto Balanceado
CARAT – *Computerized Adaptive Reporting and Testing*
CAST – *Computer Adaptive Sequential Testing*
CAT – *Computerized Adaptive Testing* ou *Computering Adaptive Testing* ou *Computer Adaptive Test*
CAT-MD – *Computerized Adaptive Testing on Mobile Devices*
CBAT-2 – *Content-Balanced Adaptive Testing*
CCI – Curva Característica do Item
CCT – *Computerized Classification Testing*
CD-CAT – *Cognitive Diagnosis with Computer Adaptive Testing*
CIASC-SC – Centro de Informática e Automação do Estado de Santa Catarina
CL – Correlação Linear
CNH – Carteira Nacional de Habilitação
DE – Desvio Empírico Médio,
DENATRAN – Departamento Nacional de Trânsito
DETRAN-SC – Departamento de Trânsito do Estado de Santa Catarina
DIF – *Differential item functioning*
EAP – Esperança a Posteriori
EF – Eficiência
ENEM – Exame Nacional do Ensino Médio
EP – Erro Padrão ou Erro Padrão da Medida
EPM – Erro Padrão Médio
ETS – *Educational Testing Service*
EVP – Estimação por Verossimilhança Ponderada
FIFA – *Ful Information Factor Anallysis*
FII – Função de Informação do Item
FIT – Função de Informação do Teste
FRI – Função de Resposta do Item
GAI – Geração Automática de Itens
GGUM – *Generalized Graded Unfolding Model*
GM – *Graphical Modeling*
GMAC – *Graduate Management Admission Council*
GMAT – *Graduate Management Admissions Test*

GMAT – *Graduate Management Admissions Test*
GRE – *Graduate Record Exam*
HME – Habilidade Média Estimada
IACAT – *International Association for Computerized Adaptive Testing*
IC – Intervalo de Confiança
IF – Informação de Fisher
INEP – Instituto Nacional de Estudos e Pesquisas Educacionais
IO – Informação Observada
IRT – *Item Response Theory*
KL – Kullback-Leibler
MAP – Moda a Posteriori
MCMC – *Markov Chain Monte Carlo*
MCP – Modelo de Crédito Parcial
MCPG – Modelo de Crédito Parcial Generalizado
MEC – Ministério da Educação
MEG – Modelo de Escala Gradual
MI – Máxima Informação
MIE – Máxima Informação Esperada
MIG – Máxima Informação Global
ML1 – Modelo Logístico Unidimensional de 1 Parâmetro
ML2 – Modelo Logístico Unidimensional de 2 Parâmetros
ML3 – Modelo Logístico Unidimensional de 3 Parâmetros
ML4 – Modelo Logístico Unidimensional de 4 Parâmetros
MPP – *Maximum Posterior Precision*
MRG – Modelo de Resposta Gradual
MRN – Modelo de Resposta Nominal
MV – Máxima Verossimilhança
MVC – Máxima Verossimilhança Conjunta
MVM – Máxima Verossimilhança Marginal
MVP – Máxima Verossimilhança Ponderada
NCARB – *National Council of Architectural Registration Boards*
NCLEX – *National Council Licensure Examination for registered nurses*
NCME – *National Council on Measurement in Education* (NCME)
NRET – *Number Right Elimination Testing*
NWEA – *Northwest Evaluation Association*
ONR – *Office of Naval Research*
PISA – *Programme for International Student Assessment*
PSD – *Posterior Standard Deviation*
RQEQM – Raiz Quadrada do Erro Quadrado Médio
SAEB – Sistema de Avaliação da Educação Básica

SARESP – Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo
SAT – *Scholastic Aptitude Tests*
SCAALE – Sistema Computadorizado de Avaliação Adaptativa em Larga Escala
SH – *Sympson-Hetter*
SITAI – Sistemática para a Implantação de Testes Adaptativos Informatizados
SPRT – *Sequential Probability Ratio Test*
TA – Teste Adaptativo
TAC – Teste Adaptativo Computadorizado
TAI – Teste Adaptativo Informatizado
TAI-ML3 – TAI baseado no Modelo Logístico de 3 Parâmetros
TAI-MRN – TAI baseado no Modelos de Resposta Nominal
TCM – Teoria Clássica da Medida
TCT – Teoria Clássica dos Testes
TH – Teste de Hipótese
TOEFL – *Test of English as a Foreign Language*
TRI – Teoria da Resposta ao Item
TRI-NA – Modelos Normais Assimétricos
UAB – Universidade Aberta do Brasil
UNB – Universidade de Brasília
WI – *Weighted Information*
WML –*Weighted Maximum Likelihood*

SUMÁRIO

1. INTRODUÇÃO

1.1. CONTEXTUALIZAÇÃO

Uma importante ferramenta utilizada por praticamente todas as organizações para mensurar fatores competitivos tais como qualidade, satisfação, desempenho, proficiência, conhecimento, habilidade, etc., é o instrumento de avaliação, geralmente constituído por um conjunto de questões ou itens. Esses fatores competitivos não podem ser mensurados diretamente, ou seja, não existe, por exemplo, uma régua que meça essas características, daí a necessidade da utilização de questionários. Características desse tipo são chamadas de traços latentes.

Tradicionalmente, a avaliação de traços latentes está baseada na conhecida Teoria Clássica dos Testes – TCT (do inglês, *Classical Test Theory – CTT*), também chamada de Teoria Clássica da Medida (TCM), onde geralmente atribui-se uma nota ao traço latente avaliado caracterizada por uma pontuação bruta ou ponderada de acordo com as respostas aos itens. Segundo Francisco (2005), é comum verificar que, em processos avaliativos, cuja finalidade é a seleção de candidatos, são utilizados resultados obtidos em "provas" (instrumentos avaliativos de desempenho), expressos apenas por seus escores brutos (somatório de questões corretas) ou padronizados. Dessa forma, torna-se inviável a comparação entre respondentes que não foram submetidos às mesmas provas (exceto no caso das "provas paralelas"[1]), como é o caso, por exemplo, dos vestibulares, entre diferentes instituições e em diferentes épocas, e dos exames teóricos para a obtenção da carteira de habilitação para motoristas. O escore bruto depende do nível de dificuldade do exame. Se um exame for gerado com itens mais difíceis, o escore de um indivíduo com certa habilidade provavelmente será menor do que seria se o exame fosse gerado com itens mais fáceis, o que influenciará na nota do exame. Isso significa que um candidato pode ser prejudicado, já que a sua nota não depende somente da sua habilidade, mas também do nível de dificuldade da prova.

Como não existe um instrumento universal para medir traços latentes, diferentes instrumentos de avaliação podem ser utilizados para medir a mesma característica e geralmente fornecem resultados diferentes. Nesse sentido, é importante avaliar os próprios instrumentos

[1] Conforme Gulliksen (1950), provas paralelas (ou testes paralelos) são provas diferentes elaboradas segundo critérios da TCT, de tal forma que elas possuem o mesmo grau de dificuldade e os resultados sejam comparáveis.

de avaliação para verificar se eles realmente são adequados para medir aquilo que se deseja medir. A TCT possui técnicas de análise de itens e de validação de construto adequadas para tal fim, mas que possuem algumas limitações.

Uma alternativa à TCT para medir traços latentes é a Teoria da Resposta ao item – TRI (do inglês, *Item Response Theory – IRT*). Segundo Andrade, Tavares e Valle (2000), a TRI é uma metodologia que sugere formas de representar a relação entre a probabilidade de um indivíduo dar uma certa resposta a um item e seus traços latentes. Além disso, a TRI é uma ferramenta estatística que surgiu também para suprir as necessidades decorrentes das limitações da TCT. Diferentemente da TCT, a TRI possui a propriedade de invariância dos parâmetros, que considera que a proficiência do indivíduo não depende da prova que ele realiza, e os parâmetros dos itens não dependem da proficiência do indivíduo. A TRI também permite comparar as notas de indivíduos que fizeram provas diferentes. Outra vantagem de utilizar a TRI se deve ao fato que o seu desempenho na análise dos itens do teste é mais eficiente do que a análise dos itens através da TCT.

Embora a TRI já tenha uma longa história (PASQUALI, 1996), as primeiras aplicações no Brasil começaram em 1995 na área de avaliação educacional, conforme Andrade, Tavares e Valle (2000). Segundo Klein e Fontanive (1995), avaliação educacional é um sistema de informações que tem como objetivo fornecer diagnóstico e subsídios para a implementação ou manutenção de políticas educacionais, além de prover um contínuo monitoramento do sistema educacional com o intuito de detectar os efeitos positivos ou negativos das políticas adotadas. Para tanto, necessitava-se de uma metodologia mais sofisticada e precisa, que permitisse não só a avaliação pontual mas, sobretudo, a construção de escalas de habilidades que pudessem levar a um acompanhamento do progresso do conhecimento adquirido pelos indivíduos ao longo do tempo. Francisco (2005) destaca que a interpretação qualitativa sobre instrumentos de avaliação quantitativa tem se tornado cada vez mais necessária no contexto educacional, principalmente em termos de Brasil, onde este tipo de abordagem há pouco vem sendo implantada. A TRI veio ao encontro a essa necessidade e passou a ser introduzida progressivamente no Brasil, onde atualmente é utilizada em diversas avaliações educacionais, inclusive no Sistema de Avaliação da Educação Básica (SAEB) e no Exame Nacional do Ensino Médio (ENEM) ambos realizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) do Ministério da Educação (MEC).

As avaliações educacionais geralmente são medidas através de uma aplicação em larga escala (CASSETTARI, 2008; KLEIN; FONTANIVE, 1995; RODRIGUES, 2007). Avaliações em larga escala do tipo "papel e lápis" (do inglês, *paper-and-pencil*) – que é a forma tradicional de avaliações escritas ou impressas em papel e respondidas através da utilização de lápis ou caneta – são baseadas na TCT e podem demandar elevados custos com papel, tinta de impressão, correção, equipamentos de impressão, espaço físico para armazenamento das provas, recursos humanos para elaboração, aplicação e fiscalização das provas, etc. Segundo Fernandes (2009), as despesas com avaliações vão desde a contratação de um elaborador de questões até o pagamento pelo serviço do colaborador que irá transportar os cadernos de avaliação. A impressão de uma grande quantidade de provas, além de gerar custo, também afeta questões relacionadas com a preservação ambiental e sustentabilidade, assuntos de grande importância atualmente. O espaço físico necessário para o armazenamento das provas não é pouco, devido à quantidade de provas realizadas e o tempo em que elas devem permanecer arquivadas. A correção das provas, no sistema atual, está sujeita a erros no processamento e na leitura óptica das respostas (rasuras, formas erradas de preenchimento na grade de respostas), além do tempo gasto para a correção, o que pode comprometer a divulgação dos resultados no prazo estabelecido. Esses testes tradicionais muitas vezes são longos e cansativos, o que pode provocar um desgaste no candidato, de tal forma a comprometer seu desempenho.

A aplicação das avaliações também está sujeita às fraudes, "colas", ou outras situações que constrangem os fiscais ou os candidatos, o que pode levar a processos judiciais que podem resultar na anulação dos testes, prejuízo da imagem da instituição, custos com indenizações, etc. Por exemplo, no ano de 2009, o furto de apenas um caderno da avaliação do ENEM causou um prejuízo de aproximadamente R$ 34 milhões para o governo brasileiro (FERNANDES, 2009) e o transtorno da anulação do exame e da realização de um novo exame, sem contar o prejuízo intangível à imagem do ENEM.

O examinando deve realizar o exame em locais, datas e horários pré-definidos, o que pode levar o mesmo a desmarcar ou perder compromissos, além da preocupação com a locomoção até o local do teste e com possíveis imprevistos que podem causar um atraso fazendo com que o candidato perca a prova. Ainda que esses serviços possam ser terceirizados pela empresa, existe, além do custo relacionado, a possibilidade de possíveis transtornos ou problemas relacionados com a

contratação de empresas terceirizadas (licitações, prazos, desistências, multas, etc.).

Com o advento da informática e a popularização dos computadores, versões informatizadas de vários testes do tipo "papel e lápis" passaram a ser desenvolvidas. Entretanto, na grande maioria dos casos, o que foi feito foi apenas uma mudança nos processos de aplicação do teste, que passou a ser apresentado na tela do computador, e de resolução do teste, que passou a ser realizado por intermédio do computador. Assim, um teste informatizado é o mesmo teste tradicional do tipo "papel e lápis" realizado por meio do computador. Essa foi considerada a primeira geração de testes informatizados, segundo Olea e Hontangas (1999). Uma alternativa para as avaliações do tipo "papel e lápis" tradicional ou informatizada é a utilização dos chamados Testes Adaptativos Informatizados.

Um Teste Adaptativo Informatizado – TAI (do inglês, *Computerized Adaptive Testing*[2] *– CAT*) é um teste baseado na TRI administrado via computador que tem como objetivo apresentar itens adequados para o indivíduo que está realizando o teste, buscando uma melhor estimação da habilidade ou nota desse indivíduo junto com a redução do tempo de realização do teste. Os TAIs, por serem mais rápidos do que os testes tradicionais de "papel e lápis", contribuem para a redução da fadiga do candidato, diminuindo conseqüentemente erros relacionados ao viés e fornecem o resultado do teste mais rapidamente, muitas vezes logo após a aplicação do mesmo. Na aplicação de um TAI, o candidato irá responder questões que procuram medir com eficiência a sua habilidade. Costa (2009) destaca que a criação de testes para serem aplicados por meio do computador é um exemplo de iniciativa que está obtendo êxito.

Embora sejam poucos conhecidos no Brasil, os TAIs possuem uma longa história e aplicações com sucesso em avaliações de larga escala, principalmente nos Estados Unidos e na Europa (ABAD et al., 2010) e mais recentemente na Ásia (NOGAMI; HAYASHY, 2010). Alguns programas de avaliação internacionalmente conhecidos são: o *Test of English Foreign Language* (TOEFL), o *Graduate Record Exam* (GRE), o *Armed Services Vocational Aptitude Battery* (ASVAB), o *Scholastic Aptitude Tests* (SAT), o *Graduate Management Admissions Test* (GMAT), o *National Council of Architectural Registration Boards* (NCARB), o *National Council Licensure Examination for registered nurses* (NCLEX) (OLEA et al., 2004; RECKASE; HE, 2005; TEJADA,

[2] Também são utilizados os termos *Computering Adaptive Testing* e *Computer Adaptive Test*.

2001). No Brasil, os TAIs têm sido desenvolvidos recentemente, com destaque para o programa de teste adaptativo que foi implantado na Universidade de Brasília – UNB para avaliar a proficiência de língua estrangeira (COSTA, 2009; COSTA; FERNANDES, 2009; COSTA et al., 2009; FERNANDES, 2009; KARINO; COSTA; LAROS, 2009).

Embora possa parecer simples desenvolver um TAI, muitos cuidados devem ser tomados para evitar a construção de um teste inadequado. A maneira como as etapas do algoritmo do TAI devem ser desenvolvidas e os métodos e critérios a serem adotados dependem de diversos fatores: tipo de teste, objetivo do teste, tamanho e qualidade do banco de itens, modelo de resposta ao item utilizado, público alvo para a aplicação do teste, etc.

Os trabalhos sobre TAI geralmente são desenvolvidos em estudos teóricos, simulações, aplicações práticas e comparações de métodos e voltados para aspectos particulares do TAI em estudo, como o estudo de algum método específico ou o desenvolvimento de um TAI específico. Existe uma carência em encontrar um método para a implantação de um TAI desde o princípio da sua elaboração até a sua efetiva aplicação, passando por todos os passos necessários e que também possa servir como um guia para a implantação de um TAI.

Diversos benefícios podem ser obtidos com a implantação de um TAI. O processo de impressão dos cadernos de exame será eliminado, reduzindo os custos com papel, tinta de impressão e equipamentos, contribuindo para a preservação do ambiente com a redução de resíduos e solucionando os problemas relacionados com a falta de espaço físico para armazenar os exames. Outro fator importante será a redução no tempo de execução da prova e da divulgação do resultado da prova que poderá ser apresentado imediatamente após a realização do seu exame. Como a correção do teste é realizada pelo próprio sistema imediatamente após a execução do exame, não será mais necessário o processo de correção atual, reduzindo-se custos. A correção automatizada[3] também é isenta de erros, já que não há preenchimento de gabarito, e a correção das questões é feita de forma dinâmica, de modo que o resultado do teste é disponibilizado ao final do mesmo. O resultado do teste pode sair imediatamente após o seu término, acompanhado de um relatório que apresenta o nível de habilidade estimado para o examinando e quais os conteúdos que ele domina ou não. Também existe a possibilidade de o candidato realizar o exame em

[3] Considerando avaliações que podem utilizar questões com múltipla escolha e possuem gabarito. Isso não é possível, por exemplo, na avaliação de redações.

qualquer horário e dia[4], desde que haja um terminal informatizado disponível para a realização do mesmo e condições mínimas de segurança. Diferentemente dos testes clássicos, isso é possível porque o candidato não teve contato anterior com as questões do teste, uma vez que os mesmos não são divulgados nem disponibilizados, como em uma prova impressa. A flexibilidade de data e hora torna o processo mais aceitável pelo candidato. No caso da existência de serviços terceirizados, esses praticamente serão desnecessários com a implantação do TAI, gerando uma redução ou extinção desse tipo de custo, e evitando os possíveis transtornos ou problemas relacionados com a terceirização. Outra característica que merece destaque é a segurança proporcionada ao teste, visto que o TAI não apresenta os problemas que ocorrem no caso da avaliação por "papel e lápis", tais como o vazamento de cadernos de avaliação ou o transporte dos mesmos.

O uso na TRI pode trazer vantagens para a Engenharia de Produção (TEZZA; BORNIA, 2009): a) poder de posicionar indivíduos ou processos de diferentes grupos em uma escala comum, mesmo que estes tenham respondido a itens diferentes, permitindo a identificação de oportunidades de melhoria ou até mesmo *benchmarking*, b) permitir uma avaliação mais precisa das propriedades dos itens e seus resultados e, consequentemente, permitir maior precisão na aplicação de técnicas estatísticas, c) compreender adequadamente as propriedades psicométricas dos instrumentos, d) possibilidade de desenvolver indicadores mais eficientes para avaliar diferenças individuais de processo, práticas, sistemas ou indivíduos, e) maior robustez dos resultados.

1.2. OBJETIVOS

O objetivo geral desta tese é desenvolver uma sistemática para a implantação de Testes Adaptativos Informatizados baseados na Teoria da Resposta ao Item.

Os objetivos específicos dessa tese são:

- Realizar um levantamento bibliográfico exaustivo sobre Testes Adaptativos Informatizados, o qual será utilizado para o desenvolvimento da sistemática proposta nessa tese;
- Realizar um levantamento bibliográfico sucinto sobre a Teoria da Resposta ao Item voltado para a aplicação de TAIs;

[4] Essa possibilidade não é viável para os testes que necessitam divulgar os itens após a avaliação, como acontece, por exemplo, no atual ENEM.

- Definir as etapas necessárias para a implantação de um TAI, segundo as características específicas do teste;
- Identificar os mais diferentes métodos e critérios utilizados para a elaboração de um TAI, segundo as suas características e especificações;
- Essa aplicação será feita na avaliação teórica para a obtenção da carteira de habilitação de motorista realizada pelo Departamento de Trânsito do Estado de Santa Catarina – DETRAN-SC. No Brasil, algumas avaliações já adaptaram as suas provas ao sistema informatizado, onde as mesmas são realizadas via computador. Entretanto, como mencionado na Seção 1.1, esse tipo de aplicação não se trata de um teste adaptativo, já que é apenas a reprodução da prova "papel e lápis" no computador.

1.3. JUSTIFICATIVA

A criação de uma sistemática para a implantação de um TAI servirá como um referencial para todos aqueles que quiserem desenvolver e implantar um TAI, já que estabelecerá os passos necessários para a implantação do teste, assim como a direção a ser seguida em relação aos métodos e critérios utilizados segundo as características específicas dos teste, tais como: tipo de teste, objetivo do teste, tamanho e qualidade do banco de itens, modelo de resposta ao item utilizado, público alvo para a aplicação do teste, etc. Essa tese pretende servir como um manual teórico, principalmente ao usuário leigo que deseja desenvolver e implantar um TAI, orientando-o quanto à criação das etapas do teste, salientando os cuidados necessários na elaboração do mesmo, e indicando os métodos e critérios adequados para o teste a ser construído, de acordo com as características e especificações do mesmo. Além de estabelecer um método, essa sistemática pretende servir como um guia de orientação ao usuário para as etapas do desenvolvimento e elaboração de um Teste Adaptativo Informatizado. Em decorrência disso, serão obtidas direta ou indiretamente soluções para os problemas detalhados anteriormente na Seção 1.1.

Um Teste Adaptativo Informatizado realizado sob o suporte da Teoria da Resposta ao Item irá reunir as vantagens dessas duas ferramentas, produzindo um teste com uma qualidade muito superior aos testes convencionais. Em países da Europa e nos Estados Unidos, os TAIs vêm sendo realizados com êxito. No Brasil, ainda são raras as aplicações de TAIs, como mencionado na Seção 1.1.

1.3.1. Relevância

Embora exista um algoritmo básico para o TAI, disponibilizado em diversas literaturas, por exemplo, em Costa (2009), esse não fornece as diretrizes para a elaboração de um TAI. Essa tese pretende se tornar um referencial teórico para o usuário que deseja desenvolver um TAI, orientando o mesmo quanto à criação das etapas do teste, salientando os cuidados necessários na elaboração do mesmo, e providenciando os métodos e critérios adequados para o teste construído.

Esse trabalho procura mostrar a potencialidade da utilização dos TAIs no Brasil, que podem ser aplicados em diversos ramos onde existe a necessidade ou possibilidade de realizar avaliações, tais como, avaliações educacionais e de proficiência, avaliações psicológicas ou médicas, pesquisas de satisfação e de preferências, indicadores de qualidade ou desempenho, etc. Vários sistemas de avaliações atualmente existentes no país são potenciais candidatos para a aplicação de um futuro TAI. O desenvolvimento e aplicação de um TAI pode trazer vários benefícios para o sistema de avaliação.

Nesse trabalho, são apresentadas as vantagens e desvantagens da utilização dos TAIs. O leitor poderá observar que o ganho relacionado às vantagens supera as desvantagens, na maioria dos casos. Muitos programas de avaliações internacionais em larga escala, como os mencionados na Seção 1.1, utilizam com sucesso o TAI, já faz algumas décadas. No Brasil, assim como a TRI, que começou a ser utilizada tardiamente em avaliações de larga escala (comparando com as avaliações internacionais), porém com sucesso, os TAIs estão começando a ser utilizados. A tendência é que no futuro, essas e outras avaliações nacionais sejam feitas baseadas num TAI.

1.3.2. Ineditismo

Existe uma extensa literatura sobre TAIs publicada principalmente na Europa e nos EUA. Os trabalhos são desenvolvidos em estudos teóricos, simulações, aplicações práticas, desenvolvimento e comparações de métodos. Normalmente, os trabalhos são voltados para um aspecto particular do TAI, como o estudo de algum método específico ou o desenvolvimento de um TAI específico, o qual possui características particulares. Geralmente, apenas o algoritmo básico do TAI é mencionado. Na literatura estrangeira, existem poucos guias para a elaboração de testes adaptativos (CHALHOUB-DEVILLE, 2000; GREEN et al., 1984; MILLS; STOCKING, 1996; RENOM; DOVAL, 1999; THISSEN et al., 2007; THOMPSON; PROMETRIC, 2007;

THOMPSON; WEISS, 2011; WEISS, 2011; WISE; KINGSBURY, 2000). Entretanto, muitos são específicos a um determinado tipo de avaliação, restritos a algumas características, ou incompletos, ou voltados para a parte mais técnica do teste ou abordam métodos antigos. O desafio desse trabalho é construir uma sistemática teórica para a implantação de um TAI, por meio da utilização dos principais estudos e dos métodos atualmente utilizados. A intenção desse trabalho é elaborar um método, constituído por diretrizes, que sistematiza etapas para a elaboração de um TAI qualquer, independente das características particulares de cada um, onde o pesquisador poderá utilizá-lo como um guia para o desenvolvimento do TAI que deseja elaborar. O ineditismo desse trabalho é justificado pela definição dessa sistemática.

1.4. DELIMITAÇÕES

Comparada com outras técnicas estatísticas, a TRI pode ser considerada relativamente um técnica nova. Entretanto, já existe bastante literatura disponível sobre a TRI, principalmente em língua inglesa, e por esse motivo, não serão apresentados muitos detalhes sobre essa teoria, sendo suficiente as referências para consultas que serão mencionadas.

São diversos os modelos matemáticos da TRI que podem ser utilizados na elaboração de um TAI, segundo a abordagem do teste, a natureza do traço latente e a forma como os itens são elaborados. Entretanto, esse trabalho delimita-se a apresentar com mais detalhes apenas o modelo logístico unidimensional de três parâmetros (ML3), o qual foi utilizado na aplicação desse estudo, já que esse se refere a um traço latente unidimensional com itens de múltipla escolha que serão dicotomizados em “certo” ou “errado”, justificando a escolha do modelo.

A elaboração e análise dos itens é uma etapa fundamental para a construção do banco de itens a ser utilizado num TAI. Entretanto, como nesse trabalho foi analisado um banco de itens existente, as etapas relacionadas com a elaboração dos itens não serão apresentadas. A criação de itens para o uso da TRI é realizada através dos mesmos procedimentos teóricos da TCT o que pode ser consultado em diversas literaturas (ANASTASI, 1977; FAYERS; MACHIN, 2007; GARRET, 1979; GULLINKSEN, 1950; MIGUEL, 1974; OSTERLIND, 1997; PASQUALI, 1996; 1998; PRIETO; DELGADO, 1996; WILSON, 2005; etc.).

Esse trabalho também delimita-se à calibração do banco de itens, à elaboração do algoritmo do TAI e a sua validação através de

simulações. Embora conste na sistemática a etapa de implantação do teste, o mesmo não foi efetivamente implantado nesse momento devido à falta de recursos financeiros, físicos (computadores, provedores) e humanos (programadores, *designers*) e de tempo.

1.5. ESTRUTURA DO TRABALHO

Esse trabalho está estruturado em seis capítulos.

O primeiro capítulo apresenta uma introdução ao tema proposto para a tese, a contextualização, o problema envolvido, o objetivo geral e os objetivos específicos, a justificativa, a relevância e o ineditismo do projeto, as limitações e a estrutura do trabalho.

O segundo capitulo apresenta, de forma resumida, uma introdução sobre a Teoria da Resposta ao Item (TRI), sua conceituação, definições e características básicas, uma seção exclusiva apresentando as características do Modelo Logístico Unidimensional de Três Parâmetros (ML3), a análise dos itens e do teste, os métodos de estimação dos itens e da habilidade e a construção da escala de habilidade.

O terceiro capítulo apresenta os Testes Adaptativos Informatizados (TAIs) por meio de uma ampla revisão bibliográfica. Essa parte abrange um breve histórico, a conceituação e as definições, as características básicas, o banco de itens, os principais métodos utilizados na seleção dos itens, na estimação das habilidades, no controle de exposição dos itens e no balanceamento de conteúdo do teste e os critérios de parada.

No quarto capítulo é apresentada a metodologia utilizada no desenvolvimento desse trabalho, a caracterização do tipo de pesquisa utilizado e o resultado principal dessa tese, que é a Sistemática para a Implantação de Testes Adaptativos Informatizados, denominada SITAI, constituída por oito etapas que foram detalhadas e explicadas.

No quinto capítulo é apresentado um estudo de caso através do desenvolvimento de um Teste Adaptativo Informatizado da avaliação teórica para a obtenção da carteira de habilitação do Departamento de Trânsito do Estado de Santa Catarina – DETRAN-SC, elaborado a partir da sistemática proposta no quarto capítulo. Esse estudo envolve a calibração do banco de itens e a elaboração do algoritmo do TAI com a definição dos métodos de seleção dos itens, das restrições, dos métodos e estimação do traço latente e do critério de parada do teste. A validação do teste é verificada por meio de critérios adequados com o uso de simulações. Não foram desenvolvidas as etapas da SITAI relacionadas

com a elaboração dos itens, e a implementação, aplicação e manutenção do teste, porém algumas considerações foram feitas.

No sexto capítulo são apresentadas as considerações finais, as conclusões, as dificuldades encontradas e as propostas para os trabalhos futuros.

No final, são relacionadas as referências utilizadas para a elaboração desse trabalho e os apêndices.

2. TEORIA DA RESPOSTA AO ITEM - TRI

A Teoria da Resposta ao Item (TRI) é uma metodologia que sugere formas de representar a relação entre a probabilidade de um indivíduo dar uma certa resposta a um item, os traços latentes do indivíduo e as características dos itens, por meio de modelos matemáticos (ANDRADE; TAVARES; VALLE, 2000). Traços latentes são características do indivíduo que não podem ser medidas diretamente, como, por exemplo, proficiência, grau de satisfação ou nível de ansiedade. Não existe um aparelho capaz de medir características desse tipo. Dessa forma, essas características são mensuradas através de variáveis secundárias relacionadas com o traço latente em estudo.

Conforme Nunes e Primi (2005), a TRI não entra em contradição com os princípios da psicometria clássica e traz uma nova proposta estatística, tendo como unidade de análise o item, formalizando a relação que existe entre a probabilidade de acertar o item e a capacidade latente requerida na sua resolução. Segundo Pasquali (1996), a TRI surge para substituir grande parte da Teoria Clássica dos Testes (TCT), também chamada de Teoria Clássica da Medida (TCM), ao fato de ela superar certas limitações teóricas graves que a psicometria tradicional contém. Mais detalhes sobre a TCT podem ser encontrados, por exemplo, em Anastasi (1977), Crocker e Algina (1986), Cronbach (1996), Garret (1979), Guilford (1954), Gulliksen (1950), Lord e Novick (1968), Miguel (1974), Nick (1963), Nunnally (1970), Osterlind (1997), Pasquali (1996, 1998), Prieto e Delgado (1996), Vianna (1987) e Wilson (2005).

Comparações entre a TRI e a TCT podem ser encontradas em várias literaturas (BORTOLOTTI, 2003; 2010; EMBRETSON; REISE, 2000; HARVEY; HAMMER, 1999). Muitas são as vantagens da utilização da TRI em relação à TCT. Entre elas, destacam-se: (1) a TRI fornece informações mais precisas do desempenho dos respondentes já que o traço latente do indivíduo não depende da dificuldade das questões do teste, enquanto que na TCT o escore do indivíduo depende essencialmente dos itens que compõe o teste (ANDRADE; TAVARES; VALLE, 2000; VENDRAMINI; SILVA; CANALE, 2004); (2) a TRI permite obter índices de precisão do item (função de informação do item - FII) e do teste (função de informação do teste - FIT) mais ricos do que os índices utilizados pela TCT (ANDRADE; TAVARES; VALLE, 2000; BACKER, 2001); (3) a TRI permite utilizar modelos que consideram a possibilidade de acerto casual (ANDRADE; TAVARES; VALLE, 2000); (4) a TRI permite a comparação através do escore entre

os indivíduos que responderam questionários com itens diferentes para medir o mesmo traço latente, pois os itens e os indivíduos são colocados numa mesma escala (ANDRADE; TAVARES; VALLE, 2000; EMBRETSON; REISE, 2000); (5) na TRI, uma vez estimada a proficiência do indivíduo, é possível verificar qual a probabilidade de acertar um determinado item que ele não respondeu (VENDRAMINI; SILVA; CANALE, 2004); (6) na TRI, cada respondente tem seu próprio erro padrão, relacionado à sua habilidade, onde a estimação desse erro é mais precisa (EMBRETSON; REISE, 2000); (7) na TRI, testes curtos podem ser mais confiáveis que os testes longos (EMBRETSON; REISE, 2000); (8) a TRI permite a utilização de formatos mesclados de itens (por exemplo, dicotômicos e politômicos nominais e graduais) sem causar um impacto desequilibrado nos escores total do teste (EMBRETSON; REISE, 2000).

Não obstante, a TRI apresenta algumas desvantagens. Entre elas, destacam-se: (1) dificuldades encontradas com a aplicabilidade da sua própria natureza, tanto do ponto de vista teórico, devido a problemas de difícil solução no campo da estimação, como do ponto de vista computacional (ANDRADE; TAVARES; VALLE, 2000); (2) as análises da TRI geralmente exigem um software especial e em geral diversos programas diferentes são necessários para executar os testes das suposições e estimação dos parâmetros, bem como os testes para ajuste do modelo (SCHERBAUM; FINLINSON; TAMANINI, 2006); (3) em alguns casos, as análises da TCT do item poderão fornecer toda a informação que é necessária e a utilização da TRI pode não contribuir em nada para a análise, sendo, assim, desnecessária (BORTOLOTTI, 2010); (4) em muitos casos existe a necessidade de uma amostra bastante grande, dependendo da quantidade de parâmetros que serão estimados (ANDRADE; TAVARES; VALLE, 2000).

Existem vários modelos matemáticos utilizados na TRI, diferentes quanto à sua função e à quantidade de parâmetros, e cada um deles é específico para um tipo de situação. Esses modelos podem ser classificados quanto à sua dimensão (unidimensionais ou multidimensionais), quanto ao tipo de traço latente (cumulativo ou não cumulativo), quanto ao tipo de item (dicotômico ou politômico) e quanto ao número de populações envolvidas.

Os modelos unidimensionais são utilizados quando se supõe que o objeto de estudo é composto por um único traço latente que está relacionado com a capacidade ou habilidade do indivíduo em responder aos itens do teste, ou quando o traço latente pode ser representado por uma única dimensão ou fator. Os modelos unidimensionais podem ser

encontrados na maioria das literaturas sobre TRI, por exemplo, em Andrade, Tavares e Valle (2000), de Ayala (2008), de Boeck e Wilson (2004), Boomsma, van Duijn e Snijders (2001), Embretson e Reise (2000), Lord e Novick (1968), Muñiz (1997), Ostini e Nering (2006), Reckase (2009), Revuelta, Ponsoda, e Abad (2006), Thissen e Wainer (2001) e van der Linden e Hambleton (1997).

Os modelos multidimensionais são adequados quando se estuda mais de um traço latente ou quando o traço latente não pode ser representado por uma única dimensão. Para os modelos multidimensionais, recomenda-se de Ayala (2008), Nojosa (2002), Reckase (2009), Silva (2005) e van der Linden e Hambleton (1997).

Os modelos cumulativos devem ser utilizados quando o traço latente analisado possui característica acumulativa. Isso ocorre quando um indivíduo, com um determinado traço latente (proficiência, habilidade, atitude, satisfação, etc.), domina (ou concorda ou possui – depende do que está sendo medido) os itens que estão situados abaixo dele na escala e não domina (ou não concorda ou não possui) os que estão situados acima dele. Segundo Arias e Rivas (1991), a característica acumulativa é uma relação de dominância entre itens e sujeitos: se um indivíduo domina certo item, consequentemente ele domina também todos os itens que estão posicionados abaixo desse item na escala. Na avaliação educacional, por exemplo, quanto maior o traço latente, mais conhecimento tem o aluno e mais itens ele domina. No nível de satisfação, quanto maior o traço latente, maior é a satisfação do indivíduo e com mais itens ele vai concordar. Os modelos cumulativos abrangem a classe dos modelos logísticos, os modelos de resposta nominal, de resposta gradual e de crédito parcial generalizado, entre outros.

Por outro lado, entre os modelos não cumulativos, destacam-se os modelos de desdobramento são adequados quando o traço latente não possui característica cumulativa, ou seja, as diversas manifestações do traço latente estão distribuídas ao longo da escala. Segundo Bortolotti e Andrade (2007), são situações em que nem sempre a probabilidade de responder as categorias de resposta mais alta aumenta quando aumenta o traço latente do indivíduo. Dessa forma, o indivíduo posicionado na escala possuirá maior probabilidade de ter aquela característica que está posicionada mais próxima dele nessa escala. É o caso, por exemplo, de atitudes e comportamentos. Entre os modelos de desdobramento, destacam-se o modelo de desdobramento qualitativo e o modelo de desdobramento generalizado graduado, entre outros. Para os modelos de

desdobramento, recomenda-se Bortolotti (2003; 2010) e Bortolotti e Andrade (2007).

Os modelos dicotômicos são utilizados quando os itens são de natureza dicotômica ou quando eles podem ser dicotomizados ou avaliados de forma dicotomizada, enquanto que os modelos politômicos são utilizados quando os itens possuem mais de duas alternativas de respostas. Os modelos politômicos possuem mais parâmetros para serem estimados do que os dicotômicos, conforme a quantidade de alternativas dos itens. Dessa forma, para utilizar um modelo politômico, é necessário ter número maior de respondentes do que para o modelo dicotômico. Quando a amostra não é grande o suficiente para a análise com os modelos politômicos, uma alternativa é dicotomizar os itens e utilizar os modelos dicotômicos. Os modelos dicotômicos são representados pelos modelos logísticos (ou pela ogiva normal), enquanto que os modelos politômicos abrangem uma classe maior de modelos: os modelos de resposta nominal, de resposta gradual e de crédito parcial generalizado, modelos de desdobramento, entre outros (ANDRADE; TAVARES; VALLE, 2000).

A definição da população depende dos objetivos do estudo. Populações diferentes podem ter que responder a itens diferentes que medem o mesmo traço latente. Por exemplo, alunos de quarta série não podem responder itens que são adequados para alunos de oitava série. Embora a TRI permita que esses dois grupos possam ser comparados, eles possuem níveis de proficiência diferentes que exigem itens diferentes, constituído, assim, duas populações diferentes. Porém, em algumas situações, pode ocorrer que uma única população em um estudo pode ser considerada como mais de uma em outro. Por exemplo, um estudo pode considerar os alunos de terceiro ano do Ensino Médio de Porto Alegre como sua população, porém outro estudo pode considerar esses mesmos alunos como duas populações, sendo uma referente ao turno diurno e outra referente ao turno noturno. Sobre situações com duas ou mais populações, sugere-se Bock e Zimowski (1997) e Andrade, Tavares e Valle (2000).

A TRI foi utilizada pela primeira vez no Brasil em 1995 no SAEB, para montagem de instrumentos, tratamento de dados e construção de escalas a partir de resultados apresentados por alunos em provas de rendimento (SOUZA, 2005), o que permitiu que alunos de diferentes séries pudessem ser comparados e colocados na mesma escala, que não era possível fazer através da TCT. A maioria das aplicações tem sido na avaliação educacional (MOREIRA JUNIOR, 2010), onde ENEM (Exame Nacional do Ensino Médio) tem sido,

atualmente, o maior exemplo que mostra os benefícios da TRI. Nos últimos anos, a TRI tem sido aplicada em diversas áreas no Brasil, inclusive dentro do contexto da Engenharia de Produção, como por exemplo, na Gestão da Qualidade Total (ALEXANDRE et al., 2001; 2002a; 2002b; 2003a; 2003b; BATISTA; ALEXANDRE, 2004; BATISTA et al., 2002; VASCONCELOS et al., 2002), na avaliação de intangíveis nas organizações (VARGAS, 2007; VARGAS et al., 2008; MOREIRA JUNIOR; VARGAS; ANDRADE, 2010), na avaliação do nível de satisfação (BORNIA et al., 2009; BORTOLOTTI, 2003; BORTOLOTTI; ANDRADE, 2007; BORTOLOTTI; MOREIRA JUNIOR; SOUZA JUNIOR, 2010a; 2010b; BORTOLOTTI; SOUZA JUNIOR; ANDRADE, 2009; COSTA; CHAVES NETO, 2002; CUNHA; SENA JUNIOR; MATOS, 2002; MOREIRA JUNIOR et al., 2010), na gestão organizacional (MORAIS, 2009; SANTOS et al., 2009,), usabilidade de sites (MOREIRA JUNIOR TEZZA; ANDRADE, 2010; TEZZA, 2009; TEZZA; BORNIA, 2009; TEZZA; BORNIA; MOREIRA JUNIOR, 2009; TEZZA; BORNIA; ANDRADE, 2011), na gestão do conhecimento (ALMEIDA, 2009), na avaliação da maturidade empresarial (CASTRO JUNIOR, 2007; PEREIRA, 2007; RORIZ JUNIOR, 2008), na ergonomia (VERGARA, 2005), na avaliação da resistência à mudança (BORTOLOTTI, 2010), etc.

No exterior, a TRI vem sendo aplicada há muito mais tempo e existe uma quantidade muito grande de publicações. Algumas aplicações famosas em larga escala que utilizam a TRI são o *PISA – Programme for International Student Assessment* (OECD, 2005) e o *TOEFL – Test of English as a Foreign Language* (YAMAMOTO, 1995).

2.1. CONCEITOS BÁSICOS

Andrade e Valle (1998) destacam que dois pressupostos teóricos fundamentais para a utilização dos modelos unidimensionais devem ser verificados: independência local e unidimensionalidade.

A independência local assume que, para uma dada habilidade, as respostas aos diferentes itens da prova são independentes. Isso significa que a resposta do indivíduo dada para um item qualquer não deve ser influenciada pelas respostas fornecidas a outros itens, ou seja, os itens devem ser independentes. Assim, os parâmetros de cada item não dependem dos outros itens do teste, e a pontuação do teste se faz em função das respostas do sujeito a cada item. Dessa forma, é possível verificar se os respondentes são mais ou menos hábeis, e da mesma forma, se os itens podem ser considerados mais fáceis ou mais difíceis,

já que os itens e os respondentes são posicionados na mesma escala de desempenho. Andrade, Tavares e Valle (2000) ressaltam que a unidimensionalidade implica independência local (a recíproca não é verdadeira), portanto basta somente verificar a suposição da unidimensionalidade. Não somente a unidimensionalidade, mas a dimensionalidade correta também implicará na independência local.

A unidimensionalidade pressupõe que os itens de um teste devem medir uma única dimensão (um único traço latente) ou, pelo menos, deve haver um fator dominante, embora Soares (2005) considere esse critério subjetivo. Andrade, Tavares e Valle (2000) admitem que provavelmente qualquer desempenho humano é sempre multideterminado ou multimotivado, dado que mais de um traço latente entra na execução de qualquer tarefa. Contudo, para satisfazer o postulado da unidimensionalidade, é suficiente verificar que existe uma habilidade dominante responsável pelo conjunto de itens, que é o que se supõe estar sendo medido pelo teste. A quebra das suposições necessárias em relação à dimensionalidade e independência local pode ter como efeito resultados totalmente inválidos (MISLEVY; CHANG, 2000).

Conforme Pasquali (1998), a dimensionalidade de um atributo refere-se à sua estrutura interna. O objetivo é definir quantas dimensões estão sendo analisadas pelo construto a ser construído. Para isso, são discutidas questões do tipo:

- O construto mede um ou mais traços latentes?
- O construto mede um traço latente com uma ou mais dimensões?
- Essas dimensões podem ser analisadas separadamente do construto, cada uma de forma unidimensional, ou não podem ser separadas?
- As características do traço latente pesquisado permitem uma composição homogênea (em um único instrumento de coleta) ou é necessário subdividi-las (em vários instrumentos de coleta)?

Essas questões devem ser respondidas pela literatura e pelas análises dos dados empíricos existentes, que podem ou não ser confirmadas através da verificação da dimensionalidade durante os procedimentos estatísticos. Os modelos da TRI mais utilizados assumem que os dados são unidimensionais, mas na prática é difícil conseguir uma unidimensionalidade absoluta, já que há muitos fatores não controlados que afetam as respostas dos sujeitos (PASQUALI, 1998).

A questão da dimensionalidade, muitas vezes omitida, é um fato importante na TRI, porque o uso adequado do modelo depende da dimensão. Silva (2005) adverte que, quando, supostamente, o traço

latente é multidimensional e é utilizado um modelo unidimensional, os erros de medida aumentam e podem ser retiradas conclusões incorretas sobre a habilidade estimada.

Normalmente, a dimensionalidade do teste é verificada através da técnica estatística multivariada conhecida como análise fatorial. Para que seja constatada uma dimensão, o fator dominante deve explicar mais de 20% da variância total (RECKASE, 1979). Esse percentual pode ser maior se, depois de uma primeira análise, forem eliminados os itens com cargas fatoriais baixas no fator dominante (LUMSDEN, 1976). Entretanto, para verificar a dimensionalidade, é necessário que se tenha dados, ou seja, o teste já deve ter sido executado, pelo menos, experimentalmente. Essa aplicação pode ser realizada através de uma amostra piloto.

Existem diferentes formas de realizar a análise fatorial para os diferentes tipos de dados. Quando os dados são quantitativos, o método mais comum é a análise fatorial clássica exploratória ou confirmatória (HAIR et al., 2009; MINGOTI, 2007), disponível na maioria dos softwares estatísticos. Quando os dados são qualitativos ordinais ou dicotômicos, os procedimentos mais comuns são: a Análise Fatorial de Informação Completa (NOJOSA, 2002), do inglês, *Full Information Factor Analysis (FIFA)*, a Análise Fatorial baseada em correlações tetracóricas (LORD, 1980) e a Análise Fatorial Booleana (HAIR et al., 2009, p. 108). Nesses casos, é necessária a utilização de softwares específicos, como o TESTFACT (WILSON et al., 1991; BOCK et al., 2003), o LISREL (JÖRESKOG; SÖRBOM, 1996a), o MicroFACT (FINGER, 2004) o PRELIS, (JÖRESKOG; SÖRBOM, 1996b), ou o BMDP (BMDP, 1992). Outros procedimentos para verificar a unidimensionalidade têm sido propostos (CUESTA, 1996; FERRANDO, 1996; REVUELTA; PONSODA, 1998b; REISE; MORIZOT; HAYS, 2007; YU et al., 2007). Uma listagem de softwares para testar a dimensionalidade encontra-se em Deng e Hambleton (2007).

Para maiores detalhes nas análises sobre dimensionalidade, recomenda-se: de Ayala (1992); Fayers e Machin (2007), Thissen e Wainer (2001), Hair et al. (2009), Corrar et al. (2009), Mingoti (2007), Joaristi e Lizasoain (2008), Vitória, Almeida e Primi (2006) e Richaud (2005).

Os modelos de resposta ao item só podem ser considerados vantajosos quando o ajuste do modelo aos dados de interesse for satisfatório, entretanto o ajuste do modelo não é suficiente para verificar a dimensionalidade (WISE; KINGSBURY, 2000). Um modelo mal-

ajustado não fornecerá parâmetros invariantes para os itens e para as habilidades. Tradicionalmente, o teste utilizado para verificar a adequação do modelo tem sido o teste Qui-Quadrado de ajuste (*Chi-square Goodness-of-fit*) (BAKER, 2001), entretanto esse teste não se mostra adequado quando a amostra é grande. Porém, mais recentemente, outros métodos têm sido propostos (GLAS; DAGOHOY, 2007; PINA; MONTESINOS, 2005) inclusive dentro do contexto dos TAIs (BERGSTROM, 1990; GLAS, 2000; 2010; NERING, 1997; REISE; DUE, 1991; VAN KRIMPEN-STOOP; MEIJER, 1999; 2000; 2001; MEIJER; VAN KRIMPEN-STOOP, 2010).

Entre os modelos utilizados na TRI, destacam-se: os modelos logísticos[5] com 1 parâmetro – ML1 (WRIGHT, 1968), 2 parâmetros – ML2 (LORD, 1980; BIRNBAUM, 1968), 3 parâmetros – ML3 (LORD, 1980; BIRNBAUM, 1968) e 4 parâmetros – ML4 (WANG; HANSON, 2001), o modelo de resposta nominal – MRN (BOCK, 1972), modelo de resposta gradual – MRG (SAMEJIMA, 1969), o modelo de escala gradual – MEG (ANDRICH, 1978), o modelo de crédito parcial (MASTERS, 1982), o modelo de crédito parcial generalizado (MURAKI, 1992), o modelo de desdobramento graduado generalizado – GGUM (ROBERTS; DONOGHUE; LAUGHLIN, 2000), modelos normais assimétricos – TRI-NA (BAZÁN, 2005), modelos logísticos multidimensionais (SILVA, 2005), modelos não paramétricos (MOKKEN, 1971; SIJTSMA; MOLENAAR, 2002), o modelo cosseno hiperbólico de desdobramento (ANDRICH; LUO, 1993), o modelo Parella (HOIJTINK, 1990), o modelo seqüencial (TUTZ, 1990), modelos logísticos para mais de uma população (BOCK; ZIMOWSKI, 1997), etc. Para a exploração desses e outros modelos, recomenda-se Boomsma, van Duijn e Snijders (2001), de Ayala (2008), de Boeck e Wilson (2004), Embretson e Reise (2000), Lord (1980), Muñiz (1997), Ostini e Nering (2006), Reckase (2009), Revuelta, Ponsoda e Abad (2006), Thissen e Wainer (2001) e van der Linden e Hambleton (1997).

Os modelos logísticos unidimensionais mais conhecidos e utilizados na TRI são:

1) Modelo Logístico de Um Parâmetro (ML1) ou o Modelo de Rasch (1960): supõe que todos os itens possuem o mesmo nível de discriminação, porém dificuldades diferentes, e que não há possibilidade de acerto casual. Rasch desenvolveu seu modelo

[5] Estes modelos foram primeiramente desenvolvidos na forma de uma função ogiva normal e, depois, foram descritos para uma forma matemática mais conveniente, e que vem sendo usada até então: a logística (ANDRADE; TAVARES; VALLE, 2000).

utilizando a ogiva normal, porém Wright (1968) o descreveu na métrica logística.

2) Modelo Logístico de Dois Parâmetros (ML2): proposto inicialmente por Birnbaum (1968), na métrica logística, avalia a dificuldade e a discriminação dos itens, sem a possibilidade de acerto casual. Lord (1980) o descreveu na métrica normal.
3) Modelo Logístico de Três Parâmetros (ML3): proposto inicialmente por Birnbaum (1980), na métrica logística, avalia a dificuldade e a discriminação dos itens, e o acerto casual. Lord (1968) o descreveu na métrica normal. O ML1 e o ML2 são situações particulares do ML3, como será visto na próxima seção.

2.2. O MODELO LOGÍSTICO DE TRÊS PARÂMETROS

O modelo logístico unidimensional de 3 parâmetros (ML3) é um dos mais utilizados na TRI (ANDRADE; TAVARES; VALLE, 2000; WAINER, 2000a). O ML3 considera a dificuldade e a discriminação do item e a probabilidade de acerto casual, e é dado por:

$$P(U_{ij}=1 \mid \theta_j) = c_i + (1-c_i)\frac{1}{1+e^{-a_i(\theta_j - b_i)}}, \qquad (1)$$

para $i = 1,\ 2,\ ...,\ I$, e $j = 1,\ 2,\ ...,\ n$, onde

U_{ij} é uma variável dicotômica (assume o valor 1 quando o indivíduo *j* responde corretamente o item *i*, ou assume o valor 0, caso contrário);

θ_j é o traço latente (parâmetro da habilidade) do indivíduo *j*;

$P(U_{ij}=1 \mid \theta_j)$, também chamada de Função de Resposta do Item (FRI), é a probabilidade do indivíduo *j* responder corretamente o item *i*, dado que ele tem habilidade θ_j, ou seja, é a proporção de respostas corretas do item *i* dos indivíduos da população com habilidade θ_j;

a_i é o parâmetro de discriminação (ou de inclinação) do item *i*;

b_i é o parâmetro de dificuldade (ou de posição) do item *i*, medido na mesma escala da habilidade;

c_i é o parâmetro de acerto casual, que representa a probabilidade de indivíduos com baixa habilidade responderem corretamente o item *i* (ou parâmetro de acerto casual);

e é a conhecida constante matemática igual a 2,718281...;

I é o número total de itens; e

n é a quantidade total de indivíduos na amostra.

A Figura 1 apresenta um exemplo de uma Curva Característica do Item (CCI) de um ML3 e a sua relação existente com os parâmetros dos itens a_i (inclinação da curva), b_i (posição do item na escala) e c_i (probabilidade de acerto casual de indivíduos com baixa habilidade). A CCI é o gráfico da função do modelo matemático, onde o eixo Y é a probabilidade de resposta correta de um indivíduo segundo a sua habilidade (eixo X). No exemplo da Figura 1, foram utilizados *a* = 1,2, *b* = 1 e *c* = 0,2, considerando uma escala (0, 1), ou seja, com média igual a zero e desvio padrão igual a 1. Maiores detalhes sobre a construção e interpretação da escala são discutidos na Seção 2.6.

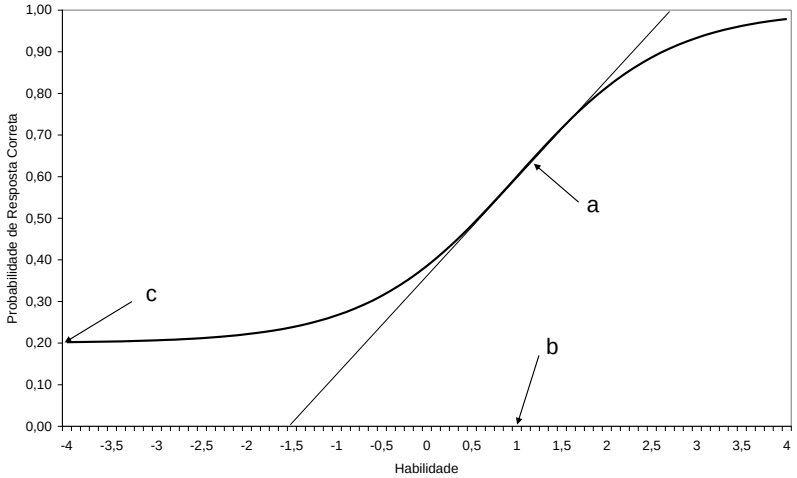


Figura 1. Relação entre os parâmetros dos itens e a CCI

O traço latente (habilidade, proficiência, atitude, preferência, satisfação, etc.) do indivíduo (θ_j) é medido em uma escala arbitrária que varia teoricamente entre $-\infty$ e $+\infty$. Porém, o importante nessa escala não é a sua magnitude, mas as relações de ordem existentes (ANDRADE; TAVARES; VALLE, 2000). O traço latente, no modelo acumulativo, é especificado como um tipo de característica que apresenta uma probabilidade maior para indivíduos com θ_j maior, e uma probabilidade menor para indivíduos com θ_j menor. Ou seja,

quanto maior for θ_j, maior será a probabilidade do indivíduo *j* acertar o item.

Segundo Baker (2001), o traço latente θ_j do indivíduo *j* é invariante em relação aos itens utilizados para estimá-lo, desde que os itens sejam adequados, isto é, estejam calibrados, em uma métrica comum e medindo o mesmo traço latente (unidimensionalidade). Isso justifica o fato do resultado do θ_j ser o mesmo, independente dos itens que formam o questionário, o que não ocorre na TCT. Portanto, não importa se o teste é composto por itens difíceis ou fáceis, a estimativa da habilidade é a mesma. Isso é condizente com a realidade, já que a habilidade de um indivíduo, num determinado tempo *t*, é a mesma independente do grau de dificuldade do teste. Essa é a chamada propriedade de invariância do parâmetro de habilidade da TRI. Condé e Laros (2007) afirmam que essa propriedade é a maior distinção da TRI em relação à TCT.

Também deve-se tomar cuidado no caso de qualquer alteração no banco de itens (inclusão ou remoção de itens) se os parâmetros dos itens forem estimados novamente, pois isso irá alterar a escala e os valores dos θ_j. Geralmente, se a calibração dos itens no banco é confiável, faz-se a inclusão de itens sem alterar a escala.

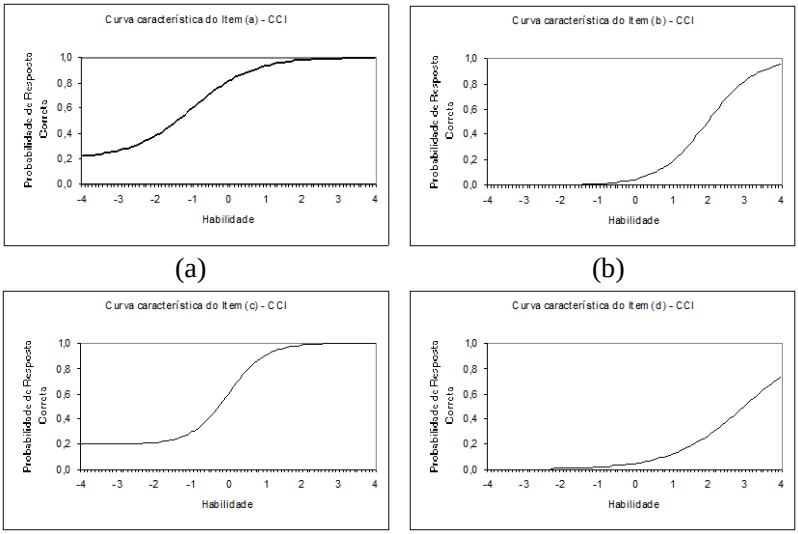
O parâmetro a_i é proporcional à inclinação da curva no ponto de inflexão, assim, valores negativos de a_i não são esperados nesse modelo. O parâmetro a_i mede a discriminação do item. Valores baixos de a_i indicam que o item tem pouco poder de discriminação, ou seja, a probabilidade de um indivíduo responder corretamente o item ou concordar com ele é aproximadamente a mesma para indivíduos com baixa ou alta proficiência. Por outro lado, valores altos de a_i indicam que o item tem grande poder de discriminação, dividindo os indivíduos praticamente em dois grupos: os que possuem habilidades abaixo do valor de b_i (com baixa probabilidade de acertar o item), e os que possuem habilidades acima do valor de b_i (com alta probabilidade de acertar o item). Não existe um valor exato de a_i para decidir se um item discrimina bem ou não. Em geral, na métrica logística, um item com a_i maior que 0,7 pode ser considerado aceitável, mas um valor maior ou

igual a 1,0 indica que o item discrimina bem. Valores extremamente altos de a_i também não são adequados, pois provavelmente dividiria os indivíduos em dois grupos distintos (os que têm θ_j maior que b_i e os que têm θ_j menor que b_i), mas não faria distinção entre os indivíduos dentro dos grupos.

O parâmetro b_i é o parâmetro de dificuldade ou proficiência do item, que é medido na mesma unidade da escala da habilidade do indivíduo (θ_j). Ele representa o grau de dificuldade do item, ou seja, quanto maior seu valor, mais difícil o item é (somente indivíduos com habilidade alta terão uma alta probabilidade de acertá-lo), e vice-versa. Esse valor de b_i é que vai definir a posição do item na escala, por isso ele também é chamado de parâmetro de localização ou posição. Teoricamente, b_i pode assumir qualquer valor entre $-\infty$ e $+\infty$, entretanto, para valores muito altos ou baixos, o item pode não ser adequado, sendo usual os valores entre -3 e 3, na escala (0, 1), com média igual a zero e desvio padrão igual a um.

O parâmetro c_i é a probabilidade de um indivíduo com baixa proficiência ou com pouco (ou nenhum) conhecimento, em relação ao assunto que está sendo avaliado, responder corretamente ao item *i*. O parâmetro c_i é considerado quando existe a possibilidade de acerto casual e o seu valor depende da quantidade de alternativas que o item apresenta e do parâmetro de dificuldade do item.

Os gráficos da Figura 2 apresentam 4 itens que são adequados para um teste representado na escala (0,1) onde a habilidade dos indivíduos segue uma Distribuição Normal Padrão.



(a) (b)

(c) (d)

Figura 2. Exemplos de itens adequados

Pode-se observar que o item (a) é fácil (b = -1) e discrimina bem (a = 1,2), o item (b) é mais difícil (b = 2) e discrimina melhor que o (a) (a = 1,5), o item (c) é mediano (b = 0) e discrimina melhor que os demais (a = 2), e o item (d) é o mais difícil (b = 3) mas discrimina menos que os demais (a =1). Observa-se também o parâmetro de acerto casual nos itens (a) e (c), onde c = 0,2. Nesses itens, percebe-se claramente a variação da probabilidade de resposta correta ao longo da escala do traço latente, pois são itens que discriminam bem e estão posicionados dentro do intervalo adequado na escala (0, 1).

Os gráficos da Figura 3 apresentam 4 itens que não são adequados para um teste que utiliza escala (0,1) onde a habilidade dos indivíduos segue uma Distribuição Normal Padrão.

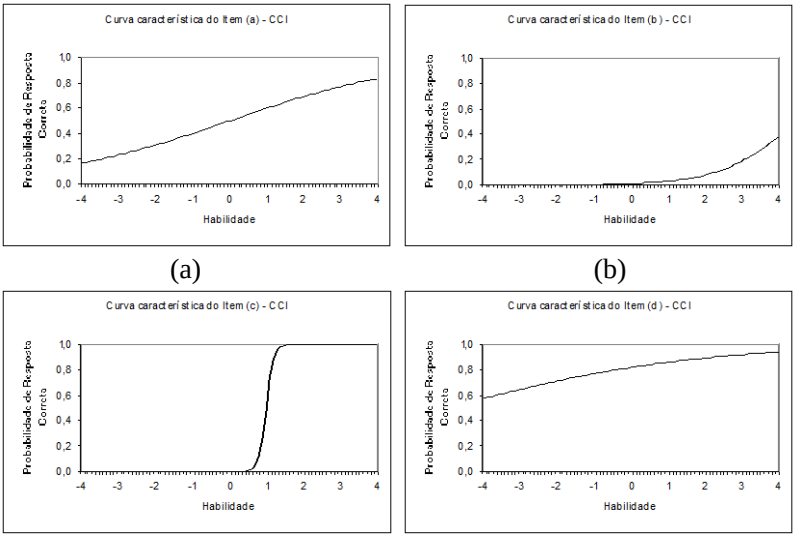


Figura 3. Exemplos de itens inadequados

Pode-se observar que o item (a) é de média dificuldade (b = 0), mas não discrimina bem (a = 0,4), parecendo uma reta quase horizontal, isto é, o poder de discriminação dele é muito pequeno, o que faz com que a probabilidade de acerto seja semelhante entre indivíduos com diferentes traços latentes. O item (b) discrimina bem (a = 1), mas é muito difícil (b = 4,5), onde quase todos os respondentes têm baixa probabilidade de acertar. O item (c) é pouco difícil (b = 1), mas não discrimina bem (a =10), parecendo uma reta quase vertical, pois o poder de discriminação praticamente divide o grupo em duas partes, mas não permite discriminação entre os indivíduos dentro dessas partes. O item (d) discrimina pouco (a = 0,3) e é muito fácil (b = -5), onde quase todos os respondentes têm alta probabilidade de acertar. Itens como esses não devem fazer parte do teste, pois eles podem prejudicar o desempenho do teste, resultando numa estimativa pouco precisa do traço latente.

2.2.1. Função de Informação do Item

A Função de Informação do Item (FII) permite analisar a quantidade de informação que um item fornece para a medida do traço latente analisado e reflete a qualidade do item. Segundo Andrade, Tavares e Valle (2000), a FII de um ML3 é dada por:

$$I_i(\theta) = a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{1 - c_i} \right]^2, \qquad (2)$$

onde

$I_i(\theta)$ é a informação fornecida pelo item *i* no nível de habilidade θ,

$P_i(\theta) = P\left(X_{ij} = 1 \mid \theta\right)$, e

$Q_i(\theta) = 1 - P_i(\theta)$.

Desenvolvendo-se as expressões da Equação 2, chega-se a seguinte expressão, conforme Francisco (2005):

$$I_i(\theta) = \frac{a_i^2 \left(1 - c_i\right)}{\left[c_i + e^{a_i(\theta - b_i)}\right] \left[1 + e^{-a_i(\theta - b_i)}\right]^2}. \qquad (3)$$

Quanto maior for a informação de um item, melhor será a sua qualidade. Quanto maior for o valor de a_i e menor for o valor de c_i, maior será a informação do item (curva será mais alta e estreita) e mais acentuada será a CCI. A Figura 4 apresenta diferentes itens com a CCI (linha contínua) e a sua curva de informação (linha pontilhada).

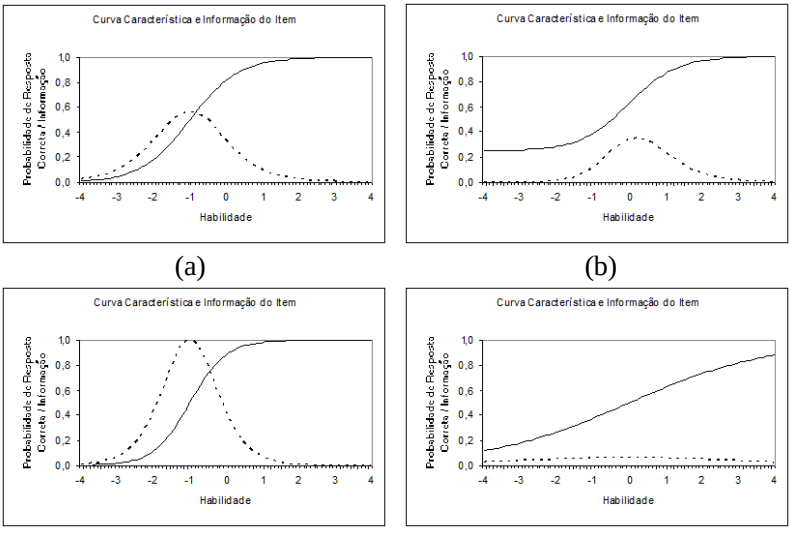


Figura 4. CCI e curva de informação dos itens

Observa-se na Figura 4 que o item (c) é o mais informativo (a = 2 e c = 0). Os itens (a) e (b) possuem a mesma discriminação (a = 1,5), entretanto nota-se o impacto na diminuição da informação em (b) devido a existência do parâmetro c (c = 0,25). Porém, dentre esses itens, o único que não é adequado é o (d), que praticamente não oferece informação nenhuma (a = 0,5 e c = 0). O valor máximo da curva de informação será obtido no ponto b_i do item, ou seja, quando $b_i = \theta$. Quando isso ocorre, a expressão da Equação 3 se reduz a:

$$I_i(\theta) = \frac{a_i^2(1-c_i)}{4(1+c_i)}. \qquad (4)$$

Isso significa que o valor máximo da FII depende apenas dos parâmetros a_i e c_i, ou seja, dois itens com os mesmos valores de a_i e c_i, independente do valor de b_i, apresentarão o mesmo valor máximo na FII. O gráfico da Figura 5 apresenta a influência dos parâmetros a_i e c_i, (para os valores de c_i = 0; 0,1; 0,2 e 0,25; e $0{,}5 \leq a \leq 3$) na FII do teste. Observa-se que quando o valor de a_i aumenta, o valor máximo da FII aumenta de forma quadrática. Também percebe-se que a medida que o valor de c_i aumenta, o valor máximo da FII diminui proporcionalmente.

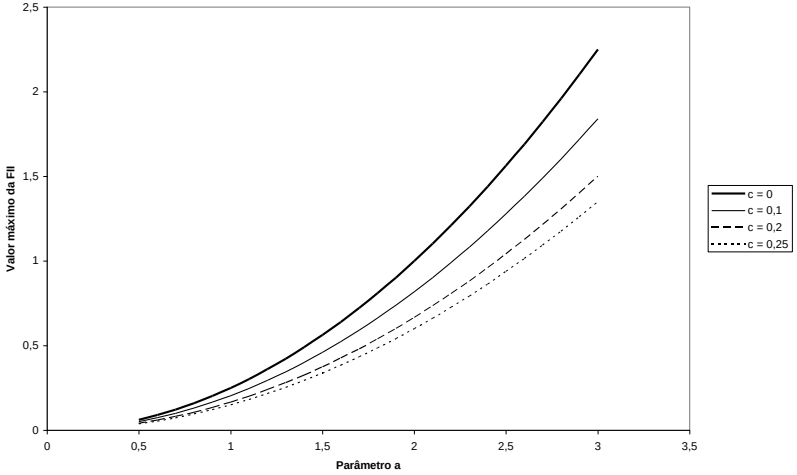


Figura 5. Relação entre o máximo da FII e os parâmetros a_i e c_i

A FII pode ser usada para identificar itens pobres ou pouco significativos para o teste. Segundo Reeve e Fayers (2005), um item que tem um grau de informação baixo, pode significar que o item: mede algo diferente em relação aos outros itens na escala (outra dimensão); ou não está bem elaborado e precisa ser reescrito; ou é muito complexo para os indivíduos; ou está fora do contexto do teste.

2.2.2. Função de Informação do Teste

Segundo Andrade, Tavares e Valle (2000), a Função de Informação do Teste (FIT) é simplesmente a soma das informações fornecidas por cada item que compõe o teste:

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta) \qquad (5)$$

A FIT está inversamente relacionada com o erro-padrão de medida (EP) que é o erro padrão do estimador de θ, ou seja:

$$EP(\theta) = \frac{1}{\sqrt{I(\theta)}} \qquad (6)$$

Observa-se uma relação inversa entre a FIT e o erro padrão de estimação: quanto maior for a FIT, menor será o erro padrão de estimação e maior será a precisão da estimação da habilidade. Quanto maior o erro padrão de estimativa, menor a precisão com que é estimado θ. Segundo Hambleton, Swaminathan e Rogers (1991) a magnitude do EP(θ) depende, em geral:

- Do número de itens no teste (EP baixos estão associados com testes maiores);
- A qualidade dos itens no teste (EP baixos estão associados com discriminação alta dos itens); e
- Itens com dificuldade adequada à habilidade dos examinandos (EP baixos estão associados com testes compostos por itens com parâmetros de dificuldade próximos da habilidade do examinando).

A FIT mede a qualidade do banco de itens (BI) do teste. Um teste adequado deve apresentar informação considerável em toda a extensão desejável de θ. A Figura 6 apresenta um exemplo hipotético de quatro padrões de gráficos de FIT (linha contínua) com o respectivo EP(θ) (linha pontilhada) de testes onde foram aplicados dez itens com diferentes valores estimados para os parâmetros do ML3. O teste do gráfico (a) mostra que o teste oferece mais informação no intervalo de θ entre –3 e 1, aproximadamente, o que significa que esse teste é

adequado para estimar proficiências de indivíduos com baixa habilidade, pois é um teste formado mais por itens fáceis. Por outro lado, o teste do gráfico (b) mostra que o teste oferece mais informação no intervalo de θ entre –1 e 3, aproximadamente, o que significa que esse teste é adequado para estimar proficiências de indivíduos com alta habilidade, pois é um teste formado mais por itens difíceis. O teste do gráfico (c) não parece adequado para estimar qualquer proficiência, pois a informação fornecida é muito baixa e o EP, em relação à informação, é alto em todo intervalo da habilidade, o que indica que esse teste é formado por itens que não possuem boa discriminação. O teste do gráfico (d) mostra que o teste oferece informação adequada no intervalo de θ entre –3 e 3, aproximadamente, o que significa que esse teste é adequado para estimar proficiências de indivíduos com baixa, média e alta habilidade. Esse é o tipo de teste que possui itens fáceis, médios e difíceis com boa capacidade de discriminação. Segundo Pasquali (1996), para que um teste seja considerado de qualidade, ele precisa ter itens fáceis, médios e difíceis para estimar populações com diferentes níveis de dificuldade.

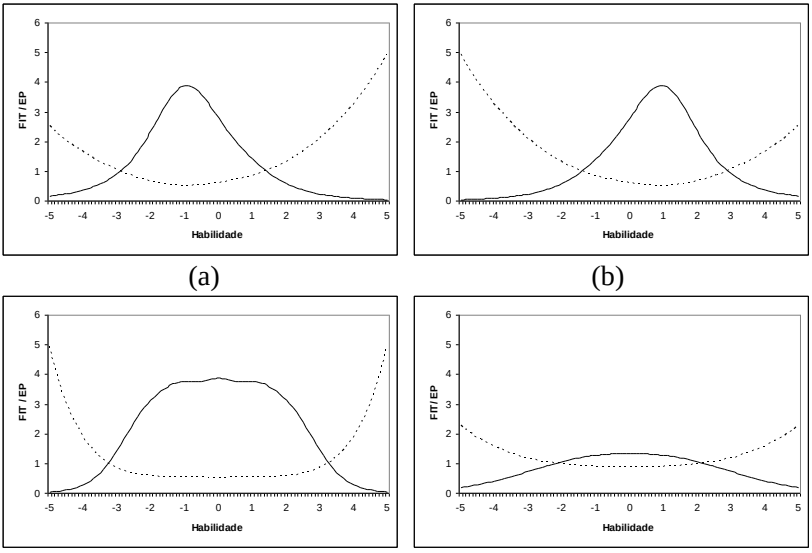


Figura 6. Exemplo de Gráficos da Função de Informação do Teste

No TAI, como cada indivíduo tem seu próprio teste personalizado, cada teste terá uma FIT diferente, segundo os itens que foram aplicados. Nesse caso, não é necessário que o teste tenha uma FIT semelhante ao gráfico (d) da Figura 6, já que ela deverá apresentar informação em torno da proficiência daquele indivíduo. Porém o banco de itens deverá apresentar uma FIT semelhante à do gráfico (d) da Figura 6.

2.3. CONSTRUÇÃO E CALIBRAÇÃO DO BANCO DE ITENS

Pode-se considerar um banco de itens (BI) uma base de dados de itens formada por uma parte descritiva (enunciado, opção correta, opções incorretas), uma parte de informação psicométrica (parâmetros estimados dos itens, tanto os da TCT quanto os da TRI) e qualquer outra informação relevante (por exemplo, conteúdo que cada item mede, dificuldade teórica, taxas de exposição do item). Oliveira (2002) define banco de itens como um grande conjunto de itens que armazenados de tal maneira, facilitam a sua recuperação em determinado momento.

O processo de calibração do banco dos itens consiste em estimar os parâmetros dos itens utilizando a TRI. Um banco de itens é considerado bem calibrado se as estimativas dos parâmetros dos itens forem adequadas e seus respectivos erros padrões forem baixos.

Olea, Ponsoda e Prieto (1999) destacam sete passos para a elaboração de um banco de itens:

1. Definição da estrutura do banco de itens: definem-se os tipos e os formatos de itens de acordo com as diferentes áreas de conteúdo.
2. Desenvolvimento dos itens: elaboração dos itens, onde pode-se aproveitar itens pré-existentes ou construir novos itens, procedendo com a análise de conteúdo clássica (PASQUALI, 1996; 1998).
3. Coleta de dados: definição do processo de coleta de dados para a calibração dos parâmetros dos itens por meio da TRI.
4. Administração dos itens: todos os itens deverão ser respondidos para a calibração dos parâmetros, mas não necessariamente pelos mesmos indivíduos, como será visto nessa Seção, ainda mais se o banco de itens for extenso. Essa aplicação poderá ser feita por um teste administrado por computador ou por um teste tradicional "papel e lápis". Segundo Segall (2005), vários estudos encontraram diferenças insignificantes no funcionamento da resposta do item, devido ao modo de administração (computador ou teste tradicional "papel e lápis"). Segall (2005) destaca ainda que o modo de coleta de dados por meio do formato tradicional "papel e lápis" é mais

rápido e tem um custo menor do que se a coleta for feita por meio do computador.

5. Análise dos itens: após a coleta de uma amostra suficiente de respostas, é realizada uma análise preliminar dos itens utilizando-se recursos da TCT e da TRI.
6. Calibração dos itens: processo de estimação dos parâmetros dos itens por meio da TRI, o qual será melhor detalhado nessa Seção.
7. Armazenamento de informação: os indicadores da TCT e os parâmetros estimados dos itens pela TRI devem ser armazenados juntamente com os itens no banco de itens.

Para calibrar os itens, é necessário que eles já tenham sido aplicados segundo um teste tradicional. De acordo com Andrade, Tavares e Valle (2000), seis formas diferentes de aplicações de testes podem ser encontradas na prática, as quais são ilustradas na Figura 7 para uma e duas populações (ou grupos):

1. Uma única população fazendo uma única prova.
2. Uma única população, dividida em dois ou mais subgrupos, fazendo duas provas totalmente distintas (nenhum item comum).
3. Uma única população, dividida em dois ou mais subgrupos, fazendo duas provas parcialmente distintas (com alguns itens comuns).
4. Duas ou mais populações com características diferentes fazendo uma única prova.
5. Duas ou mais populações com características diferentes fazendo duas provas totalmente distintas (nenhum item comum).
6. Duas ou mais populações com características diferentes fazendo duas provas parcialmente distintas (com alguns itens comuns).

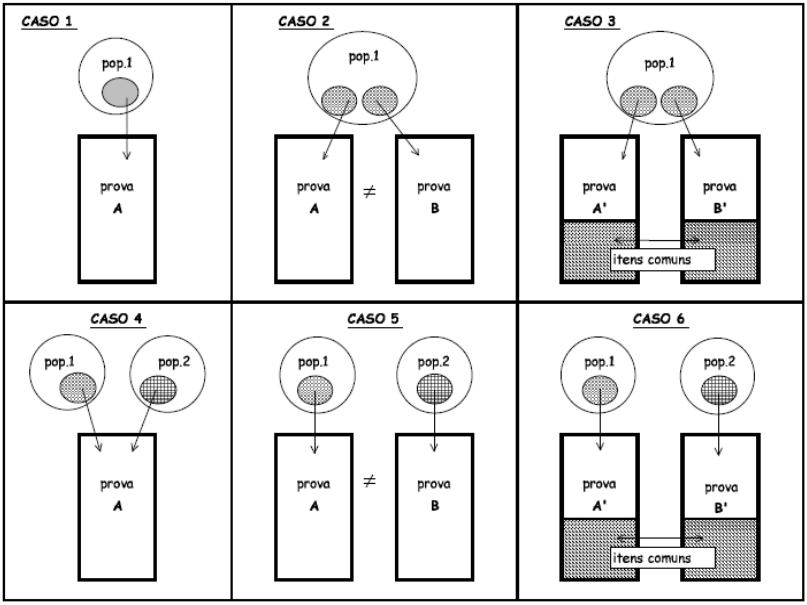


Figura 7. Representação gráfica das seis formas diferentes de aplicações de testes (Fonte: Andrade, Tavares e Valle, 2000)

No primeiro caso (mais simples) todos os itens são calibrados simultaneamente. Nos casos 2, 3, 4 e 6, é utilizado um processo conhecido como equalização, que, segundo Kolen e Brennam (1995), consiste em colocar parâmetros de itens vindos de provas distintas ou habilidades de respondentes de diferentes grupos na mesma métrica, ou seja, numa única escala comum, o que torna itens e indivíduos comparáveis. A equalização pode ser feita de duas formas:

- Equalização via população: uma única população de respondentes é submetida a provas distintas ou parcialmente distintas (casos 2 e 3 da Figura 7). Nesse caso, basta que todos os itens sejam calibrados conjuntamente para garantir que todos fiquem na mesma métrica. Além disso, deve-se garantir que a população seja a mesma, isto é, tenha a mesma característica. Por exemplo, alunos da primeira série do Ensino Médio de um determinado ano não caracterizam a mesma população em outro ano, pois os alunos mudam de série.
- Equalização via itens comuns: duas populações diferentes submetidas a provas iguais ou parcialmente distintas (casos 3 e 6 da Figura 7) com itens em comum. A calibração será garantida pelos itens em comum.

A Tabela 1 apresenta o tipo de equalização necessária para cada uma das seis situações representadas na Figura 7.

Tabela 1 Tipo de Equalização para as 6 situações analisadas

Caso	Equalização	Processo
1	Não necessita	Calibrar itens simultaneamente
2	Via população	Calibrar itens simultaneamente
3	Via população	Calibrar itens simultaneamente
4	Via itens comuns	Calibrar itens com as respostas de ambos os grupos simultaneamente.
5	Nenhuma	Nenhum. Não pode ser resolvida pela TRI.
6	Via itens comuns	Calibrar itens com as respostas de ambos os grupos simultaneamente.

Também é possível calibrar os itens separadamente nos casos 3 e 6 e utilizar a chamada "equalização a posteriori" para colocar itens e indivíduos na mesma escala. Maiores detalhes podem ser encontrados em Andrade, Tavares e Valle (2000).

O quinto caso não pode ser resolvido pela TRI pelo fato de não existir nenhum item em comum. Por outro lado, o sexto caso é o principal avanço da TRI em relação à TCT, pois que permite colocar indivíduos de diferentes populações na mesma escala e compará-los.

Nem sempre é possível aplicar todos os itens no mesmo teste, devido ao grande número de itens que o banco contém. Uma solução possível é aplicar subconjuntos diferentes de itens a diferentes indivíduos e fazer a calibração como nos casos 3 e 6 da Figura 7. Recomenda-se que os diferentes testes tenham pelo menos 20% dos itens em comum para obter-se um bom resultado na equalização (NAVAS, 1996). Nesse caso, deve-se cuidar para que os itens sejam calibrados na mesma escala. Para criar os diferentes testes, é comum a utilização de arranjos utilizados em projetos de experimentos, tais como o bloco incompleto balanceado (BIB) (ANDRADE; TAVARES; VALLE, 2000; EGGEN; VERHELST, 2011).

Os processos de calibração e equalização são matematicamente complicados e exigem métodos estatísticos complexos que precisam de auxílio computacional. Para esse fim, existem vários desenvolvidos. Alguns softwares bastante utilizados para a calibração e equalização são: o BILOG-MG (TOIT, 2003), o MULTILOG (REISE; YU, 1990; TOIT, 2003), o PARSCALE (TOIT, 2003), o R (MAZZA; PUNZO; MCGUIRE , 2011; PARTCHEV, 2011; RIZOPOULOS, 2010; WEEKS,

2011; ZOPLUOGLU, 2011) e o Xcalibre (WEISS; GUYER, 2010). Algumas listagens com vários softwares de TRI podem ser encontradas, por exemplo, em Deng (2011), Moreira Junior (2010) e Zhao e Hambleton (2009).

O tamanho da amostra necessário para calibração depende da quantidade de itens do banco, da quantidade de parâmetros do modelo da TRI a ser utilizado e do padrão de respostas da própria amostra, ou seja, é necessário que todas as categorias de respostas tenham uma quantidade de resposta suficiente para a estimação dos parâmetros dos itens. O modelo de Rasch (ML1) é o que exige a menor amostra necessária, sendo suficiente 200 observações, se a qualidade da amostra for boa. Para o Modelo de Resposta Gradual (MRG), Reise e Yu (1990) consideram que a amostra deve ser de pelo menos 500 observações. Em geral, é aceitável um mínimo de 500 observações por cada bloco de itens em que o banco está dividido (RENOM; DOVAL, 1999). Um tamanho insuficiente de observações pode gerar estimativas pouco precisas. Estudos de simulação podem ser utilizados para verificar o nível de erro na estimação dos parâmetros dos itens. Em situações reais talvez seja necessário calibrar um banco de itens com uma amostra pequena. Por exemplo, Zickar *et* al. (1999), que possuíam 164 itens para calibrar com um ML3 com uma amostra de 200 indivíduos, utilizaram dados simulados para aumentar o tamanho da amostra e calibrar o banco de itens. Primeiramente, os parâmetros dos itens foram estimados com base na amostra existente. Esses parâmetros estimados foram utilizados para simular mais respostas. Posteriormente, essas respostas foram juntadas com as respostas existentes e os parâmetros dos itens foram recalibrados, diminuindo o erro padrão das estimativas.

Deve-se eliminar do banco os itens com propriedades psicométricas inadequadas (item pouco discriminativo, com erro padrão alto ou que não se ajusta adequadamente). Por outro lado, a inclusão de novos itens pode ser feita gradualmente, sendo adicionados a um teste juntamente com os demais itens calibrados, onde eles não seriam utilizados para avaliar o respondente, mas apenas para serem calibrados. A calibração dos itens do banco pode ser atualizada quando se dispuser de mais respostas.

Para elaborar um banco de itens, além dos métodos clássicos da TCT, existe uma técnica conhecida como Geração Automática de Itens (GAI), que pode ser utilizada para criar banco de itens com características psicométricas conhecidas, sem a necessidade de calibrar os itens, conforme Revuelta (2000). A geração de itens se baseia em um conjunto de regras explícitas, programas em um computador, que

determinam como serão construídos os itens (BEJAR, 1993; IRVINE; KYLLONEN, 2002; HAMBLETON *et al*., 1996; HORNKE; HABON, 1986). Para gerar um item, é necessário aplicar algumas regras relacionadas ao enunciado do item, a alternativa correta e as alternativas incorretas, que devem ser coerentes com o item. No caso da GAI, os parâmetros do modelo não são estimados através da aplicação de amostras, devido a grande variedade de itens que podem ser gerados durante a realização do teste. Dessa forma, se tenta elaborar uma outra maneira para identificar o modo como as pessoas resolvem esse tipo de item a fim de determinar os valores para os parâmetros do item.

Uma solução possível é utilizar a metodologia de itens isomorfos (REVUELTA, 2000). Se existir um banco de itens elaborados e calibrados pelo método tradicional, cada item desse banco pode se tornar um modelo para elaborar itens isomorfos (diferentes do original, mas que possuem uma lógica similar de resolução). A GAI pode elaborar esses itens baseados na regra de solução dos itens já existentes, com dificuldades similares.

A idéia essencial dos modelos geradores de itens é captar o conhecimento que os especialistas deixaram nos itens elaborados por eles. Assim, não só é possível através de um modelo gerador produzir itens, mas também controlar seus atributos, ou seja, pode-se gerar itens onde se conhece antecipadamente seus parâmetros sempre que tivermos um modelo psicológico de como os sujeitos respondem aos itens. Isso representa uma melhor utilização do computador que, além de extrair itens de uma base de dados para a sua aplicação, passa a ter um papel ativo na elaboração dos itens. Esses modelos podem criar formas paralelas aleatorizadas a partir de modelos de itens e isomorfos. O termo "modelo de item" se refere a um conjunto de isomorfos e aos princípios e restrições que controlam a produção desses isomorfos. Isto supõe que na hora de criar um teste, levando em conta as especificações que o definem, pode-se gerar um grande número de versões de itens (modelos de itens) que medem cada uma das especificações. Assim, tem-se a idéia de realizar formas paralelas aleatorizadas através dos distintos conjuntos de modelos de itens e cada modelo gera diferentes isomorfos aleatoriamente. Portanto, cada indivíduo responde a um teste diferente, porém, se todos forem bem, todos os testes produzidos dessa maneira são equivalentes tanto psicometricamente quanto em termos da estimação da habilidade. Com o uso de modelos de itens, existe uma quantidade enorme de formas que se geram aleatoriamente na hora de elaborar um teste, o que contribui para o controle da superexposição de

um item (TEJADA, 2001). Entretanto, Wainer (2000a) afirma que esse tipo de procedimento não é aplicável em todos os tipos de testes.

2.4. MÉTODOS DE ESTIMAÇÃO DOS PARÂMETROS DOS ITENS

O processo de calibração dos itens é muito importante para o bom desempenho do uso da TRI. Segundo Bazán (2005), os problemas de estimação na TRI podem ser agrupados em três categorias: estimação por máxima verossimilhança (frequentista ou clássica), estimação Bayesiana e estimação Bayesiana com MCMC (*Markov Chain Monte Carlo*).

Os principais métodos de estimação são: o método de Máxima Verossimilhança Marginal (MVM), o método bayesiano da Moda a Posteriori (MAP), o método bayesiano da Média a Posteriori (EAP) e o método da Máxima Verossimilhança Conjunta (MVC). Esses métodos não possuem solução explícita, o que torna necessária a utilização de algum método numérico iterativo[6], como o Algoritmo Newton-Raphson, o Método Scoring de Fisher e o Algoritmo EM. Devido à dificuldade de integração das equações nos métodos de estimação, também é muito comum o uso do método numérico de integração de quadratura gaussiana, o qual consiste em aproximar as funções contínuas à funções discretas através de retângulos. Essas soluções envolvem cálculos bastante complexos e, conseqüentemente, necessitam de programas computacionais específicos.

Nessa seção esses métodos serão descritos de forma sucinta. Para maiores detalhes, sugerem-se os trabalhos de Andrade, Tavares e Valle (2000), Azevedo (2003; 2008), Baker (2001), Baker e Kim (2004), Bazán (2005), Costa (2009), de Ayala (2008), Embretson e Reise (2000), Francisco (2005), Glas (2010), Hambleton, Swaminathan e Rogers (1991), Thissen e Wainer (2001) e van der Linden e Glas (2000b; 2010).

2.4.1. Método da Máxima Verossimilhança Conjunta (MVC)

Segundo Andrade, Tavares e Valle (2000), esse método foi mais utilizado nos primeiros estudos da TRI, onde parâmetros dos itens e das proficiências eram estimados e maximizados simultaneamente (DE AYALA, 2009). Entretanto, por envolver uma quantidade muito grande

[6] Geralmente um método iterativo começa fornecendo um valor inicial para estimar os parâmetros. Entretanto, um problema que pode ocorrer é que o valor pode convergir para um máximo local e não para o máximo global. Para obter um resultado mais confiável, pode-se executar esse método com diferentes valores iniciais ou utilizar simulações.

de parâmetros a serem estimados, existem grandes problemas computacionais na utilização desse método.

Para contornar os problemas computacionais, foi proposto um processo que divide o MVC em duas etapas: (1) é iniciado com determinadas estimativas das habilidades (consideradas conhecidas) para a estimação dos parâmetros dos itens; (2) são estimadas as habilidades considerando os parâmetros dos itens conhecidos baseado nas estimativas da primeira etapa. Porém, esse procedimento pode produzir estimativas viciadas.

No método MVC existe o problema de identificabilidade do modelo, no que se trata da escala do mesmo. O fato é que existem diferentes valores de θ e de b que fornecem o mesmo valor para a probabilidade P_{ij}. A solução para esse impasse é fixar uma métrica. Geralmente se estabelece uma escala com média μ e desvio padrão σ, sendo muito comum utilizar $\mu = 0$ e $\sigma = 1$.

Conforme Andrade, Tavares e Valle (2000), o método MVC pode apresentar problemas de indeterminação e problemas na estimação do parâmetro de acerto casual (valores fora do intervalo [0, 1]), e da discriminação (valores negativos). Além disso, esse método não está definido para alguns padrões de resposta (itens respondidos corretamente ou incorretamente por todos os respondentes e respondentes que acertaram ou erraram todos os itens).

2.4.2. Método da Máxima Verossimilhança Marginal (MVM)

Em 1970, com o objetivo de resolver o problema de inconsistência do método MVC, foi proposto o método da máxima verossimilhança marginal para a estimação dos parâmetros em duas etapas (BOCK; LIEBERMAN, 1970): (1) estima-se os parâmetros dos itens, assumindo-se uma certa distribuição para as proficiências; e (2), estimam-se as proficiências assumindo os parâmetros dos itens conhecidos. Apesar do avanço que esse método trouxe para o problema, ele requeria que todos os parâmetros dos itens fossem estimados simultaneamente. Em 1981, Bock e Aitkin (1981) propuseram a utilização do algoritmo EM (DEMPSTER; LAIRD; RUBIN, 1977) – um processo iterativo para determinação de estimativas de máxima verossimilhança – que permitiu que os itens pudessem ter seus parâmetros estimados em separado, facilitando em muito o aspecto computacional do processo de estimação.

O método MVM, assim como o MVC, procura encontrar os valores dos parâmetros que fazem com que a probabilidade de ter dado a

resposta que foi encontrada seja a maior possível, ou seja, a mais provável de ter ocorrido.

Quando as habilidades são consideradas conhecidas, um procedimento alternativo de estimação dos itens consiste em agrupar as habilidades em categorias, o que pode reduzir bastante a exigência computacional. Para maiores detalhes, recomenda-se Andrade, Tavares e Valle (2000).

Conforme Andrade, Tavares e Valle (2000), o método da MVM pode apresentar problemas de indeterminação e problemas na estimação do parâmetro de acerto casual, obtendo valores fora do intervalo [0, 1], e da discriminação, obtendo valores negativos. Além disso, esse método não está definido para alguns padrões de resposta (itens respondidos corretamente ou incorretamente por todos os respondentes).

2.4.3. Métodos Bayesianos

Mais recentemente, os métodos bayesianos foram propostos para, entre outras coisas, resolver dois problemas das estimações por Máxima Verossimilhança: (1) estimação dos parâmetros dos itens respondidos corretamente ou incorretamente por todos os respondentes, (2) estimação das proficiências dos respondentes que acertaram ou erraram todos os itens da prova.

Nos métodos de Máxima Verossimilhança também há a possibilidade de que as estimativas dos parâmetros dos itens caiam fora do intervalo esperado, por exemplo, valores negativos para a discriminação ou valores estimados para o acerto casual fora do intervalo [0, 1]. A utilização de prioris nos métodos bayesianos é a solução para esses problemas.

A estimação bayesiana consiste em estabelecer distribuições a priori para os parâmetros, construir uma nova função denominada distribuição a posteriori e estimar os parâmetros de interesse com base em alguma característica dessa distribuição. Os métodos bayesianos mais utilizados para estimar os parâmetros são o da Média a Posteriori (EAP), que utiliza a média da distribuição a posteriori; e o da Moda a Posteriori (MAP), que utiliza a moda da distribuição a posteriori.

As prioris utilizadas são definidas conforme a característica do parâmetro a ser estimado. Para o parâmetro de discriminação a_i, geralmente adota-se como priori a distribuição Log-Normal ou a Qui-quadrado, justificando-se pelo fato que na prática os a_i devem ser positivos. Para o parâmetro de dificuldade b_i geralmente adota-se como

priori a distribuição Normal, por se supor que os b_i tenham distribuição Normal, já que estão na mesma escala das habilidades. Para o parâmetro de acerto casual c_i, que é uma probabilidade, geralmente adota-se como priori a distribuição Beta, uma vez que essa distribuição varia dentro do intervalo [0, 1] (ANDRADE; TAVARES; VALLE, 2000).

2.4.4. Método Bayesiano com MCMC

Conforme Bazán (2005), os métodos MCMC são um conjunto de métodos de simulação de amostras aleatórias de uma distribuição multivariada usualmente desconhecida, baseados na construção de uma cadeia de Markov cuja distribuição estacionária é a distribuição multivariada de interesse. No contexto da Inferência Bayesiana, a distribuição multivariada de interesse é uma distribuição a posteriori. Assim, estatísticas da distribuição teórica de interesse (desconhecida) podem ser estimadas através das correspondentes estatísticas da amostra aleatória simulada. Azevedo (2008) destaca que os métodos MCMC permitem obter, de forma empírica, a estrutura de distribuições a posteriori conjuntas e marginais que são complicadas ou impossíveis de obter-se de forma explícita.

2.5. MÉTODOS DE ESTIMAÇÃO DA HABILIDADE

Existem diversos métodos de estimação da habilidade, sendo que os mais utilizados são: o método de Máxima Verossimilhança (MV), o método bayesiano da Moda a Posteriori (MAP), o método bayesiano da Média a Posteriori (EAP) e o método da Máxima Verossimilhança Conjunta (MVC). Assim como no caso da estimação dos parâmetros dos itens, a esses métodos não possuem solução explícita, sendo necessária a utilização de algum método numérico iterativo, como o Algoritmo Newton-Raphson, o Método Scoring de Fisher e o Algoritmo EM. Da mesma forma, devido à dificuldade de integração das equações, também é muito comum o método numérico de integração de quadratura gaussiana.

Nessa seção esses métodos serão descritos de forma sucinta. Para maiores detalhes, sugerem-se os trabalhos de Andrade, Tavares e Valle (2000), Azevedo (2003; 2008), Baker (2001), Baker e Kim (2004), Costa (2009), Francisco (2005), Hambleton, Swaminathan e Rogers (1991), Thissen e Wainer (2001) e van der Linden e Pashley (2000).

2.5.1. Método da Máxima Verossimilhança Conjunta (MVC)

Esse método é o mesmo descrito na seção 2.4.1, no qual a estimação das proficiências ocorre simultaneamente com a calibração dos itens. Esse método de estimação não é comum dentro do contexto dos TAIs, onde o interesse é possuir um banco de itens previamente calibrado antes de estimar as proficiências.

2.5.2. Método da Máxima Verossimilhança (MV)

Esse método consiste em maximizar a verossimilhança, ou seja, encontrar os valores dos parâmetros que fazem com que a probabilidade de ter dada a resposta que foi encontrada seja a maior possível, considerando os parâmetros dos itens conhecidos. Entretanto, segundo Andrade, Tavares e Valle (2000), o uso do método MV resultará em uma expressão sem solução explícita para θ_j, necessitando de métodos iterativos. Além disso, para o ML3, nem sempre existe um único máximo da função de Verossimilhança (SAMEJIMA, 1973), o que pode levar o estimador de MV encontrar um máximo local ao invés do máximo global da função. O método MV também não está definido para os padrões de respostas constantes dos respondentes (indivíduos que acertam ou erram todos os itens respondidos). Todavia, alguns programas computacionais[7] que utilizam estimações por Máxima Verossimilhança fazem uso de algum artifício para contornar esse problema da indeterminação das estimativas, como, por exemplo, atribuir algum valor razoável para as estimativas. Segundo Kim e Nicewander (1993), a utilização do Método MV produz um viés na estimação de valores altos e baixos da habilidade: valores altos são superestimados e valores baixos são subestimados.

A utilização desse método supõe que as proficiências dos examinandos são estocasticamente independentes entre si. Dessa maneira, como examinandos diferentes não possuem informação de outros examinandos, pode-se estimar cada proficiência separadamente.

Uma variação do Método MV é o Método da Máxima Verossimilhança Ponderada (MVP), do inglês, *Weighted Maximum Likelihood –WML*, segundo Eggen e Straetmans (2000) e Weiss e Guyer (2010) também conhecido como Método da Estimação por Verossimilhança Ponderada (EVP), segundo o seu criador Warm

[7] Por exemplo, o BILOG e o BILOG-MG. Segundo Andrade, Tavares e Valle (2000), esses problemas são contornados através de um artifício: os indivíduos que erraram todos os itens ganham um meio certo no item mais fácil, e os indivíduos que acertaram todos os itens, perdem um meio certo no item mais difícil.

(1989). Nesse método, a função de verossimilhança é ponderada por uma função da Função de Informação do Teste (FIT). O método MVP é capaz de fornecer uma estimativa para θ baseado em um único item respondido ou um vetor não-misto de itens respondidos, semelhante aos métodos bayesianos. No entanto, não produz estimativas tendenciosas como os métodos bayesianos (WEISS; GUYER, 2010).

2.5.3. Métodos Bayesianos

Assim como na estimação por MV, a estimação bayesiana das habilidades é feita em uma segunda etapa, considerando os parâmetros dos itens conhecidos. Através da suposição de independência entre as habilidades de diferentes indivíduos, pode-se fazer as estimações de cada respondente individualmente. Ao contrário do Método MV, o Método Bayesiano está definido para os padrões de respostas constantes dos respondentes. Conforme Kim e Nicewander (1993), a utilização do Método Bayesiano produz um viés na estimação de valores altos e baixos da habilidade, inverso ao que ocorre no Método MV: valores altos são subestimados e valores baixos são superestimados.

Segundo Hambleton e Swaminathan (1985), o método bayesiano pode fornecer resultados significativos quando existe qualquer informação prévia sobre a habilidade dos respondentes. Geralmente, é utilizada uma distribuição a priori Normal para estimar a habilidade (Owen (1975) sugere uma Normal Padrão), porém para o ML3 uma priori Normal não fornecerá uma distribuição a posteriori Normal se o teste for curto (VAN DER LINDEN; PASHLEY, 2000).

A estimação pela moda da posteriori (MAP), introduzida na TRI em Lord (1986) e Mislevy (1986), consiste em obter o máximo da função de probabilidade a posteriori e necessita de algoritmos iterativos para isso, sendo mais complexa do que o método MV, e é mais eficiente entre os Métodos Bayesianos por necessitar de menos itens para alcançar certa precisão (REVUELTA; PONSODA, 2001). Entretanto, entre os Métodos Bayesianos, o EAP é o que obtém o menor erro padrão e a maior fidelidade (HONTANGAS; PONSODA; OLEA, 1999), além de não necessitar de métodos iterativos, já que são dados os pontos de quadratura e não é necessário calcular as integrais (COSTA, 2009), motivo pelo qual é mais recomendada. O estimador EAP apresenta-se mais estável do que o MAP para todos os tamanhos de testes adaptativos, incluindo o primeiro item administrado (BOCK; MISLEVY, 1982).

Para uma priori não informativa, a distribuição a posteriori será equivalente à função de verossimilhança e o estimador MAP apresentará

as mesmas propriedades do estimador MV. Para uma priori não Uniforme, as propriedades do estimador MAP em um teste curto dependerão da verossimilhança e da distribuição a posteriori que poderá ser multimodal, o que poderá fazer com que o estimador MAP encontre um máximo local. Por outro lado, o estimador EAP sempre existirá para uma distribuição a priori própria (VAN DER LINDEN; PASHLEY, 2010).

2.6. CONSTRUÇÃO DA ESCALA DE HABILIDADE

A calibração dos itens, devido à facilidade computacional, geralmente é feita na escala (0,1), ou seja, numa escala com média igual a zero e desvio padrão igual a 1. Após essa etapa, é realizada a construção da escala de habilidade ou de proficiência, que geralmente é colocada na mesma métrica dos itens. Conforme, Fontanive, Elliot e Klein (2007), as escalas de habilidade ordenam o desempenho dos indivíduos do menor para o maior de forma contínua e são cumulativas, isto é, os alunos que situam-se em um determinado nível da escala são capazes de demonstrar as habilidades descritas nesse nível e nos níveis anteriores dessa escala.

Segundo Andrade, Tavares e Valle (2000), os valores da escala de habilidade podem assumir teoricamente qualquer valor real entre $-\infty$ e $+\infty$, diferentemente da TCT, onde a escala geralmente varia entre 0 e a quantidade total de questões do teste. Dessa forma, é preciso estabelecer valores para a média e para o desvio padrão que representem a escala de habilidade dos indivíduos na população. Nesse sentido, é bastante comum fazer uma transformação linear em todos os parâmetros envolvidos antes da construção das escalas. Por exemplo, Vergara (2005) utilizou a escala (50, 15), considerando que a grande maioria dos respondentes possuem uma habilidade que varia entre $\pm$ 3 desvios padrões, obtendo assim uma escala que varia praticamente dentro do intervalo de 0 a 100. Vargas (2007), por sua vez, utilizou a escala (500, 50), enquanto que Oliva (2008) preferiu a escala (200, 20), Tezza (2009) utilizou (50, 10) e Costa (2009) a escala (100, 25). Entre as avaliações em larga escala no Brasil, o SARESP (Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo) utilizou no início a escala (50, 16) (VALLE, 2001) e depois passou a utilizar a escala do SAEB, que é a escala (250, 50) (FONTANIVE; ELLIOT; KLEIN, 2007), e o ENEM que utiliza a escala (500, 100) (BRASIL, 2010). Essas são formas de representar a habilidade em valores numéricos que tornem mais fácil o entendimento pelas pessoas, uma vez que, na prática, existe uma dificuldade em compreender os valores negativos e decimais que

existem na escala (0,1), onde muitas vezes os valores das proficiências são interpretados inadequadamente (VALLE, 2001). Mesmo assim, alguns autores utilizaram a escala (0,1) (ALEXANDRE et al., 2002a; FRANCISCO, 2005; GUEWEHR, 2007). De qualquer forma, a relação para a transformação da escala (0,1) em uma outra escala qualquer (μ, σ) não muda a relação de ordem entre os indivíduos e itens na escala e nem a probabilidade de resposta ao item pelo indivíduo (a habilidade do indivíduo é invariante à escala de medida). Essa transformação se dá da seguinte forma, para o parâmetro a:

$$a = \frac{a_{(0,1)}}{\sigma}, \tag{7}$$

para o parâmetro b:

$$b = \left(\sigma . b_{(0,1)}\right) + \mu, \tag{8}$$

e para a habilidade θ:

$$\theta = \left(\sigma . \theta_{(0,1)}\right) + \mu, \tag{9}$$

onde $\theta_{(0,1)}$ é a habilidade do indivíduo na escala (0,1).

Segundo Vergara (2005), a construção da escala de habilidade é efetuada após a calibração e equalização dos itens, com o objetivo de encontrar uma interpretação qualitativa dos valores obtidos pela aplicação do modelo da TRI, possibilitando assim, a interpretação pedagógica dos valores das habilidades. Nesse sentido, surge a idéia dos níveis âncoras e a técnica conhecida como ancoragem.

2.6.1. Níveis Âncoras

Andrade, Tavares e Valle (2000) definem níveis âncora como pontos selecionados pelo analista na escala da habilidade para serem interpretados pedagogicamente. Valle (2001) ressalta que esses níveis âncoras não podem ser muito próximos nem muito distantes, podendo-se tomar como base a média e o desvio padrão. Em geral, a maioria dos pesquisadores (VERGARA, 2005; VARGAS, 2007; OLIVA, 2008; TEZZA, 2009) determina os níveis âncoras em função do desvio padrão em relação à média, obtendo, assim, cerca de até 7 níveis âncoras ($-3\sigma + \mu$, $-2\sigma + \mu$, $-1\sigma + \mu$, μ, $1\sigma + \mu$, $2\sigma + \mu$, $3\sigma + \mu$). Outras configurações de níveis âncoras podem ser definidas (PEREIRA, 2004; KLEIN; FONTANIVE; ELLIOT, 2007), porém não são muito utilizadas na prática. No entanto, não se sabe a priori quantos níveis âncoras serão representados, já que isso depende da quantidade de itens

âncoras, que serão definidos na próxima seção. Para a definição e obtenção de níveis âncoras, sugere-se Beaton e Allen (1992).

A caracterização ou interpretação de um nível âncora se dará em função dos itens âncoras pertencentes ao nível, no sentido de identificar qual é o conjunto de habilidades de um descritor que esses itens tratam dentro desse nível. Espera-se que níveis mais altos estejam relacionados com assuntos que exigem uma proficiência maior e níveis mais baixos relacionados com assuntos que exigem uma proficiência menor. Assim, um sujeito que se situa acima de um determinado nível dominará os conteúdo caracterizados por esse nível e pelos níveis inferiores à esse, porém não dominará os assuntos dos níveis superiores a esse.

2.6.2. Itens Âncoras

De acordo com Costa (2009), após a fixação dos níveis âncoras, a técnica de ancoragem é utilizada para identificar itens que discriminam sucessivos pontos da escala chamados de itens âncoras.

Kolen e Brennan (1995) definem item âncora da seguinte forma: considere dois níveis âncora consecutivos Y e Z sendo que Y < Z. Um determinado item é âncora para o nível Z se e somente se as 3 condições abaixo forem satisfeitas simultaneamente:

$$P(U=1 \mid \theta = Z) \geq 0{,}65 \text{ e} \quad (13)$$

$$P(U=1 \mid \theta = Y) \leq 0{,}50 \text{ e} \quad (14)$$

$$P(U=1 \mid \theta = Z) - P(U=1 \mid \theta = Y) \geq 0{,}30. \quad (15)$$

Isso significa que, para um item ser considerado âncora em um determinado nível âncora, ele precisa ser respondido corretamente por um grande percentual de indivíduos (pelo menos 65%) com este nível de habilidade e por um percentual menor de indivíduos (no máximo 50%) com o nível de habilidade imediatamente anterior. Além disso, a diferença entre essas duas proporções deve ser de pelo menos 0,30. Assim, para um item ser âncora em um nível, ele deve ser um item "típico" desse nível, ou seja, bastante acertado por indivíduos com aquele nível de habilidade e pouco acertado por indivíduos com um nível de habilidade imediatamente inferior (ANDRADE; TAVARES; VALLE, 2000).

Na prática, às vezes um item não se caracteriza âncora por violar "levemente" uma das três condições necessárias. Nessas situações, pode-se considerar esse item como sendo âncora, se ele for importante ou se existirem poucos itens no instrumento de pesquisa. Outra alternativa é dividir os itens em categorias segundo a quantidade de

condições satisfeitas, método adotado por Costa (2009). Valle (2001) salienta que alguns níveis âncoras extremos podem ser mal caracterizados por serem definidos por itens muito fáceis ou muito difíceis, que geralmente são poucos.

Depois de identificados os itens âncoras de cada nível âncora, os especialistas no traço latente estudado devem caracterizar os níveis âncoras, segundo o conteúdo abordado no conjunto de itens que compõem cada nível. Após essa etapa, a escala está pronta para ser utilizada, por exemplo, para o posicionamento das populações ou dos indivíduos a fim de verificar os conteúdos dominados ou para identificar o percentual de indivíduos em cada nível de habilidade (VALLE, 2001). Outras formas de interpretação da escala na TRI são discutidas por Primi (2004) e Oliveira (2008).

3. TESTES ADAPTATIVOS INFORMATIZADOS – TAI

Uma nova forma de avaliação, que surgiu na década de 50, foi o chamado Teste Adaptativo (TA). Segundo Piton-Gonçalves (2004), um Teste Adaptativo (TA) tem como característica principal aplicar questões para o indivíduo de acordo com a sua habilidade, gerando assim um teste personalizado para cada indivíduo. Entretanto, aplicar um teste adaptativo naquela época não era tão simples assim. A grande dificuldade era apresentar para o indivíduo a próxima questão, de acordo com a resposta que ele desse à questão atual. Outra dificuldade era a utilização de um procedimento adequado para a seleção das questões, devido aos complexos cálculos. Porém, com o advento da informática, os testes adaptativos tornaram-se viáveis para a aplicação prática, surgindo assim, os Testes Adaptativos Informatizados (TAI).

Os Testes Adaptativos Informatizados[8] (TAI) – do inglês, *Computerized Adaptive Test*[9] (CAT) – são Testes Adaptativos (TA) administrados via computador. Dessa forma, cada questão é apresentada isoladamente ao indivíduo e, de acordo com a sua resposta, uma próxima questão é selecionada no banco de itens do teste para ser administrada ao indivíduo. O objetivo do TAI é apresentar itens ao indivíduo que sejam adequados ao seu nível de habilidade. A consequência disso é uma estimação mais precisa da proficiência com menos itens aplicados e em menos tempo do que nos testes convencionais do tipo "papel e lápis" onde todos os indivíduos devem responder todas as questões de um mesmo teste[10] (BARTRAM; HAMBLETON, 2006; DRASGOW; OLSON-BUCHANAN, 1999; EGGEN, 2004; FETZER et al., 2008; GEORGIADOU; TRIANTAFILLOU; ECONOMIDES, 2006; MEIJER; NERING, 1999; MILLS et al., 2002; MUÑIZ, 1997; OLEA; PONSODA, 1996; OLEA; PONSODA; PRIETO, 1999; PARSHALL et al., 2002; RENOM, 1993; SANDS; WATERS; MCBRIDE, 1997; SEGALL, 2005; VAN DER LINDEN; GLAS, 2000b; 2010; WAINER, 2000b; WEISS, 1983; 1985; WEISS; SCHLEISMAN, 1999).

---

[8] No Brasil, a maior parte dos pesquisadores (COSTA, 2009; DALPIAZ, 2007; DAMANDO, 2003; MOREIRA JUNIOR; TEZZA; ANDRADE, 2010; OLIVEIRA, 2002; PITON-GONÇALVES, 2004; SILVA; CURI, 2009) tem utilizado, em português, o termo Teste Adaptativo Informatizado (TAI). Recentemente, o termo Teste Adaptativo Computadorizado (TAC) tem sido utilizado (FERNANDES, 2009).

[9] Também são utilizados os termos *Computering Adaptive Testing* e *Computer Adaptive Test*.

[10] Nem todos os testes convencionais são dessa forma. Por exemplo, nas utilizações de provas paralelas (vide Seção 1.1) ou dos Blocos Incompletos Balanceados (BIB), os indivíduos não são submetidos a testes com exatamente as mesmas questões.

A Figura 8 apresenta um exemplo típico de um TAI para um teste com itens dicotômicos do tipo acerta/erra. O examinando inicia o teste com uma habilidade mediana, considerando a escala (0, 1). O primeiro item é administrado, o examinando acerta e sua habilidade estimada aumenta. O segundo item é administrado, o examinando acerta e sua habilidade estimada aumenta. O terceiro é administrado, o examinando erra e sua habilidade estimada diminui. O teste continua seguindo essa lógica até que seja encontrado um ponto de equilíbrio, onde o examinando domina o conhecimento que está abaixo desse ponto, mas não domina o conhecimento que está acima. É nesse ponto de equilíbrio que a sua habilidade deverá estar situada.

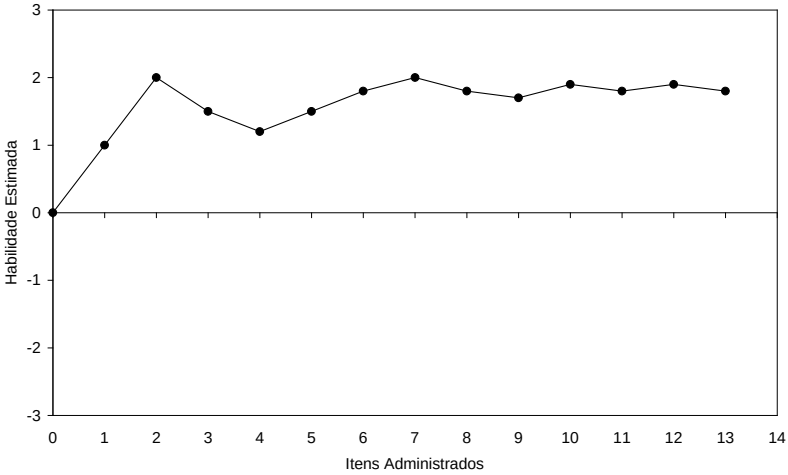


Figura 8. Exemplo de um TAI

Conforme Fernandes (2009), o TAI é utilizado para avaliações com intuito de medir o nível de conhecimento de um examinando ou verificar o nível mínimo de conhecimento. Segundo Fetzer et al. (2008), entre todos os métodos de testes disponíveis atualmente, o TAI é o que oferece o melhor equilíbrio entre precisão e eficiência. Cada indivíduo responde um teste personalizado (GREEN, 2000), com menos itens, o qual produz um resultado mais preciso que os tradicionais testes de "papel e lápis". Diversos estudos comparam o desempenho entre os TAIs e os testes tradicionais (AL-AMRI, 2008; BERGSTROM; LUNZ, 1992; BODMANN; ROBINSON, 2004; BUGBEE JUNIOR, 1996; CHOI; TINKLER, 2002; EIGNOR, 1993; EIGNOR; WAY; AMOSS, 1994; HALEY et al., 2008; HALKITIS, 1996; 1998; JONES, 2000;

KIM-KANG; WEISS, 2007; KINGSBURY, 2002; KINGSBURY; HOUSER, 1988; HOL; VORST; MELLENBERGH, 2005; LIGHTSTONE; SMITH, 2009; LUNZ; DEVILLE, 1996; MEAD; DRASGOW, 1993; OLEA et al., 2000; OLSEN; MAYNES; SLAWSON, 1986; ORTNER; CASPERS, 2011; PAEK, 2005; PARSHALL; KROMREY, 1993; POMMERICH, 2004; 2007; POMPLUN; FREY; BECKER, 2002; PONSODA et al., 1997; POWERS, 1999; ROTOU et al., 2003; SAWAKI, 2001; SCHAEFFER et al., 1995; 1998; THOMASSON, 1997; THOMPSON; WAY, 2007; VISPOEL; ROCKLIN; WANG, 1994; VISPOEL; WANG; BLEILER, 1997; WANG; KOLEN, 2001; WANG; SHIN, 2010; WANG et al., 2008; WEISS, 1973; 1985).

Sands e Waters (1997) salientam que administrar itens fáceis para examinandos com alta habilidade é desgastante e, as respostas corretas a esses itens agregam pouca informação para a estimação da proficiência desses indivíduos. Além disso, o examinando pode ficar entediado com os itens do teste que não oferecem nenhum desafio a ele e pode responder sem maiores cuidados os demais itens, introduzindo uma medida adicional de erro na estimação da proficiência. Similarmente, a administração de itens difíceis para indivíduos de baixa proficiência também é desgastante e as respostas incorretas não oferecem muita informação às estimativas. Diante de itens difíceis, os indivíduos de baixa proficiência estão mais propícios a se sentirem frustrados e acabam por responder aleatoriamente aos itens, incorporando erro adicional ao processo de estimação. Por outro lado, um instrumento como o TAI procura ajustar o teste a cada examinando diferente. Birnbaum (1968) mostrou que o item ideal é aquele com o maior valor do parâmetro de discriminação e com um valor para o parâmetro de dificuldade igual à capacidade do examinando. Devido às características da TRI, Van der Linden e Glas (2010) consideram que foi a TRI que forneceu uma base sólida para o desenvolvimento dos TAIs.

Desde os testes de Binet (BINET; SIMON, 1905), tem-se criado procedimentos de avaliação adaptativos mediante testes psicológicos, e nas últimas décadas, dado os avanços psicométricos alcançados pela utilização da TRI e os avanços técnicos no campo da informática, tem-se desenvolvido instrumentos informatizados para apresentar apenas os itens que são altamente informativos para estimar a habilidade de cada indivíduo. O nível de informação depende, dentre outros fatores, de que os itens selecionados possuam uma dificuldade apropriada para avaliar o nível de habilidade do indivíduo (OLEA; HONTANGAS, 1999).

Os TAIs também permitem a elaboração dos chamados “tipos inovadores de itens” (DRASGOW; OLSON-BUCHANAN, 1999; OLEA; ABAD; BARRADA; 2010; OLEA; PONSODA, 1998; PARSHALL, DAVEY; PASHLEY, 2000; PARSHALL et al., 2010), que são aqueles que se beneficiam do suporte computacional em relação à sua construção (uso de sons, gráficos, animação ou vídeo) e em relação ao procedimento de resposta (por exemplo, marcar em partes de figuras ou gráficos complexos, selecionar trechos de um texto, mover objetos, escrever o resultado de um problema, responder usando microfone, etc.). Isso representa um importante avanço, pois permite avaliar melhor alguns tipos de habilidade, tais como, a aptidão musical ou o rendimento de um controlador de tráfego aéreo. Alguns tipos desses itens podem reduzir significativamente a chance do “acerto casual”, por exemplo, um item construído para marcar em uma parte de um gráfico pode ter várias opções de resposta.

Embora, teoricamente, o TAI seja uma idéia relativamente simples, a realidade do planejamento, implementação e manutenção de um programa de TAI é substancialmente mais complexa (WISE; KINGSBURY, 2000). O desenvolvimento de um teste consistente requer uma grande equipe, composta por profissionais de várias áreas e recursos computacionais com hardware e software apropriados. O desenvolvimento do software de um TAI requer implementação de algoritmos complexos que utilizam métodos estatísticos. Esses algoritmos consomem muitos recursos, pois um extenso cálculo deve ser realizado nos modelos mais complexos de seleção de questões. Dependendo do tipo de algoritmo escolhido para a seleção das questões, a cada resposta dada pelo examinando, o sistema aplicador deve realizar um cálculo estatístico com todas as outras questões envolvidas no banco de questões (FERNANDES, 2009). O suporte informático é necessário para a aplicação de um TAI baseado na TRI (OLEA; PONSODA, 1996).

A utilização dos TAIs tem sido particularmente útil quando (WAINER, 2000a; WAINER; EIGNOR, 2000):

a) A natureza do construto é de tal forma que a administração informatizada ajuda na sua avaliação (por exemplo, na utilização de sons ou animação);
b) O teste é administrado freqüentemente (várias vezes ao ano);
c) As pessoas que fazem o teste têm interesse em obter o nível de habilidade com uma alta precisão (por exemplo, identificar o nível de estresse a fim de fazer um tratamento adequado).

As aplicações de TAIs podem ser realizadas em computadores independentes de uma rede local ou por meio da internet (BARTRAM; HAMBLETON, 2006). No primeiro caso, cada computador deve ter instalado o software necessário para a avaliação. Se eles estiverem ligados a uma rede, uma unidade central pode controlar o processo de avaliação e registrar os resultados. Alguns aspectos práticos devem ser observados (OLEA; PONSODA, 1996):

a) O tempo entre a emissão da resposta e a apresentação do próximo item deve ser imperceptível para o respondente.
b) Necessita-se de uma grande capacidade de armazenamento de dados, onde resida a informação sobre o banco de itens, os resultados da calibração, as instruções, os exemplos de prova, os resultados de cada respondente, e o software.
c) Garantir uma boa resolução de tela no caso de itens com conteúdo gráfico.
d) Simplificar o procedimento de resposta do respondente, onde ele possa marcar a sua resposta com o mouse ou com o teclado e pressionar a tecla ENTER para confirmar o envio da resposta.

No caso da aplicação via internet (Educação à distância), deve-se levar em consideração principalmente a questão da segurança do teste, por exemplo, verificar se o indivíduo que está respondendo o teste é ele mesmo e não outro (BARTRAM; HAMBLETON, 2006; LOZZIA et al., 2009; MUÑIZ, 2005; OLEA; ABAD; BARRADA, 2010; WAINER, 2000a). Esses testes também estão sujeitos a limitações da tecnologia informática, tais como, velocidade de conexão, falhas na conexão, problemas com o computador, incompatibilidade, etc. Porém existem várias vantagens, tais como, a rapidez do resultado, o baixo custo e a facilidade de acesso (LOZZIA et al., 2009).

Para implantar um TAI, existe a opção de programar um algoritmo ou de utilizar um software disponível. Os algoritmos têm sido programados em Turbo Pascal (OLEA et al., 1996), Basic (LINACRE, 2000), linguagem C (OLEA et al., 2004; PITON-GONÇALVES, 2004), linguagem ASP (DAMANDO; GUEDES, 2004; 2005), no SAS (RAÎCHE; BLAIS, 2006) e em R (DIAO; VAN DER LINDEN, 2011; MAGIS; RAÎCHE, 2010). Alguns softwares criados foram o *AddChart Application* (YI; ZHANG; CHANG, 2006), o ADEPT – *Adaptive English Proficiency Test for the Web* (PITON-GONÇALVES, 2004), o *ADTEST* (PONSODA; OLEA; REVUELTA, 1994), o *APT-System* (OLEA; PONSODA, 1996; PONSODA; OLEA; ALCÁZAR, 1994), o *Assessment Center*^{SM} (CELLA; GERSHON, 2010), o *CATGlobal* (PROMISSOR, 2003), o CAT *Software System* (COMPUTER

ADAPTIVE TECHNOLOGIES, 1994), o *CatR* (MAGIS; RAÎCHE, 2010; 2011), o *CATSys* (DE LA TORRE; VISPOEL, 1991), o *CATSim* (THOMPSON, 2009b; WEISS; GUYER, 2010), o *CPLEX* (ILOG, 2003; VAN DER LINDEN, 2010a), o *DEMOTAC* (OLEA; PONSODA, 1996; RENOM, 1993), o *FastTEST Professional* (WEISS, 2006), o *FastTEST Web* (THOMPSON, 2009b), o *Firestar* (CHOI, 2009), o *GenTAI* (LÓPEZ-CUADRADO; FERNÁNDEZ, 2005; LÓPEZ-CUADRADO et al., 2008), o *METRIX Engine* (RENOM, 1993), o *MicroCAT* (ASSESSMENT SYSTEMS CORPORATION, 1994; KINGSBURY, 1990), o *MISTRAL* (SALCEDO; PINNINGHOFF; CONTRERAS, 2005), o *Perception* (QUESTION MARK CORPORATION, 1998), o *PETA* (NITKO; HSU, 1984), o *POSTSIM* (WEISS, 2008), o *Question Mark for Windows* (QUESTION MARK CORPORATION, 1997), SCAALE – Sistema Computadorizado de Avaliação Adaptativa em Larga Escala (FERNANDES, 2009), o *SIETTE* (CONEJO et al., 2001; GUZMÁN; CONEJO, 2004; GUZMÁN; CONEJO; GARCÍA-HERVÁS, 2005), o *SIMCAT* (RAÎCHE; BLAIS, 2006; RAÎCHE; BLAIS; RIOPEL, 2006), o *SimulCAT* (HAN, 2010), o *Test Editor* (ROMERO et al., 2006) e o *UCAT* (LINACRE, 1987a; 1987b; 2000). Um software para o TAI se caracteriza por conter uma série de módulos que processam diferentes passos do teste de forma independente e se encontram em uma relação hierárquica (OLEA; PONSODA, 1996). Um software completo para a avaliação adaptativa deve ter (HAMBLETON; ZAAL; PIETERS, 1991):

- Procedimentos de identificação dos respondentes e das provas aplicadas;
- Textos e parâmetros do banco de itens;
- Um módulo de construção de testes;
- Um módulo de apresentação de itens (início do teste, seleção de itens, finalização do teste, estimação final da habilidade e sua precisão, e armazenamento dos resultados);
- Um módulo de atualização do banco de itens (rendimento dos indivíduos, informação histórica dos itens);
- Um módulo para oferecer ao usuário uma informação escrita do seu rendimento.

Antes que um TAI seja efetivamente implantado, é necessário avaliá-lo mediante algum controle psicométrico de qualidade para verificar a precisão e a validade do teste, através de dados empíricos ou simulações. Muñiz e Hambleton (1999) comentam sobre os seguintes

aspectos: a) em relação à precisão (erro padrão de estimação, erro quadrado médio, desvio empírico médio, eficiência, correlações entre as estimações do banco e da aplicação, procedimentos provenientes da TCT), e b) em relação à validade (de conteúdo, de construto e de predição), os quais serão explorados na seção 3.5.

Quando o objetivo do TAI for realizar certificações (proficiência ou avaliações educacionais), Way (1998) propõe que: a) o banco deve ter 8 vezes mais itens do que a quantidade que vai ser aplicada, b) cada item não pode ser apresentado para mais do que 15% dos candidatos, c) para qualquer par de candidatos, não se deve apresentar, em média, mais de 20% de itens em comum, d) para qualquer par de candidatos com nível de habilidade semelhante, não se deve apresentar mais de 40% de itens comuns.

3.1. BREVE HISTÓRICO

O desenvolvimento e aperfeiçoamento das técnicas psicométricas necessárias para implantar testes adaptativos levou várias décadas. Muitos modelos de TAs foram criados ao longo dos anos, sendo que inicialmente o interesse era na aplicação tradicional "papel e lápis" (VAN DER LINDEN; GLAS, 2010). Weiss (1985) destaca os seguintes TAs desenvolvidos:

- Teste Binet *(Binet Test)*: é um teste de inteligência baseado em níveis de dificuldade, criado por Binet e Simon (1905). As questões são classificadas por especialistas segundo os níveis de dificuldade. Se todos os itens de um nível de dificuldade forem respondidos corretamente, são fornecidos itens de um nível mais alto, até que todos eles sejam respondidos incorretamente, identificando, assim, o Nível Superior. Após alcançar esse nível, são fornecidos itens de um nível mais baixo, até que todos eles sejam respondidos corretamente, identificando, assim, o Nível Inferior. Pode-se iniciar o teste tanto para identificar primeiramente o Nível Inferior quanto o Nível Superior. O teste termina quando os dois níveis são identificados.
- Teste Adaptativo de Dois Estágios (*Two-Stage Adaptive Testing*): é basicamente dividido em dois subtestes, sendo o primeiro com itens de dificuldade média (*Routing Test*). Baseado nas respostas do primeiro subteste, os indivíduos recebem o segundo teste (*Measurement Test*), o qual é mais adaptado à habilidade dos indivíduos (CRONBACH; GLESER, 1957).

- Teste Adaptativo Piramidal (*Pyramidal Adaptive Test*): proposto por Larkin e Weiss (1975), onde os itens são organizados pelo grau de dificuldade, numa estrutura que lembra a forma de uma pirâmide. O primeiro item, de dificuldade média, encontra-se no topo da pirâmide e é administrado a todos os indivíduos. Se o indivíduo acerta, no próximo nível recebe um item um pouco mais difícil, se erra, recebe um item um pouco mais fácil. Dessa forma, o teste vai administrando itens gradativamente mais difíceis (ou mais fáceis) até chegar ao final na base da pirâmide.
- Teste Adaptativo Estratificado (*Stratified Adaptive Test*): é uma melhoria do Teste Binet, proposto por Weiss (1973). Quando o aluno responde corretamente um item, o próximo item a ser apresentado é de um nível dificuldade maior, porém se o aluno responde incorretamente o item, o próximo item a ser apresentado é de um nível de dificuldade menor. O teste termina quando for identificado o Nível Superior de dificuldade (nível no qual todos os itens foram respondidos incorretamente).
- Teste Adaptativo Baseado na Teoria de Resposta ao Item: a Teoria de Resposta ao Item (TRI) vem sendo muito utilizada ultimamente nos TAIs devido à possibilidade computacional e às vantagens apresentadas. A TRI permite estimar os parâmetros dos itens (dificuldade, discriminação e acerto casual) para a calibração do banco de itens, estimar a habilidade do respondente durante o teste, calcular a informação dos itens e construir uma escala única onde os indivíduos que responderam itens diferentes podem ser comparados. Os parâmetros dos itens permitem estimar a habilidade do respondente com maior precisão do que os testes mencionados anteriormente, onde a seleção do item se baseava apenas na dificuldade do item.

As primeiras experiências sobre testes adaptativos foram desenvolvidas no início do século XX por Binet e Simon (1905), que construíram testes de inteligência segundo a idade e escolaridade dos estudantes. Porém, a idéia de ajustar um teste a um indivíduo de forma automatizada, veio de uma sugestão de Bill Turnbull. Por volta dos anos 60, ele se junta a Frederick Lord com o propósito de implementar essa idéia. O trabalho de Frederick Lord foi primordialmente subsidiado pela ONR (*Office of Naval Research*), a mesma agência que subsidiou as pesquisas de David Weiss, que formou um grupo de trabalho que produziu as primeiras implantações de TAIs em Minesota, EUA. No entanto, foram Lord e Novick (1968) que estabeleceram as bases da TRI

e os fundamentos estatísticos para colocar diversos examinandos na mesma escala mesmo que tenham respondido itens diferentes. As idéias originais de TAIs fundamentados na TRI são de Lord, mas o desenvolvimento se deu nos anos 80 através de um convênio entre a Universidade de Minesota (dirigido por Weiss) e o exército americano, para elaborar versões adaptativas do *Armed Services Vocational Aptitude Battery* (ASVAB). Durante uma década se estudaram as vantagens de uma versão adaptativa do teste, aplicado anualmente a mais de 500.000 pessoas. Na metade dos anos 80 se aplicou a primeira versão adaptativa do ASVAB. Segundo Fetzer et al. (2008), o ASVAB é o único programa de TAI em grande escala que tem sido utilizado para fins de contratação, o qual avalia habilidades profissionais relevantes para uma grande variedade de postos de trabalho no serviço militar. Essa foi considerada a primeira geração dos TAIs. Porém, as deficiências computacionais daquela década e outros problemas enfrentados, como a complexa relação entre a dificuldade do item e o tempo necessário para respondê-lo, causaram uma certa decepção em relação ao TAI (TEJADA, 2001; OLEA; HONTANGAS, 1999). Por outro lado, em 1993, foi constatado que o TAI economizava mais de três milhões de dólares anualmente no processo de seleção militar (RENOM; DOVAL, 1999).

Após o ASVAB, com o advento de computadores mais poderosos, outros programas de avaliação passaram a substituir a avaliação tradicional "papel e lápis" por testes adaptativos, tais como o *National Council Licensure Examination for registered nurses* (NCLEX) e o *Graduate Record Exam* (GRE). Um grande número de programas passaram a utilizar os TAIs, não só na educação mas também na psicologia e, mais recentemente, áreas como marketing e pesquisa de resultados em saúde (VAN DER LINDEN; GLAS, 2010).

Nas décadas de 70 e 80 os procedimentos para a seleção de itens não podiam depender de procedimentos numéricos muito complexos. Renom (1993) destaca os seguintes procedimentos utilizados:

a) Estratégia de duas etapas (LORD, 1971c): na primeira etapa, os indivíduos respondiam um teste de tamanho curto, a fim de estimar o nível de habilidade. Na segunda etapa, cada um respondia um teste com itens selecionados de acordo com a habilidade estimada na primeira etapa.
b) Estratégia do nível flexível (LORD, 1971a; 1971b): mais apropriado para bancos com poucos itens. Divide-se o banco de itens em duas metades, uma com os itens mais fáceis e outra com os itens mais difíceis. Começa-se com o item de dificuldade média, se for

acertado, passa-se ao próximo item mais difícil, caso contrário, apresenta-se o seguinte mais fácil (entre os itens que ainda não foram apresentados). A prova termina quando se responde a metade do banco.

c) Estratégias alternativas de ramificação fixa (WEISS, 1974): possuem um estabelecimento prévio de uma ordenação dos itens, no formato de pirâmide, segundo a sua dificuldade. A seleção de um item depende da resposta anterior, porém as possibilidades seqüenciais a apresentar estão pré-fixadas de antemão.

d) Estratégia de ramificação variável (LORD, 1980): se estima um nível provisório de habilidade após se responder uma quantidade fixa de itens. Escolhe-se o item mais informativo para o nível, surgindo, assim, o método da máxima informação.

Nos últimos anos, a aplicação dos TAIs tem sido cada vez mais frequente nos Estados Unidos e na Europa (OLEA et al., 2004; TEJADA, 2001). De acordo com Fetzer et al (2008), cerca de 30 programas de TAIs avaliam de 4 a 6 milhões de indivíduos por ano em todo o mundo. Uma lista de 25 grandes programas de TAIs aplicados no mundo está disponível em Fetzer et al. (2008). Entre as principais aplicações de TAIs, merecem destaque: o *Test of English Foreign Language* (TOEFL) (EIGNOR et al., 1998; HICKS, 1989; KIRSCH, 1998; MCKINLEY; WAY, 1992; OLEA et al., 2004; STRICKER; WILDER, 2001; TANG; EIGNOR, 1997; TANG; WAY; CAREY, 1993; WAINER; WANG, 2000; YAMAMOTO, 1995), o *Graduate Record Exam* (GRE) (BEJAR et al., 2003; BRIDGEMAN; CLINE, 2000; BRIDGEMAN; CLINE; HESSINGER, 2003; EIGNOR et al., 1993; GOLDBERG; PEDULLA, 2002; MILLS; STEFFEN, 2000; OLEA et al., 2004; POWERS, 1999; SCHAEFFER et al., 1993; 1995; 1998; STOCKING; SMITH; SWANSON, 2000; WAINER; WANG, 2000) e o *Scholastic Aptitude Tests* (SAT) (EIGNOR, 1993; EIGNOR et al., 1993; TEJADA, 2001) da *Educational Testing Service* (ETS), o *Armed Services Vocational Aptitude Battery* (ASVAB) (FETZER et al., 2008; HETTER; SEGALL, 1997; SEGALL; BLOXOM, 1997; HETTER; SYMPSON, 1997; MCBRIDE; MARTIN, 1983; MCBRIDE et al., 2001; MORENO, 1997; MORENO; SEGALL, 1997; OLEA et al., 2004; SANDS; WATERS, 1997; SEGALL, 1993; 1997; SEGALL; MORENO, 1999; SEGALL; MORENO; HETTER, 1997; SEGALL et al., 1997; STICHA; BARBER, 2003; WOLFE; MCBRIDE; SYMPSON, 1997; WOLFE; MORENO; SEGALL, 1997), o *Graduate Management Admissions Test* (GMAT) (FETZER et al., 2008; RUDNER, 2010; TEJADA, 2001), o *National Council of Architectural*

Registration Boards (NCARB) (TEJADA, 2001) e o *National Council Licensure Examination for registered nurses* (NCLEX) (EIGNOR; WAY; AMOSS, 1994; RECKASE; HE, 2005; TEJADA, 2001; ZARA, 1999). Os TAIs têm sido usados em exames de admissão e contratação (FETZER et al., 2008; XING; HAMBLETON, 2004), em avaliações educacionais ou de conhecimentos (BOYD; DODD; FITZPATRICK, 2003; BURGHOF, 2001; DAVIS; DODD, 2001; 2003; FINKELMAN; WEISS; KIM-KANG, 2010; GARCIA; REVUELTA, 2003; GLOWACKI; MCFADDEN; PRICE, 1995; GUZMÁN; CONEJO, 2004; HARMES; KROMREY; PARSHALL, 2001; HALKITIS, 1993; KIM-KANG; WEISS, 2007; 2008; KINGSBURY, 1990; KINGSBURY; HOUSER, 1999; KREITER; FERGUSON; GRUPPEN, 1999; LI; SCHAFER, 2003a; 2005b; LUECHT; NUNGESTER, 2000; MELICAN; BREITHAUPT; ZHANG, 2010; MILLS; STEFFEN, 2000; MORRISON; NUNGESTER, 1995; OLEA et al., 1996; 2000; 2004; POMPLUN; FREY; BECKER, 2002; RUDNER, 2010; SUKAMOLSON, 2002; TRUELL; ZHAO; ALEXANDER, 2005; TEJADA, 2001; VAS, 2007; VERSCHOOR; STRAETMANS, 2000, 2010; VISPOEL, 1993; 1998a; 1998b; VISPOEL; HENDERICKSON; BLEILER, 2000; WANG et al., 2008; WAY; DAVIS; FITZPATRICK, 2006; WEISS; KINGSBURY, 1984), em testes de proficiência de línguas estrangeiras (ABAD et al., 2010; ABERNATHY, 1986; CHALHOUB-DEVILLE, 2000; CHALHOUB-DEVILLE; DEVILLE, 1999; CHOI; KIM; BOO, 2003; CISAR et al., 2010; GIOUROGLOU; ECONOMIDES, 2005; HAYDEN, 2003; LILLEY; BARKER; BRITTON, 2004; MADSEN, 1986; MEUNIER, 1994; NOGAMI; HAYASHY, 2010; OLEA et al., 1996; 2004; PONSODA et al., 1997; SAWAKI, 2001; SUMBLING et al., 2007), na medição do potencial de aprendizagem (DE BEER, 2002; 2003; 2005), em testes psicológicos (AGUADO et al., 2005; FINKELMAN; WEISS; KIM-KANG, 2010; FLIEGE et al., 2005; FORBEY; BEN-PORATH, 2007; FORBEY; BEN-PORATH; GARTLAND, 2009; GARCIA et al., 2000; HOL; VORST; MELLENBERGH, 2005; 2007; 2008; MCBRIDE, 1988; MORLEY, 2000; REISE; HENSON, 2000; RUBIO; SANTACREU, 2003; SCHOONMAN, 1989; SIMMS; CLARK, 2005; WALLER; REISE, 1989; WALTER; HOLLING, 2008), na medicina e áreas da saúde (ANATCHKOVA, 2009; BJORNER; KOSINSKI; WARE JUNIOR, 2003; CELLA *et al.*, 2007; CHAKRAVARTY; BJORNER; FRIES, 2007; COOK et al., 2003; 2005; COSTER et al., 2008; FAYERS, 2007; FRIES; BRUCE; CELLA, 2005; GARDNER; KELLEHER; PAJER, 2002; GARDNER et al., 2004; GIBBONS et al.,

2008; HALEY et al., 2006; 2008; 2009a; 2009b; HART et al., 2008a; 2008b; 2008c; 2009; 2010; HART; MIODUSKI; STRATFORD, 2005; HWANG et al., 2005; IP et al., 2010; JACOBUSSE; VAN BUUREN, 2007; JANSKY; HUANG, 2009; JETTE; HALEY, 2005; JETTE et al., 2008; KOCALEVENT et al., 2009; KOSINSKI et al., 2003; 2006; LAI et al., 2003; 2005; LEWIS et al., 1988; MCHORNEY, 1997; 2003; PETERSEN et al., 2006; 2010; REBOLLO et al., 2009; REEVE et al., 2007; REVICKI; CELLA, 1997; RILEY; DENNIS; CONRAD, 2010; RILEY et al., 2007; SCHWARTZ et al., 2006; SMITS; CUIJPERS; VAN STRATEN, 2011; VOGELS; JACOBUSSE; REIJNEVELD, 2008; WALTER, 2010; WALTER et al., 2007; WANG et al, 2009; 2010; WARE et al., 2003; 2005; YOUNT et al., 2011) e em outras áreas (BAEK, 1995; VISPOEL; WANG; BLEILER, 1997).

No Brasil, poucos têm sido os estudos sobre TAIs baseados na TRI, em algumas áreas, tais como proficiência de línguas estrangeiras (COSTA, 2009; COSTA; FERNANDES, 2009; COSTA et al., 2009; CURI et al., 2009; FERNANDES, 2009; KARINO; COSTA; LAROS, 2009; OLIVEIRA, 2002; PITON-GONÇALVES, 2004; PITON-GONÇALVES; MONZÓN; ALUÍSIO, 2009), avaliação educacional (CURA JUNIOR et al., 2007; DAMANDO, 2003; DAMANDO; GUEDES, 2004; 2005; DAMANDO et al., 2004a; 2004b; DESCOVI, 2009; GROENWALD; BECHER, 2009; GROENWALD; RUIZ, 2006; SASSI; AMARAL; CURI, 2009), na medicina e áreas da saúde (SILVA; CURI, 2009), avaliação da usabilidade de sites (MOREIRA JUNIOR; TEZZA; ANDRADE, 2010), além da avaliação da prova teórica do DETRAN-SC proposta nesse trabalho.

Devido aos recentes avanços tecnológicos, o TAI tem sido usado em aplicações de larga escala (HAMILTON; KLEIN; LORIE, 2000; MILLS; STOCKING, 1996; MELICAN; BREITHAUPT; ZHANG, 2010; MILLS; STEFFEN, 2000; NOGAMI; HAYASHY, 2010; RUDNER, 2010; VAN DER LINDEN; GLAS, 2000b; 2010; VERSCHOOR; STRAETMANS, 2000, 2010; WAINER; EIGNOR, 2000; WALTER, 2010; XING; HAMBLETON, 2004) e também na internet e educação à distância (ABAD et al., 2010; BARTRAM; HAMBLETON, 2006; CONEJO et al., 2004; GUZMÁN; CONEJO, 2004; HAMILTON; KLEIN; LORIE, 2000; HO; YEN, 2005; HOCKEMEYER; ALBERT, 1994; LEE; PARK; PARK, 2006; OLEA; ABAD; BARRADA, 2010; KUO; LIN; YUAN, 2006; MCBRIDE et al., 2001; PHANKOKKRUAD; WORARATPANYA, 2009; SALCEDO; PINNINGHOFF; CONTRERAS, 2005; SHERMIS et al., 1997; TAO; WO; CHANG, 2008; WAINER; EIGNOR, 2000), onde permite uma

máxima segurança no teste ao mesmo tempo em que fornece estimativas precisas da habilidade do indivíduo com a máxima eficiência. Nenhum outro método de avaliação permite a criação instantânea de um teste voltado para um único indivíduo, ou seja, um teste individualizado (FETZER et al., 2008).

Ultimamente, tem ocorrido uma maior preocupação com a redução do custo de elaboração, manutenção e renovação do banco de itens e com a estimação dos parâmetros dos itens no sentido de melhorar as estimativas e simplificar os procedimentos de estimação. O enfoque dos estudos atuais tem tido duas vertentes: a estimação dos parâmetros dos itens e a geração automática de itens (SIERRA-MATAMOROS, 2007). Nos últimos anos, os estudos sobre TAIs também têm sido direcionados a restrições nos testes adaptativos, regras para geração de itens, modelos de itens não dicotômicos, testes multidimensionais, testes para classificação, tempo de resposta ao item, testes com múltiplos estágios e otimização da sequência de baterias de testes (VAN DER LINDEN, 2008a).

3.2. VANTAGENS E DESVANTAGENS DE UM TAI

Muitas vantagens são obtidas com a implantação de um TAI em lugar de um teste convencional do tipo "papel-e-lápis" em relação a dois fatores: a informatização do teste e a adaptação dos itens ao examinando. Entre elas, destacam-se:

- A correção automática, reduzindo o tempo de correção dos testes, e diminuindo consequentemente a ocorrência de erros nesse processo, pois um sistema de avaliação por computador reduz os erros que podem ocorrer em processos de correção que utilizam scanners ópticos, além de eliminar possibilidade de erros de transcrição como as que ocorrem em testes que são corrigidos à mão (CLARÉS, 2008; COSTA, 2009; LINACRE, 2000; OLEA; ABAD; BARRADA, 2010; OLEA; HONTANGAS, 1999; OLIVEIRA, 2002; RUDNER, 1998; SANDS; WATERS, 1997; SUKAMOLSON, 2002).
- Monitoramento do teste e controle do tempo de exposição do item. Isso permite ao avaliador saber não só se o examinando acertou ou errou um item, mas quanto tempo ele dispensou em seu desenvolvimento (COSTA, 2009; OLEA; HONTANGAS, 1999; OLIVEIRA, 2002; WAINER, 2000a). A informação adicional do tempo de resposta ao item (*Item Response Times*) pode ser de interesse em alguns estudos (BRIDGEMAN; CLINE, 2000;

BRIDGEMAN; CLINE; HESSINGER, 2003; HORNKE, 2000; SCHNIPKE; SCRAMS, 1997; VAN DER LINDEN, 2006a; 2008a; 2008b; 2009; VAN DER LINDEN; GUO, 2008; VAN DER LINDEN; SCRAMS; SCHNIPKE, 1999; VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003).

- A "individualização" ou personalização do teste. O indivíduo terá um teste personalizado com questões adequadas a ele, segundo o seu desempenho durante o próprio teste (CLARÉS, 2008; COSTA, 2009; RUDNER, 1998). A aplicação de itens inadequados (muito fáceis ou muito difíceis) pode provocar comportamentos indesejáveis, por exemplo, o examinando pode "chutar" as respostas ou responder a mesma alternativa em todos os itens (LINACRE, 2000).
- Redução do tempo e do tamanho do teste (geralmente em 50%), mantendo o mesmo nível de confiança de um teste convencional. Isso reduz o desgaste e a fadiga do indivíduo que poderia prejudicar seu desempenho em testes longos, devido ao cansaço (CLARÉS, 2008; FAYERS; MACHIN, 2007; HAMBLETON; ZAAL; PIETERS, 1991; LINACRE, 2000; OLEA; ABAD; BARRADA, 2010; RENOM; DOVAL, 1999; SANDS; WATERS, WAINER, 2000b; 1997; SUKAMOLSON, 2002; TIAN et al., 2007; WEISS, 1985; WEISS; KINGSBURY, 1984; WISE; KINGSBURY, 2000).
- Possibilidade de aplicação do teste nas ferramentas de Educação à Distância, através da internet (BARTRAM; HAMBLETON, 2006; SALCEDO; PINNINGHOFF; CONTRERAS, 2005; WAINER, 2000a).
- Possui ótimo controle sobre as questões expostas e equilíbrio das avaliações para todos os níveis de habilidades.
- Gera relatórios rápidos que proporcionam ótimo retorno ao avaliado, pois o teste pode fornecer o resultado imediatamente após o término (FETZER et al., 2008; OLEA; ABAD; BARRADA, 2010; OLEA; HONTANGAS, 1999; RENOM; DOVAL, 1999; RUDNER, 1998; SMITH, 1994; SUKAMOLSON, 2002; VAN DER LINDEN; GLAS, 2010; TIAN et al., 2007; WAINER, 2000a; YEH, 2006). Estudos mostram que o fornecimento de um relatório de desempenho imediato é bem visto pelos examinandos (LILLEY; BARKER, 2007), principalmente para aqueles que não foram aprovados no teste (JULIAN, 1993).
- Melhoria na segurança do teste e maior rigidez no controle das regras do teste. Um exame feito pelo computador está menos sujeito

à burla de regras. Se um banco de itens é suficientemente grande, um examinando que tenha acesso a ele terá pequena vantagem sobre os demais. E ainda, há a possibilidade de criptografar os dados, de forma que somente o administrador do teste tenha a chave para decodificar as informações do banco (COSTA, 2009; FETZER et al., 2008; FERNANDES, 2009; OLEA; ABAD; BARRADA, 2010; OLEA; PONSODA, 1996; RENOM; DOVAL, 1999; RUDNER, 1998; TIAN et al., 2007; WAINER, 2000a).

- Não necessita fazer impressão de provas e, consequentemente, nem espaço físico e sigilo para o armazenamento das mesmas (FERNANDES, 2009).
- Novos itens podem ser adicionados, onde o processo de calibração se dá à medida que eles vão sendo administrados (durante o processo de calibração, eles não são utilizados para avaliar a proficiência) (RENOM; DOVAL, 1999; WAINER, 2000a).
- Itens que se tornarem obsoletos ou que começarem a apresentar problemas podem ser retirados do banco de dados a qualquer momento (WAY; DAVIS; FITZPATRICK, 2006).
- Estimativas mais precisas das habilidades dos indivíduos e com menos itens aplicados. Testes adaptativos ajustam adequadamente o nível de dificuldade das questões aos examinandos, sem prejudicar a acurácia das estimativas (OLEA; ABAD; BARRADA, 2010; OLIVEIRA, 2002; RENOM; DOVAL, 1999; SUKAMOLSON, 2002; WAINER, 2000a; WEISS; KINGSBURY, 1984).
- Os indivíduos podem se mostrar mais receptivos pelo fato das questões serem adaptadas conforme suas habilidades, e serem mostradas uma a uma, ao invés de todas ao mesmo tempo (SUKAMOLSON, 2002).
- A criação de itens em formatos multimídia (uso de *mouse*, sons, vídeos, etc.), os chamados itens inovadores (PARSHALL et al., 2010), o que torna o teste mais atrativo através da utilização de itens que não podem ser aplicados nos testes tradicionais (COSTA, 2009; FERNANDES, 2009; MUÑIZ, 1997; OLEA; ABAD; BARRADA, 2010; OLEA; PONSODA, 1998; RENOM; DOVAL, 1999; SANDS; WATERS, 1997; SUKAMOLSON, 2002).
- Aplicação simultânea em diferentes examinandos de diversos lugares do país e do mundo, seja por meio da internet ou de terminais instalados em locais fixos.
- Flexibilidade para realizar baterias de testes. Ao contrário do exame tradicional, um TAI não requer que todos os examinandos façam a

prova ao mesmo tempo. Em uma bateria de testes, por exemplo, o examinando que terminar a prova pode passar diretamente para a prova seguinte sem precisar aguardar os outros. Além disso, o administrador do teste pode fornecer as instruções do teste virtualmente (COSTA, 2009; VAN DER LINDEN; GLAS, 2010).

- Os testes podem ser realizados em ambientes mais confortáveis e com menos pessoas (VAN DER LINDEN; GLAS, 2010).
- A maioria dos examinados preferem responder um teste adaptativo do que um teste tradicional "papel e lápis" (VAN DER LINDEN; GLAS, 2010).
- Embora cada estudante possa responder a diferentes itens de uma mesma prova no TAI sob o suporte da TRI, os resultados são comparáveis entre si, pois os itens foram calibrados conjuntamente e estão na mesma métrica (OLEA; ABAD; BARRADA, 2010; WEISS, 1982; WISE; KINGSBURY, 2000).
- Permite acompanhar o desempenho do estudante ao longo do tempo, pois todos os itens estão na mesma métrica (THOMPSON, 2008; YEH, 2006).
- Um eventual erro no gabarito afeta a todos que respondem a uma prova tradicional, entretanto, nos TAIs, o impacto do erro é menor, pois apenas uma pequena parte dos respondentes terá sido submetida a esse item (LINACRE, 2000).

Entretanto, os TAIs não apresentam vantagens em todos os aspectos ou condições. Dentre as desvantagens ou limitações dos TAIs, destacam-se:

- Requer um Banco de Itens cuidadosamente calibrado, confiável, verificado (testado) e preferencialmente grande, para garantir segurança e confiabilidade (OLIVEIRA, 2002). Um banco com poucos itens pode resultar numa seleção de itens inadequados, fazendo com que o traço latente estimado do examinando se diferencie bastante do seu verdadeiro traço latente (RUDNER, 1998). Um banco com muitos itens necessita de uma amostra grande para o processo de calibração (WAINER; MISLEVY, 2000).
- Pode necessitar de consideráveis recursos financeiros, físicos (equipamentos, computadores, energia elétrica) e humanos para a sua organização e de muito tempo (implementação e manutenção). Os bancos de itens devem ser continuamente atualizados para garantir a segurança dos testes. Itens que não atendem mais às especificações e objetivos da avaliação, ou que foram utilizados constantemente em diferentes testes. devem ser eliminados do

banco de forma definitiva ou temporaria. Além disso, novos itens podem ser incluídos ao banco, o que aumenta o custo de implementação e operacionalização (COSTA, 2009; RENOM; DOVAL, 1999; MUÑIZ, 1997). No entanto, os benefícios de um TAI superam os custos envolvidos no seu desenvolvimento, na sua implementação e na sua manutenção (FETZER et al., 2008; HORNKE, 1999).

- No caso de um teste realizado pela internet, necessita de condições mínimas de segurança e de velocidade da conexão. Indivíduos mal-intencionados (por exemplo, *hackers*) podem tentar encontrar vulnerabilidades no sistema com o objetivo de "roubar" itens (FERNANDES, 2009). Além disso, deve existir uma forma de garantir que o indivíduo que está respondendo o teste é ele mesmo e não outro (LOZZIA et al., 2009; OLEA; ABAD; BARRADA, 2010; WAINER, 2000a).
- Limitação da quantidade de informação que pode aparecer na tela do computador e da sua capacidade gráfica e velocidade na apresentação do próximo item (RUDNER, 1998; SUKAMOLSON, 2002; WISE; KINGSBURY, 2000). Entretanto, para os computadores modernos, não existem mais limitações computacionais desse tipo para o TAI (VAN DER LINDEN; PASHLEY, 2010).
- Os indivíduos podem não estar familiarizados com o uso do computador ou do teclado (EIGNOR et al., 1998; SUKAMOLSON, 2002; TIAN et al., 2007; WISE; KINGSBURY, 2000). Indivíduos pertencentes às classes sociais mais baixas podem não ter acesso a computadores ou internet (SEGALL, 2005). Entretanto, essa é uma preocupação que deve desaparecer, ao longo do tempo, uma vez que a aquisição de computadores tem se tornado cada vez mais comum entre os diversos segmentos da população.
- Não podem ser aplicados a todos os assuntos ou habilidades. A maioria dos TAIs são baseados na TRI, a qual não é aplicável a todas habilidades e tipos de itens (CISAR et al., 2010; GIOUROGLOU; ECONOMIDES, 2005; RUDNER, 1998).
- Os respondentes podem se sentir injustiçados com a aplicação de questionários personalizados (RUDNER, 1998; OLIVEIRA, 2002; TIAN et al., 2007) e respondendo a quantidades diferentes de itens.
- Os respondentes podem ter uma impressão equivocada do seu nível de habilidade, possivelmente acreditando que seu desempenho foi

mediano, já que ele irá obter uma quantidade semelhante de acertos e erros (WAINER, 1993).

- Usualmente não é permitido ao respondente reavaliar ou revisar um item, ou seja, retornar a um item já respondido e alterar sua resposta (OLEA; PONSODA, 1996; OLIVEIRA, 2002; RUDNER, 1998; TIAN et al., 2007; WEISS; SCHLEISMAN, 1999), assim como não é permitido passar para a próxima questão sem responder a atual (WAINER, 2000a; WISE; KINGSBURY, 2000). Entretanto, os examinandos preferem ter a opção de alterar suas respostas nos testes adaptativos, uma vez que podem repensar suas respostas, bem como corrigir questões que tenham sido mal interpretadas (BOWLES; POMMERICH, 2001; VISPOEL, 1993; VISPOEL; ROCKLIN; WANG, 1994). No sentido de avaliar a possibilidade de revisão e alteração da resposta de um item já respondido, vários estudos têm sido elaborados (BOWLES; POMMERICH, 2001; LUNZ; BERGSTROM; WRIGHT, 1992; KINGSBURY, 1996; OLEA et al., 2000; PAPANASTASIOU, 2002; 2005; PAPANASTASIOU; RECKASE, 2007; STONE; LUNZ, 1994; VISPOEL, 1998b; WISE et al., 1994). No entanto, essa discussão é controversa e não há um consenso nos resultados (BOWLES; POMMERICH, 2001; WAY; DAVIS; FITZPATRICK, 2006; WISE; KINGSBURY, 2000), enquanto alguns argumentam que a revisão da resposta ao item pode prejudicar, diminuir a precisão das estimativas e aumentar o viés, além de permitir trapaças (GERSHON; BERGSTROM, 1995; STOCKING, 1997; VISPOEL et al., 1999; WAINER, 1993; WEISS; KINGSBURY, 1984; WISE, 1996; WISE; KINGSBURY, 2000), outros, ao contrário, acreditam que a revisão da resposta pode aumentar a precisão das estimativas e diminuir o viés (OLEA et al., 2000; PAPANASTASIOU, 2002; STONE; LUNZ, 1994; VISPOEL, 1998a; 1998b; VISPOEL; HENDERICKSON; BLEILER, 2000). O fato do examinando saber que pode rever um item, diminui a sua ansiedade e contribuiu para um melhor desempenho no teste (OLEA et al., 2000; WISE et al., 1994). Por outro lado, a revisão do item aumenta a complexidade do algoritmo do TAI (WAY; DAVIS; FITZPATRICK, 2006; WISE; KINGSBURY, 2000). Alguns TAIs permitem que a resposta de um item seja omitida e apresentam outro item com nível de dificuldade semelhante, ou ainda, permitem a alteração de alguma resposta, porém, como a habilidade é novamente estimada, todas as questões respondidas posteriormente a esse item podem ser perdidas (OLEA; PONSODA; PRIETO, 1999).

3.3. ALGORITMO DE UM TAI

A maioria dos TAIs utiliza uma estratégia que necessita estabelecer: a) um critério de partida, para determinar o primeiro item a ser apresentado, b) um método estatístico (bayesiano ou MV) para estimar a proficiência do indivíduo e a precisão associada, c) um procedimento para selecionar o próximo item, e d) um critério para finalizar o teste (OLEA et al., 2004).

A lógica da seleção dos itens no teste, em geral, acontece da seguinte forma: se o indivíduo acerta o item atual, o próximo item deverá ser de um nível mais difícil; se o indivíduo erra o item atual, o próximo item deverá ser de um nível mais fácil. Por mais simples que possa parecer, a ciência por trás do TAI não é nada simples, são utilizados uma série de análises complexas e algoritmos sofisticados (FETZER et al., 2008).

O desenvolvimento de um TAI é um processo trabalhoso e exige conhecimentos e técnicas importantes. Em primeiro lugar, necessita de um grande banco de itens devidamente calibrado através da TRI. Em segundo lugar, deve-se programar um conjunto de algoritmos para a seleção progressiva dos itens, para a estimação dos níveis de habilidade e sua respectiva precisão. Em terceiro lugar, um TAI deve submeter-se aos testes aplicados para garantir as propriedades desejáveis das estimações, assim como a sua precisão e validade. Se a aplicação do TAI for através da internet, ainda existe um trabalho adicional de programação para preservar a segurança do banco de itens e para realizar o processo de seleção do próximo item em um tempo imperceptível para o avaliado (OLEA et al., 2004).

A lógica de um teste adaptativo pode ser representada segundo a Figura 9, e compreende aos seguintes passos (CISAR et al., 2010; COSTA, 2009; DE AYALA, 2008; OLEA; ABAD; BARRADA, 2010; SEGALL, 2005; SIERRA-MATAMOROS, 2007; THISSEN; MISLEVY, 2000; VAN DER LINDEN; GLAS, 2000b; WISE; KINGSBURY, 1984):

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

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

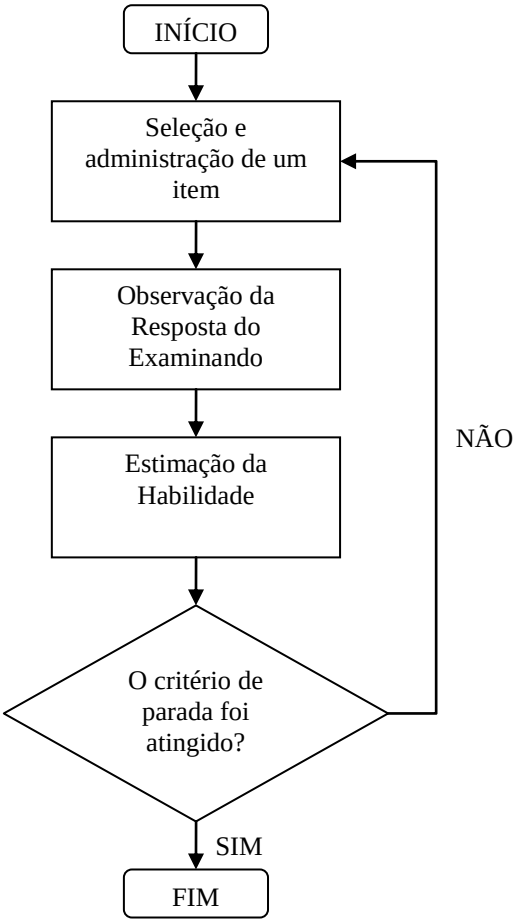


Figura 9. A Lógica de um TAI

A maior vantagem de um TAI em relação a um teste tradicional consiste na maior eficiência na estimação do nível de habilidade do respondente com um menor número de itens do que os testes tradicionais. Para atingir esse objetivo, dois aspectos básicos são

necessários: o método de estimação das habilidades dos respondentes e o critério de seleção dos itens, que serão discutidos nas Seções 3.3.5 e 3.3.4.

Conejo et al. (2001) e Segall (2005) definem os seguintes elementos básicos que devem compor os TAIs:

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

3.3.1. Modelo de Resposta ao Item

O primeiro passo é determinar qual modelo será utilizado no teste a ser construído, que depende do tipo de item. O Modelo de Resposta ao Item é uma função matemática que fornece a probabilidade de resposta correta condicionada ao nível do traço latente (SEGALL, 2005). Como visto na seção 2.1, existem diversos tipos de modelos para diversos tipos de aplicações. Também é possível montar um teste misto com diferentes tipos de itens e modelos (LINACRE, 2000), como pode-se verificar, por exemplo, no estudo de Oliveira (2002) e Tang e Eignor (1997). Diversos modelos da TRI têm sido utilizados em TAIs, por exemplo, o Modelo Logístico de Um Parâmetro (GERSHON, 2005; HALKITIS, 1998; JACOBUSSE; VAN BUUREN, 2007; KOCH; DODD, 1995; MORRISON; NUNGESTER, 1995; RILEY et al., 2007; WISNIEWSKI, 1986) o Modelo Logístico de Dois Parâmetros (MOREIRA JUNIOR; TEZZA; ANDRADE, 2010; MORRISON; NUNGESTER, 1995; SILVA; CURI, 2009), o Modelo Logístico de Três Parâmetro (COSTA, 2009; DE AYALA, 1989; 1992; EIGNOR et al., 1993; GUZMÁN; CONEJO, 2004; GUZMÁN; CONEJO; PÉREZ-DE-LA-CRUZ, 2007; PITON-GONÇALVES, 2004; PITON-GONÇALVES; MONZÓN; ALUÍSIO, 2009; SEGALL; MORENO; HETTER, 1997; TANG; WAY; CAREY, 1993; WAINER; BRADLOW; DU, 2000), o Modelo de Resposta Gradual (CHOI; SWARTZ, 2009; DE AYALA; DODD; KOCH, 1992; DODD; KOCH;

DE AYALA, 1989; GARCIA et al., 2000; GARCÍA-PÉREZ; ALCALÁ-QUINTANA; GARCÍA-CUETO, 2010; GARDNER et al., 2004; HOL; VORST; MELLENBERGH, 2005; 2008; HOU et al., 1996; PASSOS; BERGER; TAN, 2008; SILVA; CURI, 2009), o Modelo de Crédito Parcial (BAEK, 1995; CHEN; HOU; DODD, 1998; DAVIS; DODD, 2008; DAVIS et al., 2003; DE AYALA; DODD; KOCH, 1992; DODD; KOCH; DE AYALA, 1993; GORIN et al., 2005; KOCH; DODD, 1989; LANGE, 2008; PENFIELD, 2006; VAN KRIMPEN-STOOP; MEIJER, 2002), o Modelo de Crédito Parcial Generalizado (BURT; DAVIS; DODD, 2003; DAVIS, 2004; KOCALEVENT et al., 2009; LAU; WANG, 1998; 1999; LI; LI; WANG, 2010; PASTOR; DODD; CHANG, 2002; VAN RIJN et al., 2002; VELDKAMP, 2003; WANG; WANG, 2001), o Modelo de Escala Gradual (CHEN et al., 1997; DODD, 1990), o Modelo de Resposta Nominal (DE AYALA; SAVA-BOLESTA, 1999; PASSOS; BERGER; TAN, 2007), o Modelo de Desdobramento Graduado Generalizado (ROBERTS; LIN; LAUGHLIN, 2001; WANG; LIU, 2011) e Modelos Não-Paramétricos (XU; DOUGLAS, 2006).

A escolha adequada do modelo a ser utilizado é fundamental para a calibração dos itens. Estudos preliminares e simulações podem ajudar na escolha do modelo. O ajuste do modelo pode ser verificado mediante o cumprimento da suposição da dimensionalidade, a invariância dos parâmetros e a predição do modelo ajustado. Quanto à dimensionalidade, é recomendada a sua verificação quando for a primeira aplicação, quando se aplica a amostras diferentes, e quando se adicionam novos itens (WISE; KINGSBURY, 2000). Quanto à invariância das estimações do parâmetro da habilidade, pode-se obter as correlações entre os níveis de habilidade que se estimam para toda a amostra com duas subamostras distintas de itens, que devem ser próximas de 1. Quanto à invariância das estimações dos parâmetros dos itens, deve-se fazer a calibração do banco com duas subamostras distintas, onde as correlações entre os valores estimados de dificuldade deverá ser próxima de 1. Quanto ao ajuste do modelo, um dos procedimentos mais usuais (BOCK, 1972; MUÑIZ, 1997; BAKER, 2001; WISE; KINGSBURY, 2000) é comparar as CCIs empíricas e teóricas de cada item particularmente por meio de uma análise gráfica ou utilizando a distribuição Qui-quadrado, embora atualmente outros procedimentos têm sido utilizados, conforme mencionado na Seção 2.1. Além disso, o ajuste utilizando com a distribuição Qui-quadrado não se mostra adequado quando a amostra é grande, o que ocorre na maioria dos TAIs.

A maioria dos TAIs tem sido elaborada para medir proficiência, conhecimentos ou habilidades intelectuais e os modelos dicotômicos unidimensionais são os mais utilizados. Entretanto, os modelos multidimensionais também têm sido utilizados nos TAIs (ACKERMAN, 1996a; 1996b; BLOXOM; VALE, 1987; CURI et al., 2009; DAVEY; OSHIMA; LEE, 1996; DIAO; VAN DER LINDEN; YEN, 2001; FREY; SEITZ, 2009; 2011; GARDNER; KELLEHER; PAJER, 2002; GLAS; VOS, 2010; LEE; IP; FUH, 2008; HALEY et al., 2009; LI; SCHAFER, 2003a; 2004; 2005b; LUECHT, 1996; MCKINLEY; WAY, 1992; MULDER; VAN DER LINDEN, 2009; 2010a; PETERSEN et al., 2006; RECKASE, 2007; 2009; RILEY et al., 2010; ROUSSOS; STOUT, 1996; RULISON; LOKEN , 2009; SEGALL, 1996; 2000; 2001; 2010; SPRAY et al., 1997; STOUT et al., 1996; THOMAS, 1990; TSENG; HSU, 2001; VAN DER LINDEN, 1996; 1998b; 1999b; VAN DER LINDEN; HAMBLETON, 1996; VELDKAMP; VAN DER LINDEN, 2002; 2008a; WANG; CHANG, 2011a; WANG; CHANG; BOUGHTON, 2011; WANG; CHEN, 2004; WANG; WANG, 2001; WRIGHT, 1988; YAN; LEWIS; STOCKING, 2004) para avaliar o rendimento que depende de habilidades múltiplas, que podem ou não estarem relacionadas entre si.

Os modelos multidimensionais se dividem em dois grupos: Modelos Não Compensatórios (a probabilidade de acerto é um produto de probabilidades) e Modelos Compensatórios (a probabilidade de resposta é função de uma combinação linear de dimensões). Um exemplo de Modelo Não Compensatório é a avaliação da auto-eficácia, onde uma dimensão é a cognitiva e outra é a afetividade. Nesse caso, um indivíduo que tem alto desempenho cognitivo pode ter baixa afetividade, ou seja, uma dimensão não compensa a outra. Somente aqueles que possuem um alto desempenho nas duas dimensões tem uma probabilidade alta de resposta. Um exemplo de Modelo Compensatório é a proficiência em Matemática, onde um indivíduo com baixa proficiência em trigonometria e alta proficiência em álgebra e em compreensão de texto, pode ter a probabilidade de acerto da questão aumentada, ou seja, uma dimensão compensa a outra (DE AYALA, 2008). Hontangas et al. (2008a) destaca que os Modelos Não Compensatórios são mais complexos e pouco utilizados.

Os modelos multidimensionais permitem avaliar o rendimento em tarefas complexas, incluir conteúdos diferentes e estimar simultaneamente a habilidade nos distintos traços latentes, geralmente são mais eficientes que os modelos unidimensionais (reduzem em até um terço o tamanho do teste) e produzem estimativas com melhores

propriedades. As suas desvantagens incluem a necessidade de grandes amostras para a estimação dos parâmetros e todas as especificações que necessitam, por exemplo, número de dimensões, o tipo de dimensionalidade dos itens (intra-item e inter-itens), a interação entre as dimensões (modelos compensatórios ou não compensatórios), as correlações entre os traços latentes (HONTANGAS et al., 2000a).

Segall (2001) comparou várias condições de TAIs multidimensionais via simulação onde, segundo o autor, é possível conseguir uma medida quase perfeita para a habilidade geral. Além disso, o autor mostra que também pode-se utilizar os modelos multidimensionais para medir uma única dimensão. Como uma alternativa para medir o traço latente multidimensional, Schnipke e Green (1995) propuseram a construção dos chamados mini-TAIs, que consistem em dividir o banco de itens em tantos bancos unidimensionais quanto o necessário, de forma que cada TAI seleciona itens em um deles. Dessa forma, esses autores consideram que a pontuação final pode ser obtida, por exemplo, pela média da habilidade estimada em cada TAI. Entretanto, essa forma de pontuação não é aconselhada, pois uma dificuldade seria a interpretação da escala obtida. Outra questão que desfavorece o uso dos mini-TAIs é que alguns bancos de itens podem ficar muito pequenos de modo a aumentar os problemas relacionados com a superexposição dos itens.

Conforme Segall (2005), num TAI multidimensional, os itens são selecionados para maximizar a informação em várias dimensões simultaneamente. Segundo Frey e Seitz (2009), um TAI multidimensional pode apresentar de 30% a 50% menos itens na sua aplicação em comparação a um TAI unidimensional.

Têm se provado também a viabilidade dos TAIs com a utilização de modelos politômicos, que têm a vantagem de proporcionar mais informação do que os modelos dicotômicos para estimar a habilidade (LINACRE, 2000). Segundo Dodd et al. (1995), os TAIs politômicos podem proporcionar bons níveis de precisão da habilidade mesmo que o banco contenha poucos itens. De Ayala (1989) considera intuitivo presumir que o indivíduo possui um conhecimento parcial da resposta correta do item e que ele usa esse conhecimento incompleto para escolher uma particular alternativa incorreta. O uso dos modelos dicotômicos no TAI ignora essa informação parcial na sua estimação da habilidade. De fato, a habilidade será estimada com maior precisão se o conhecimento parcial for utilizado do que se for ignorado. Isto é, a habilidade estimada para um respondente, que tem um conhecimento parcial suficiente para selecionar uma alternativa que é mais atrativa

para indivíduos com habilidade superior, pode ser maior do que a habilidade estimada para um respondente que seleciona uma alternativa que é mais atraente para quem tem habilidade baixa. Além disso, os TAIs politômicos possuem a vantagem de administrar tanto itens politômicos quanto dicotômicos.

De Ayala (1989) fez um estudo comparativo entre um TAI baseado no Modelos de Resposta Nominal (TAI-MRN) e um TAI baseado no Modelo Logístico de 3 Parâmetros (TAI-ML3). Ambos TAIs mostraram um pequeno viés na sua estimativa da habilidade na região da curva de habilidade onde seu respectivo modelo não proporcionou uma quantidade grande de informação. Porém, apesar da assimetria positiva da função obtida pelo TAI-MRN, foi possível obter estimativas de habilidade relativamente precisas no intervalo superior de habilidade.

3.3.2. Banco de Itens

Os bancos de itens, definidos na Seção 2.3 já eram utilizados antes da existência dos TAIs. Entretanto, o advento da informática permitiu um melhor desempenho na construção, na administração e no armazenamento dos itens (OLIVEIRA, 2002), além da possibilidade da elaboração dos tipos inovadores de itens, que utilizam os recursos multimídia do computador (PARSHALL; DAVEY; PASHLEY, 2000). O surgimento da TRI também contribuiu para o desenvolvimento dos bancos de itens, adicionando informações psicométricas (parâmetros dos itens) que descrevem mais precisamente as características de cada item (HAMBLETON; SWAMINATHAN, 1985).

Vários estudos têm sido feitos sobre a elaboração e a calibração de um banco de itens para um TAI (ABAD et al., 2004; ARIEL; VAN DER LINDEN; VELDKAMP, 2006; ARIEL; VELDKAMP, 2006; ARIEL; VELDKAMP; VAN DER LINDEN, 2004; BAN et al., 2000; BARBERO, 1999; BARRADA; ABAD; OLEA, 2011; BARRADA; OLEA; ABAD, 2008; BARRADA et al., 2011; BELOV; ARMSTRONG, 2005; 2009; BJORNER; KOSINSKI; WARE JUNIOR, 2003; BJORNER et al., 2007; BOEKKOOI-TIMMING, 1991; BREITHAUPT; ARIEL; HARE, 2010; CHANG; LU, 2009; CHANG; VAN DER LINDEN, 2003; DAVEY; OSHIMA; LEE, 1996; DODD; KOCH; DE AYALA, 1993; EGGEN, 2004; FLAUGHER, 2000; GARCÍA-PÉREZ; ALCALÁ-QUINTANA; GARCÍA-CUETO, 2010; GLAS, 2000; 2010; GLAS; VAN DER LINDEN; GEERLINGS, 2010; GEORGIADOU; TRIANTAFILLOU; ECONOMIDES, 2006; GORIN et al., 2005; GU; RECKASE, 2007; HALEY et al., 2009b; HALKITIS,

1996; 1998; HARMES; KROMREY; PARSHALL, 2001; HETTER; SEGALL; BLOXOM, 1997; MILLMAN; ARTER, 1984; MOLINA; PAREJA; SANMARTIN, 2008; PARSHALL; DAVEY; PASHLEY, 2000; PARSHALL et al. (2010); RECKASE, 2003; RECKASE; HE, 2003; SEGALL; MORENO; HETTER, 1997; POMMERICH, 2007; STOCKING, 1994; STOCKING; SWANSON, 1998; THISSEN et al., 2007; VAN DER LINDEN; ARIEL; VELDKAMP, 2006; VAN DER LINDEN; GLAS, 2000a; 2001; VAN DER LINDEN; VELDKAMP; REESE, 2000; VELDKAMP; VAN DER LINDEN, 2000; 2010; WAINER; MISLEVY, 2000; WEISS, 1982; WRIGHT; BELL, 1984; YAO, 1991). Essa seção apresentará algumas características básicas para a elaboração de um banco de itens.

Fetzer et al. (2008) consideram a criação do banco de itens como a etapa mais longa no desenvolvimento de um TAI. Segundo Costa (2009), a criação do banco de itens compreende quatro etapas: (1) a construção dos itens, (2) a revisão pedagógica dos itens, (3) a pré-testagem e a calibração dois itens por meio de um modelo da TRI e (5) a incorporação das informações psicométricas (parâmetros dos itens) no banco de itens.

3.3.2.1. A Construção dos Itens

A criação dos itens é realizada através dos mesmos procedimentos teóricos abordados na TCT, o que pode ser consultado em diversas literaturas (ANASTASI, 1977; FAYERS; MACHIN, 2007; GARRET, 1979; GULLINKSEN, 1950; MIGUEL, 1974; OSTERLIND, 1997; PASQUALI, 1996; 1998; PRIETO; DELGADO, 1996; WILSON, 2005; etc.). Os itens devem ser elaborados por especialistas no assunto do teste. Entretanto, dentro do contexto dos TAIs, outros cuidados devem ser levados em conta.

Para preservar a condição de independência local, deve-se evitar o uso de enunciados compartilhados e cuidar na elaboração do item, de modo que um item não proporcione "pistas" para responder corretamente outro item. Caso existam itens no banco de itens com evidentes relações entre si, devem-se estabelecer certas restrições no algoritmo de seleção para que eles não sejam apresentados a um mesmo indivíduo.

Segundo Flaugher (2000), deve-se observar se os itens não apresentam funcionamento diferenciado (*Differential item functioning* -

DIF[11]) baseado em outras características específicas do examinado, tais como gênero ou etnia. Existem diversos métodos para detectar DIF em TAIs (NANDAKUMAR; ROUSSOS, 1997a; 1997b; 2004; WALKER; BERETVAS; ACKERMAN, 2001; ZWICK; THAYLER, 2002), sendo que uma revisão sobre esses métodos foi realizada por Zwick (2000; 2010). Entretanto, muitas vezes, um item com DIF só é detectado no processo de calibração.

Na TRI, existem vários formatos de itens que podem ser utilizados: dicotômico, politômico nominal, escala gradual, tipos mesclados e item aberto (com posterior categorização). A definição do tipo de item é muito importante, pois afetará na escolha e adequação do modelo de resposta e no seu desempenho (RENOM; DOVAL, 1999). No caso de avaliações educacionais ou de proficiência, os itens abertos reduzem praticamente para zero a probabilidade de um acerto casual, porém são mais trabalhosos na posterior categorização. Itens com alternativas múltiplas possibilitam o acerto casual, porém são mais fáceis de administrar, analisar, modificar (por exemplo, mantendo o enunciado e modificando as alternativas) e de implementar as técnicas de GAI. Itens do tipo verdadeiro ou falso (V ou F), de ordenação e associação são muito pouco utilizados em TAIs.

O formato do item depende do objetivo do teste. Quando se pretende medir o rendimento máximo (por exemplo, proficiência), o mais usual é utilizar um formato de resposta de escolha múltipla. Abad, Olea e Posoda (2001) estudaram o número ótimo de opções que devem ter os itens e concluíram que tanto os índices da TCT como os da TRI se mantêm em níveis aceitáveis quando se elaboram três boas alternativas de resposta. Quando o objetivo é uma medição que não seja de rendimento (por exemplo, atitude ou preferência) utiliza-se um formato de resposta de categorias ordenadas.

Para a elaboração de um TAI, é necessário possuir um grande banco de itens com boa discriminação (HAMBLETON; ZAAL; PIETERS, 1991) e que seja diversificado, isto é, contendo itens fáceis, médios e difíceis, para cada conteúdo do teste (FLAUGHER, 2000), se possível. Isso porque cada indivíduo possui uma habilidade e devem existir itens suficientes para estimar com eficiência os mais diversos graus de proficiência. Alguns autores (BERGSTROM; LUNZ, 1999)

[11] DIF (Funcionamento Diferencial do Item) é o termo utilizado para um item quando o mesmo funciona diferentemente para grupos com distintas características sociodemográficas. Em termos da TRI, um item terá DIF se para valores iguais de θ os grupos considerados possuírem diferentes CCIs para esse item (ANDRIOLA, 2000).

aconselham que a distribuição da dificuldade dos itens seja semelhante ao nível de habilidade da população. Way, Davis e Fitzpatrick (2006) consideram que a maioria dos itens deve ser de dificuldade média, ou seja, itens que possam ser respondidos corretamente por 60% a 70% da população. Por outro lado, Flaugher (2000) e Segall (2005) destacam que um banco de itens adequado para um teste adaptativo deve conter itens com alta discriminação, uma distribuição uniforme para o parâmetro de dificuldade e baixa probabilidade de acerto casual (menor do que 0,20). Nos testes com objetivo de classificação, onde o critério de seleção dos itens é baseado no ponto de corte, necessita-se de muitos itens com parâmetro de dificuldade na região do ponto de corte (THOMPSON; PROMETRIC, 2007).

Três fatores motivam o aumento da quantidade de itens em um TAI (WISE; KINGSBURY, 2000): a melhoria do desempenho dos testes tradicionais nos últimos anos, a quantidade de restrições no algoritmo de seleção de itens, e a quantidade de aplicações do TAI. Outra questão que influencia na quantidade de itens é se os itens são sigilosos ou não. Em testes onde os itens não são sigilosos (que medem, por exemplo, atitude, diagnóstico médico), o banco de itens não precisa ser tão grande, pois não há a necessidade de controlar a exposição do item. Já em testes onde os itens são sigilosos (que medem, por exemplo, proficiência), o banco de itens precisa ser grande. Renom (1993) salienta que, apesar de alguns autores recomendarem um mínimo de 100 itens, o ideal é que se tenha mais de 500. Alguns autores consideram que deve-se ter 6 a 8 vezes mais itens no banco em relação à quantidade de itens aplicados no TAI (SEGALL, 2005), outros de 5 a 10 vezes, e outros aconselham que se tenha no mínimo 10 vezes mais itens (OLEA; PONSODA; PRIETO, 1999). Weiss (1985) destaca que um teste com 15 a 20 itens necessita de um banco com pelo menos 100 itens com boa qualidade, sendo que o ideal seria ter entre 150 e 200 itens. Van der Linden e Glas (2000b) destacam que, na prática, um banco de itens contém de 7 a 10 vezes a quantidade de itens do teste, porém Stocking (1994) considera que o banco de itens deve ser 12 vezes a quantidade de itens de um TAI. Muitos TAIs possuem bancos com mais de 1000 itens (WISE; KINGSBURY, 2000), por exemplo, o ASVAB que começou com um banco com 4000 itens, e o CARAT (*Computerized Adaptive Reporting and Testing*) que possui 6500 itens calibrados. Estudos mostram que, na prática, é mais problemático ter um banco de itens pequeno do que uma amostra pequena de respondentes para calibrá-lo (LANGE, 2008). Dessa forma, recomenda-se que seja elaborada uma quantidade de itens maior do que a necessária, já que uma parte dos

itens provavelmente será descartada no processo de calibração (SEGALL, 2005). Uma revisão de vários trabalhos sobre tamanho do banco de itens pode ser encontrada em Renom (1993).

Quando há diferentes categorias de conteúdo, convém que em cada uma haja um número parecido de itens (ou proporcional a quantidade de restrições do algoritmo), caso contrário, os itens que pertencem as categorias menos numerosas podem se sobre-expor.

Wainer (2000c) afirma que, para um teste que é aplicado várias vezes ao ano, o tamanho do banco de itens deve aumentar exponencialmente para garantir a segurança e o sigilo dos itens. O autor considera que essa estratégia não é economicamente viável e oferece dois algoritmos alternativos que podem ser encontrados em Wainer (2000c).

A técnica GAI, apresentada anteriormente na Seção 2.3, também tem sido utilizada na elaboração de itens para TAIs (BEJAR et al., 2003; EMBRETSON, 1999; GLAS; VAN DER LINDEN, 2003; GLAS; VAN DER LINDEN; GEERLINGS, 2010; PITONIAK , 2002; REVUELTA, 2000; REVUELTA; PONSODA, 1998b; 1999; VAN DER LINDEN, 2008a). Revuelta e Ponsoda (1998b) criaram um TAI de análise lógica baseado em um GAI e verificaram que a geração de itens pode ser útil para melhorar a eficácia de um teste contanto que a dificuldade desses itens seja predita adequadamente. Segundo Tejada (2001), os modelos geradores de itens também podem aumentar a durabilidade do banco de itens, reduzindo-se as despesas relacionadas com a manutenção do banco. No contexto dos TAIs, Tejada (2001) destaca duas vantagens em aplicar a GAI:

- Permite melhorar a adaptação do teste para cada indivíduo. Com a GAI, não existe um banco pré-fixado de itens, mas tenta-se construir, durante o teste (EMBRETSON, 1999), para cada indivíduo, aquele item que seja o mais informativo para seu nível de habilidade entre todos os possíveis itens que podem ser gerados.
- É uma solução para controlar a taxa de exposição dos itens. A GAI permite elaborar uma quantidade de itens muito grande, o que aumenta as possibilidades de construção de diferentes testes.

Revuelta (2000) estudou, através de simulação, o efeito da imprecisão com que são gerados os itens isomorfos na confiabilidade do teste pela pontuação de um indivíduo num TAI. Utilizou-se um banco de itens elaborados e calibrados segundo um ML3 e analisou-se a precisão do TAI. Os principais resultados mostraram que:

- O erro no processo de criação dos itens isomorfos não produziu estimativas viciadas da habilidade em nenhuma das situações propostas;
- O aumento do percentual de itens isomorfos no teste produziu um aumento no erro padrão da estimativa da habilidade.
- A informação diminui à medida que se aumenta o erro nos itens isomorfos e o percentual deles no banco. O processo de criação de itens isomorfos introduz uma nova fonte de erro nos parâmetros dos itens, pois a manipulação do conteúdo pode afetar as propriedades dos itens de uma forma não controlada. A informação de um teste com 25% de itens isomorfos é aproximadamente 80% da informação de um teste sem itens isomorfos. Essa informação cai para: 60% em um teste com 50% de isomorfos, 40% em um teste com 75% de isomorfos, e 18% para um teste com 100% de isomorfos.

3.3.2.2. A Revisão Pedagógica dos Itens

A revisão pedagógica dos itens também deve ser feita por especialistas no assunto do teste. Verifica-se a quantidade de questões elaboradas em cada conteúdo do teste e o grau de dificuldade teórico (antes da estimação dos índices de dificuldade) e se os itens estão escritos de maneira que sejam compreendidos pelo público alvo de respondentes.

Quanto melhor for a qualidade do banco, melhor será a tarefa que o algoritmo adaptativo poderá utilizar. Entretanto, se a qualidade não for boa, nem o melhor algoritmo poderá trazer bons resultados para o TAI (FLAUGHER, 2000).

3.3.2.3. A Pré-Testagem e Calibração dos Itens

O objetivo da pré-testagem é realizar uma análise psicométrica dos itens antes da aplicação do TAI verificando o nível de dificuldade das questões, se os itens possuem bons parâmetros de discriminação, entre outros quesitos, relacionados com a qualidade dos itens. Na pré-testagem, são utilizados tanto os critérios da TCT e da TRI.

Para isso, é necessário obter uma amostra significativa de respondentes para a aplicação dos itens por meio de um teste que pode ser realizado no formato tradicional "papel e lápis" (numa versão informatizada, caso tenha itens multimídia) e com a utilização de técnicas como BIB, conforme Seção 2.3.

Antes de analisar os itens, devem-se remover os dados com anomalias, tanto de itens quanto de respondentes. Antes da calibração, devem-se eliminar os padrões de respostas inapropriados (indivíduo que acerta ou erra tudo ou que responde sempre a mesma alternativa ou de maneira aleatória, item respondido correta ou incorretamente por todos) e que possam ter sido danificadas no processo de tabulação dos dados. Esses cuidados são particularmente essenciais quando os itens são sigilosos, por exemplo, em avaliações educacionais.

Na TCT, um indivíduo responde a um conjunto de questões, obtendo, no final, uma nota, denominada escore (bruto ou padronizado), que é a soma das respostas corretas e expressa a magnitude do que se desejava medir no sujeito, ou seja, indivíduos com maior escore são mais hábeis naquela área de conhecimento do que aqueles com menor escore (PASQUALI, 1996; 1998). Outros critérios da TCT podem ser utilizados, tais como os índices de discriminação e de dificuldade, a correlação bisserial e o Alfa de Crombach.

Outra questão a ser verificada é a dimensionalidade dos dados, ou seja, das respostas dos itens, como discutido na Seção 2.1. Deve-se verificar se a dimensionalidade obtida pelos dados é a mesma que foi determinada na definição do(s) traço(s) latente(s). Se o traço latente é unidimensional, os dados devem ser unidimensionais, se o traço latente é bidimensional, os dados devem ser bidimensionais, e assim por diante. Caso a dimensão não se confirme, deve-se verificar se existem itens problemáticos que estejam causando esse problema, ou então rever a dimensionalidade do traço latente. Alguns autores têm estudado os efeitos da violação das suposições de unidimensionalidade (FOLK; GREEN 1989; MARTÍNEZ-CARDEÑOSO et al., 1996; MISLEVY; CHANG, 2000) e de independência local (REESE, 1999; SPRAY; PARSHALL; THOMAS, 1997) nos TAIs.

A TCT apresenta algumas limitações, como, por exemplo, ser dependente do conjunto particular de itens que compõem a prova. Dessa forma, como visto no Capítulo 2, a TRI pode ser utilizada para suprir essas limitações sem entrar em contradição com os princípios da TCT. A analise psicométrica por meio da TRI é discutida no Capítulo 2. No contexto dos TAIs, é recomendado que a estimativa do parâmetro de discriminação (no caso dos modelos logísticos) seja alto, e que seu erro padrão seja baixo, a fim de eliminar itens com baixa informação que podem causar erros na estimativa de proficiência e nas decisões de classificação de examinandos (WISE; KINGSBURY, 2000).

Deve-se observar e eliminar itens problemáticos, por exemplo, que apresentam DIF, que são acertados por indivíduos com baixa

proficiência e são errados por indivíduos com alta proficiência ou com erro de gabarito (apresentam correlação bisserial baixa ou negativa), que possuem parâmetro de dificuldade fora do intervalo considerado da escala, que possuem discriminação baixa ou que possuem elevado erro padrão das estimativas dos parâmetros dos itens.

Na fase de construção do banco de itens, o principal interesse consiste na estimação dos parâmetros (processo chamado de calibração) dos itens pela TRI, processo discutido nas Seções 2.3 e 2.4. Piton-Gonçalves (2004) salienta que um TAI baseado na TRI necessita de um banco de itens previamente calibrado, com exceção do algoritmo CBAT-2 (*Content-Balanced Adaptive Testing*), proposto por Huang (1996a; 1996b) que autocalibra os parâmetros dos itens no decorrer do teste.

Outra questão importante é o tamanho da amostra necessário para a calibração do banco, discutido na seção 2.3, que depende do modelo de resposta ao item e da quantidade de itens (número de parâmetros a serem estimados) e das respostas dos respondentes (quantidade suficiente de resposta em cada categoria).

Após a calibração dos itens, deve-se verificar se a quantidade de itens que permaneceram no banco de itens é suficiente para a aplicação do TAI, se abrangem todos os conteúdos do teste e se estão distribuídos adequadamente em toda extensão do traço latente avaliado (itens fáceis, medianos e difíceis). Caso seja necessária uma elaboração de novos itens, deve-se verificar, por meio de simulações, se é possível começar a aplicação do TAI sem que haja um comprometimento significativo do teste e, assim, esses itens podem ser adicionados e calibrados gradativamente. Se a ausência desses itens compromete a validade do teste, é necessário elaborar mais itens e realizar a pré-testagem e a calibração (pode-se utilizar métodos de equalização, conforme Seção 2.3) novamente, antes de aplicar o TAI.

Ainda que a quantidade teoricamente necessária de itens para o TAI não tenha sido alcançada, estudos de simulação considerando o algoritmo e as restrições do TAI podem ser feitos para verificar a eficiência e a validade do banco de itens e possivelmente adequá-lo para o teste, por exemplo, diminuindo a quantidade de restrições e a complexidade do algoritmo (SEGALL, 2005).

3.3.2.4. A Incorporação das Informações Psicométricas

As informações psicométricas obtidas na calibração do banco de itens, ou seja, os parâmetros dos itens (discriminação, dificuldade e acerto casual, no caso do ML3), devem ser incorporadas ao banco de

itens, assim como qualquer outra informação que pode ser considerada importante, por exemplo, os índices de discriminação e dificuldade e a correlação bisserial obtidos pela TCT.

3.3.2.5. Manutenção do Banco de Itens

Um banco de itens grande e com itens de qualidade deve refletir fielmente a natureza do domínio do conhecimento a ser medido, seus itens devem admitir um padrão de conteúdo válido, e deve ser de fácil uso e manutenção. O estudo de procedimentos eficientes para a manutenção e a renovação do banco de itens é um dos principais desafios para que o TAI seja aplicado com êxito (ABAD *et* al., 2010). Segundo Oliveira (2002), após a construção e calibração de um banco de itens, esse deve permitir a incorporação de novas informações, como a taxa de exposição do item, que está relacionada com a frequência que um item é administrado em um teste.

A inclusão de novos itens no banco e a exclusão de itens antigos fazem parte do processo de manutenção de itens. A calibração dos novos itens se dá da seguinte forma: o item é administrado durante um TAI, mas não interfere na proficiência do examinando, sendo que a sua resposta será utilizada apenas para a calibração do item (SEGALL, 2005; THISSEN; MISLEVY, 2000). Os novos itens são incluídos e colocados na escala existente através de um procedimento de ligação (WISE; KINGSBURY, 2000).

O processo de eliminação de itens de um banco deve ser feito gradativamente, pois a remoção abrupta de vários desses itens pode comprometer a eficiência do teste. Os itens também podem ser eliminados por se tornarem obsoletos ao longo do tempo, por não estarem funcionando como deveriam, ou por se tornarem conhecidos. Wise e Kingsbury (2000) salientam que, no longo prazo, é melhor rever e retestar esses itens, mesmo que seja um processo mais caro. Alguns itens se tornam conhecidos quando são aplicados com muita frequência e em testes que são realizados várias vezes ao ano. Muitas vezes, não há computadores suficientes para todos os examinandos realizarem o teste ao mesmo tempo, assim, os examinandos que realizarem o teste posteriormente poderão ter um conhecimento a mais sobre itens que poderão ser administrados a eles, devido à troca de informações com examinados que realizaram o teste mais cedo. Consequentemente ocorre uma alteração nas informações psicométricas do item: ele se torna mais fácil, o poder de discriminação diminui e o acerto casual praticamente não existe.

Outra situação que leva à eliminação de itens é quando há a necessidade de divulgação dos itens para o público, já que alguns programas de avaliação exigem a divulgação dos itens ao final de um período de testes (WAY; DAVIS; FITZPATRICK, 2006).

Além da inclusão ou exclusão de novos itens no banco, pode ser feita uma atualização do banco de itens (OLIVEIRA, 2002). As novas respostas adquiridas em testes aplicados após a calibração do banco podem servir para atualizar as informações contidas no banco de itens, como, por exemplo, o parâmetro de dificuldade do item. Porém, isso implica em repetir o processo de calibração de todo o banco de itens.

Tejada (2001) menciona que os itens de um banco devem ser desmanchados a cada nove meses aproximadamente, para que eles não sejam demasiadamente expostos.

Caso não tenha sido desenvolvido um banco de itens para o TAI, é possível implantar uma versão limitada de TAI utilizando-se os itens existentes do teste convencional (WEISS, 1982; WEISS; KINGSBURY, 1984). Isso é uma situação comum, pois muitos TAIs surgiram a partir de uma transformação dos testes tradicionais (HOL; VORST; MELLENBERGH, 2008; WAINER, 1993; WAINER; MISLEVY, 2000; WALTER; HOLLING, 2008).

3.3.2.6. Segurança do Banco de Itens

A segurança é fundamental em testes onde os itens são sigilosos, pois se eles forem conhecidos de antemão, a validade do teste estará comprometida (WISE; KINGSBURY, 2000). Diversos estudos que se preocupam com a integridade e a segurança do banco de itens (BARBERO, 1999; BARRADA; OLEA; ABAD, 2008; BARRADA et al., 2011; DAVEY; NERING, 2002; FERNANDES, 2009; HAN, 2003; STOCKING, 1994; STOCKING; SWANSON, 1998; VELDKAMP; VAN DER LINDEN, 2000; VEERKAMP; GLAS, 2000; WAY, 1998; WAY; ZARA; LEAHY, 1996; WISE; KINGSBURY, 2000; YI; ZHANG; CHANG, 2008; ZHU; FAN, 1999). Way, Davis e Fitzpatrick (2006) destacam os seguintes aspectos relacionados à segurança: (1) os dados do banco de itens devem ser criptografados e a transmissão entre o servidor e os computadores deve ser monitorada; (2) o acesso ao sistema deve ser seguro e protegido com senha e o examinando não deve ter acesso a outros recursos do computador; e (3) o sistema deve ser capaz de retomar ao mesmo item que estava sendo exibido caso ocorra uma ruptura, tais como falta de energia ou problemas na rede.

Nos TAIs, as provas não são impressas, consequentemente, não há o risco de elas serem roubadas. Entretanto, existem outras formas de

"roubo" de item, tais como o uso de dispositivos eletrônicos (pagers, câmeras "espiãs" disfarçadas em relógios ou canetas, transmissores ou gravadores de vídeo) e a invasão de computadores ligados em rede (COLTON, 1998).

Nos TAIs que se aplicam muitas vezes (em larga escala), pode-se estabelecer múltiplos bancos de itens que vão sendo utilizados de forma rotatória (ARIEL; VELDKAMP; VAN DER LINDEN, 2004; DAVEY; NERING, 2002; STOCKING; LEWIS, 1998; SWANSON; STOCKING, 1993b; WISE; KINGSBURY, 2000). Essa é uma forma de contribuir com a segurança do banco de itens e controlar a exposição dos itens (vide Seção 3.4).

3.3.2.7. Tratamento dos Dados Omitidos na Calibração do Banco de Itens para o TAI

Uma questão que deve ser levada em conta na calibração do banco de itens é a presença ou não de dados omitidos, ou seja, itens que deveriam ter sido respondidos pelos indivíduos, mas, por algum motivo (não quis responder, não teve tempo, esqueceu, não viu o item), não foram respondidos (ABAD et al., 2004; EGGEN; VERSCHOOR, 2011; HARMES; KROMREY; PARSHALL, 2001; MISLEVY; WU, 1988; VAN DER LINDEN, 1999).

Abad et al. (2004) apresentaram um método de calibração que, aplicado às respostas com omissões, produzisse resultados válidos nas condições em que se aplicam o TAI, onde não se permite omitir respostas. Para tanto, na coleta dos dados, os respondentes foram orientados a omitir (não responder) a questão caso não entendessem ou não estivessem seguros da resposta correta, o que produziu um grande número de respostas omitidas. Eles foram advertidos que perderiam pontos pelas respostas erradas. Com a possibilidade de omissão de respostas, espera-se que sujeitos com baixa proficiência omitam muito e acertem pouco. No TAI convencional, os sujeitos com baixa proficiência são obrigados a responderem o item e espera-se que a probabilidade de acertá-lo seja inversamente proporcional à quantidade de alternativas. Por esse motivo, os parâmetros estimados com omissão podem não ser válidos para o TAI.

Abad et al. (2004) trataram os dados omitidos de duas formas, a fim de verificar o efeito nas calibrações dos itens: 1) tratamento das respostas omitidas como erradas, e 2) tratamento das respostas omitidas como parcialmente corretas. No tratamento das respostas omitidas como erradas, essa calibração seria adequada para um TAI que permite omissões. Isso evita erros relacionados com o acerto casual. Porém, as

instruções devem ser seguidas estritamente pelo respondente para o TAI funcionar adequadamente. Os resultados apresentaram 13 itens desajustados e um valor médio para o parâmetro *c* de 0,10, devido ao elevado número de omissões. No tratamento das respostas omitidas como parcialmente corretas, as omissões se substituem nas equações de verossimilhança de estimação das habilidades por uma probabilidade igual ao inverso do número de alternativas. Isso equivale a considerar que se os sujeitos tivessem respondido ao item teriam tido uma probabilidade de acerto igual a essa. Esse tipo de calibração é adequado para um TAI convencional, onde todos os indivíduos são obrigados a responder a todos os item aplicados. Se o sujeito tivesse respondido ao item omitido, o teria feito ao acaso. Nesse tratamento, as prioris (valor inicial) dos parâmetros *a* e *c* foram escolhidas de tal forma que o desvio padrão deles fosse menor do que no primeiro tratamento para minimizar a possibilidade de valores extremos. Os resultados apresentaram 18 itens desajustados e um valor médio para o parâmetro *c* de 0,21. Esse método produz valores de *c* mais altos e próximos do valor esperado da probabilidade de acerto casual.

Com base nesse experimento, os autores fizeram uma simulação para investigar se o TAI é capaz de recuperar parâmetros da habilidade de sujeitos com baixa proficiência, já que a diferença mais importante dos métodos de calibração foram os valores estimados do parâmetro *c*. Se estabeleceram dois fatores: tratamento das respostas omitidas (como erradas ou parcialmente corretas) e a quantidade de itens administrados (10, 15, 20, 25 e 30). Foram simulados 10.000 sujeitos para cada condição com respostas ao acaso. O algoritmo adaptativo utilizou os seguintes passos:

a) Início do teste: seleção aleatória de um valor entre –1 e 1 de uma distribuição Normal aplicando como primeiro item o mais informativo para o nível selecionado.
b) Estimação das habilidades: Se ocorre um padrão inicial constante (até o 5º item), utiliza a média entre a última habilidade estimada e 2 (se acerta) ou –2 (se erra). Após o 5º item aplica-se o procedimento de Herrando (1989) se o padrão se mantém constante, caso contrário, máxima verossimilhança.
c) Seleção de itens: o item mais informativo para a última habilidade estimada. Para os 5 primeiros itens, utilizou-se o método de McBride e Martin (1983) e não eram aplicados itens com parâmetro *a* maior que 1,06 (total de 49 itens). Após os 5 primeiros itens, os itens com parâmetro *a* menor ou igual a 1,06 deixavam de ser aplicados.

d) Procedimento de parada: número fixo de 30 itens.

Os resultados mostraram que, o método que considerou as respostas omitidas como parcialmente corretas produziu habilidades estimadas mais próximas das habilidades verdadeiras enquanto que o outro método superestimou as habilidades estimadas. Abad et al. (2004) concluíram que, para utilizar o TAI convencional que não oferece ao examinando uma opção para omitir respostas, deve-se calibrar o banco de itens considerando as respostas omitidas como parcialmente corretas.

3.3.3. Nível de Conhecimento Inicial

O nível de conhecimento inicial é a estimativa inicial provisória da proficiência, necessária para o início do teste e relacionada com o nível de dificuldade da primeira questão. Conejo et al. (2001) considera que o nível de dificuldade da primeira questão deve ser escolhido de forma a possibilitar uma redução no tempo de teste. Sukamolson (2002) salienta que usualmente a primeira questão possui um nível de dificuldade médio. Como a princípio não se sabe qual é a habilidade do respondente, um item médio é a melhor opção inicial.

A escolha do primeiro item também depende da existência ou não de alguma informação prévia dos indivíduos em aplicações anteriores do teste ou em outros tipos de variáveis relacionadas com a característica medida. Caso se disponha de informações prévias, elas poderão ser utilizadas para estimar o nível inicial da habilidade do indivíduo, o que influenciará a seleção do primeiro item (LÓPEZ-CUADRADO; PÉREZ; ARMENDARIZ, 2005; WEISS; KINGSBURY, 1984; WEISS; SCHLEISMAN, 1999; WISE; KINGSBURY, 2000; YANG; POGGIO; GLASNAPP, 2006). Por exemplo, Schoonman (1989) utiliza a habilidade estimada em um subteste como estimativa inicial para outros três subtestes que compõem um prova. Nas avaliações do NWEA (*Northwest Evaluation Association*), o sistema verifica se o examinando já realizou o teste anteriormente e, caso positivo, ele inicia o teste com a estimativa obtida pelo teste anterior (NWEA, 1997). Entretanto, na maioria das situações práticas, não se tem informações prévias sobre os indivíduos. Nesse caso, pode-se utilizar uma das seguintes estratégias para a escolha do primeiro item:

a) Seleção aleatória de um valor para representar a habilidade inicial, dentro de um intervalo de habilidade média, por exemplo, entre -0,4 e 0,4 ou entre -0,5 e 0,5, para posterior seleção do item de localização mais próxima do valor selecionado. Abad et al. (2004) e Olea et al. (2004) utilizaram uma seleção aleatória de um valor entre

-1 e 1 de uma distribuição Normal aplicando como primeiro item o mais informativo para o nível selecionado.

b) Aplicação um teste curto inicial comum a todos os indivíduos para estimar o nível inicial de habilidade segundo as respostas obtidas.
c) Permissão para que o respondente selecione o nível inicial de habilidade do teste.
d) Utilização da média da distribuição a priori como nível inicial, no caso de estimação bayesiana.
e) Nos testes que possuem ponto de corte (testes com o objetivo de classificação), pode-se selecionar os primeiros itens com dificuldade próxima ao ponto de corte (BERGSTROM; LUNZ, 1999; RENOM; DOVAL, 1999).
f) Pode-se começar com itens fáceis, se o interesse for que os respondentes tenham tendência em acertar os primeiros itens, para incrementar a sua motivação e diminuir a sua ansiedade na escolha dos demais itens (RENOM; DOVAL, 1999; LINACRE, 2000).
g) Para simulações, pode-se selecionar aleatoriamente o primeiro item entre os 10 mais informativos para $\theta = -1$ (YI, 2002).
h) Seleção aleatória dos primeiros itens do nível médio de dificuldade, para evitar que se repita a seqüência inicial de itens entre os diferentes indivíduos (KINGSBURY; HOUSER, 1999).
i) Ho e Yen (2005) propõem até mesmo utilizar informações tais como o nível de formação e a idade do examinando.
j) Outros métodos têm sido propostos (RILEY et al., 2007; ZHU; FAN, 1999).

Alguns desses métodos podem ter influência na estimação final da habilidade, caso sejam aplicados poucos itens (em torno de 10, como critério de parada), principalmente se o nível inicial estiver longe da verdadeira habilidade. Entretanto, pode-se conseguir recuperar o verdadeiro nível de habilidade se o TAI finalizar após a aplicação de uns 20 itens (VAN DER LINDEN; PASHLEY, 2000).

Segundo Segall (2005), a escolha dos primeiros itens depende do método de estimação do traço latente: MV ou bayesiano. No método MV, normalmente se inicia com um item de dificuldade média e, se o respondente acerta, um item mais difícil é selecionado, caso contrário um item mais fácil é selecionado. Se necessário, os itens tornam-se sucessivamente mais difíceis ou mais fáceis até que haja pelo menos um acerto e um erro. No método bayesiano, normalmente a estimativa inicial do traço latente é a média da priori e se inicia com um item com dificuldade nesse nível. No início do teste, quando a quantidade de itens do teste ainda é muito pequena para se avaliar com precisão o valor

verdadeiro da proficiência, a seleção dos itens pode ser feita buscando itens menos informativos (VAN DER LINDEN; GLAS, 2010), principalmente se a estimativa da proficiência não estiver próxima do valor verdadeiro, e economizando-se itens com boa discriminação para utilizá-los quando eles realmente forem necessários.

Como visto na Seção 2.5.2, o Método MV não obtém estimações quando um indivíduo acerta ou erra todos os itens, o que pode ocorrer no início do TAI. Nessa situação, algumas soluções alternativas podem ser propostas:

a) Herrando (1989) propõe que se considere, antes da estimação da habilidade após a primeira resposta, que o indivíduo tenha acertado um item muito fácil (b = -4) e errado um item muito difícil (b = 4), eliminando o problema de estimação do método MV desde o início do teste.
b) Até que se possam estimar as habilidades pelo método MV, Dodd (1990) propõe obter as sucessivas estimações da seguinte forma: (1) se for uma sucessão de acertos, soma-se a habilidade atual estimada com a metade da diferença entre o maior valor de dificuldade dos itens que compõem o banco e a habilidade atual estimada; (2) se for uma sucessão de erros, soma-se a habilidade atual estimada com a metade da diferença entre o menor valor de dificuldade dos itens que compõem o banco e a habilidade atual estimada. O software BILOG-MG (TOIT, 2003) utiliza uma solução semelhante, conforme Andrade, Tavares e Valle (2000): os indivíduos que erraram todos os itens ganham um meio certo no item mais fácil respondido até então, e os indivíduos que acertaram todos os itens, perdem um meio certo no item mais difícil.
c) Revuelta e Ponsoda (1997) modificaram o procedimento de Dodd (1990), propondo que o valor da habilidade atualizada seja a média ou a mediana de uma distribuição normal truncada, levando-se em conta a distribuição provável dos níveis de habilidade na população.
d) Yi (2002) sugere a aplicação de uma estimação bayesiana EAP, até que se possa utilizar o método MV. Nesse caso, a distribuição a priori influenciará diretamente na seleção do primeiro item (VAN DER LINDEN; PASHLEY, 2010).
e) Van der Linden e Pashley (2010) sugerem adiar a estimativa por MV até que haja um número maior de itens respondidos.

Apesar da estimação pelo método bayesiano ser uma alternativa para resolver os problemas causados pelo método MV (vide Seção 2.5), ele também apresenta alguns problemas:

a) A estimação da habilidade não depende apenas do rendimento do indivíduo, mas também dos valores da média e da variância da distribuição a priori.
b) A distribuição a posteriori pode resultar uma distribuição multimodal, o que pode causar problemas na utilização do método bayesiano MAP.
c) No caso de sujeitos que deixam itens sem respostas, os procedimentos bayesianos produzem estimações imprecisas.

3.3.4. Método de Seleção dos Itens

Usualmente a maioria dos TAIs utiliza a chamada estratégia de ramificação variável, o que significa que se estima o nível de habilidade, segundo as respostas dadas a cada item apresentado, a partir do qual se selecionará o próximo item a apresentar. Piton-Gonçalves (2004) menciona que deve-se selecionar o próximo item em função do nível do conhecimento estimado do aluno e da resposta do item anterior para melhorar a precisão na estimativa desse nível, reduzindo conseqüentemente o tamanho do teste.

Segundo Rudner (1998), em geral, a seleção dos itens é feita por um algoritmo formado por um processo iterativo com os seguintes passos: 1) todos os itens que ainda não foram exibidos são avaliados para verificar qual será o próximo item a ser apresentado, dado o nível de habilidade atualmente estimado; 2) o próximo item é disponibilizado e o indivíduo responde; 3) uma nova estimativa da habilidade do indivíduo é calculada baseada nas respostas de todos os itens respondidos até então; 4) os passos 1, 2 e 3 são repetidos até que o critério de parada seja alcançado.

De acordo com Lord (1980), um examinando é avaliado de forma mais eficiente quando os itens dos testes não são muito difíceis nem muito fáceis para ele. Os métodos de seleção adaptativa não só avaliam o nível de dificuldade dos itens, mas também procuram encontrar uma medida de Informação que ajude na escolha dos itens para serem administrados no teste. Wiberg (2003) destaca que se forem utilizados itens inadequados ao examinando, será necessária uma maior aplicação de itens para obter o mesmo resultado caso fossem utilizados apenas itens adequados.

Diferentes estratégias têm sido elaboradas como critério de seleção de itens (ADEMA, 1990; VAN DER LINDEN; ADEMA, 1998; AGUADO et al., 2000; BARRADA; ABAD; OLEA, 2011; BARRADA; OLEA; PONSODA, 2004; BARRADA, et al., 2006; 2009; 2010; 2011; BOWLES; POMMERICH, 2001; CHANG; QIAN; YING,

2001; CHANG; VAN DER LINDEN, 2003; CHANG; YING, 1996; 1999; 2008; 2009; CHEN; ANKENMANN, 2000; 2004; CHEN; ANKENMANN; CHANG, 2000; CHEN et al., 2000; CHENG, 2010; CHENG; CHANG; YI, 2007; CHENG; LIOU, 2000; CHOI; SWARTZ, 2009; COSTA et al., 2009; DAVEY; FAN, 2000; DAVEY; PARSHALL, 1995; DAVIS et al., 2003; DENG; ANSLEY, 2003; DODD, 1990; DODD et al., 1995; EGGEN, 1999; 2004; EGGEN; VERSCHOOR, 2006; FINKELMAN; WEISS; KIM-KANG, 2010; GARCÍA et al., 2000; GARCIA; REVUELTA, 2003; GLAS; VAN DER LINDEN, 2003; HAN, 2009; HAU; CHANG, 2001; LEUNG; CHANG; HAU, 2001; 2002b; 2005; LAU; WANG, 1998; 2000; LI; SCHAFER, 2003b; LIN, 2011; LIN; SPRAY, 2000; LINACRE, 1995; LU; RIZAVI, 2003; MEIJER; NERING, 1999; MURPHY; DODD; VAUGHN, 2010; PASSOS; BERGER; TAN, 2008; PASTOR; DODD; CHANG, 2002; PENFIELD, 2006; REVUELTA, 2004; REVUELTA; PONSODA, 1998a; SCHNIPKE; GREEN, 1995; SEGALL, 2004; SPRAY; RECKASE, 1994; THOMPSON, 2009a; VAN DER LINDEN, 1995; 1998; 2005; VAN DER LINDEN; CHANG, 2003; VAN DER LINDEN; GLAS, 2007; VAN DER LINDEN; PASHLEY, 2000; VAN RIJN et al., 2002; VEERKAMP; BERGER, 1997; VELDKAMP, 2003; WAINER, 2000c; WAINER et al., 1991; WANG; CHANG, 2011a; 2011b; WANG; VISPOEL, 1998; WEISS, 1982; WEISS; MCBRIDE, 1984; WEISSMAN, 2003; 2004; 2006, 2007; YI; ZHANG; CHANG, 2008), onde as mais comuns são baseadas no procedimento da MV que selecionam itens procurando maximizar a informação na estimativa atual da habilidade e nos procedimentos Bayesianos que selecionam itens procurando minimizar a variância da posteriori (SEGALL, 2005).

3.3.4.1. Critério da Máxima Informação (MI)

Esse método consiste em selecionar o próximo item com base na medida de Informação de Fisher (IF) avaliada na proficiência atual estimada, também conhecida como Função de Informação Local (CHANG; YING, 1996). No ML3 da TRI, a IF é a FIT apresentada na seção 2.2.1 e as considerações relacionadas a ela são as mesmas descritas na referida seção.

Segundo Costa (2009), a Informação de Fisher é naturalmente relacionada à estimação da proficiência pela MV e é inversamente proporcional ao erro-padrão do estimador MV. Maximizar a IF significa intuitivamente selecionar um item de dificuldade que corresponda exatamente ao nível de proficiência do examinando. Em relação ao TAI, a IF serve como referência para seleção de itens quando existe

conhecimento suficiente sobre a localização da proficiência. Nas aplicações atuais, esse critério tem sido o mais utilizado porque, entre outras vantagens, permite estabelecer previamente tabelas calculadas de informações, chamadas *infotable* (THISSEN; MISLEVY, 2000).

Itens com maior discriminação serão preferencialmente selecionados pelo algoritmo, o que pode causar dois tipos de problemas no início do TAI, quando a quantidade de itens do teste ainda é muito pequena para se avaliar com precisão o valor verdadeiro da proficiência. Primeiro, a aplicação do método da IF pode ser pouco eficiente se a estimativa da proficiência não estiver próxima do valor verdadeiro. Por exemplo, a Figura 10 mostra o que Van de Linden e Glas (2010) chamam de paradoxo, onde dois itens estão posicionados no valor atual estimado da proficiência. O critério MI selecionaria o item mais informativo para a proficiência atual estimada (Item1), entretanto esse item praticamente não fornece informação onde o verdadeiro valor da proficiência está. No início do TAI, critérios de seleção de itens que não se baseiam na estimativa provisória de θ podem ser mais eficientes do que os critérios de Máxima Informação. À medida que o teste avança, a estimação da habilidade se torna mais precisa, de modo que os critérios de seleção que consideram a estimativa provisória de θ serão mais eficientes. Segundo, esses itens deveriam ser utilizados no final do teste, para estimar a habilidade de indivíduos que realmente estejam nesse nível de habilidade.

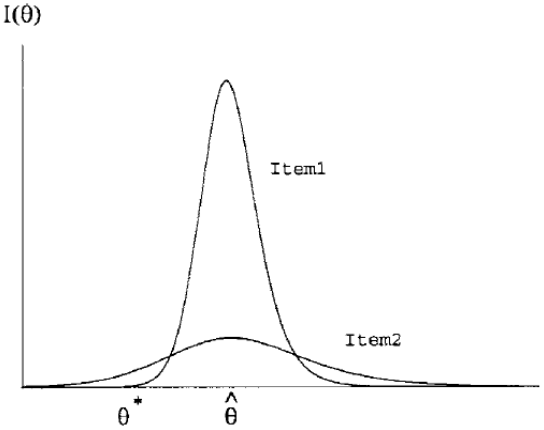


Figura 10. Paradoxo na seleção de itens em TAI (Fonte: Van der Linden e Glas (2010))

O critério MI seleciona como melhor item aquele que produz a menor variância das estimativas. A eficácia dessa estratégia nos TAIs tem sido comprovada através de estudos de simulação, onde se verificou que é possível obter uma boa estimação da habilidade com um número reduzido de itens, em média, 20 itens (OLEA et al., 1996).

Outro problema do método MI é a impossibilidade de estimar θ pela máxima verossimilhança se o respondente sempre responder nas categorias extremas. Nesse caso um procedimento comumente utilizado é o algoritmo *step-size* (DODD; DE AYALA; KOCH, 1995). Os problemas relacionados com o critério MI são piores quando o banco de itens é pequeno ou quando o TAI é de tal forma que precisa ter poucos itens administrados (GARCIA et al., 2000). Estudos recentes (CHANG; YING, 2009) mostram bons resultados da combinação do uso do critério MI com o método de estimação MV.

A utilização "pura" desse critério selecionará sempre os mesmos itens para indivíduos que apresentaram as mesmas respostas. Isso causará um problema de superexposição dos itens, principalmente os primeiros, que poderão tornar-se conhecidos. Para eliminar esse problema, outros métodos que podem ser combinados com esse critério serão mencionados na Seção 3.3.4.4 e na Seção 3.4.

3.3.4.2. Critério da Máxima Informação Global (MIG)

No início do teste, quando a amostra de respostas ainda é pequena e o erro da estimativa da proficiência é alto, Chang e Ying (1996) sugerem substituir a medida de IF pela Informação de Kullback-Leibler (KL), que fornece uma Informação Global, ideal para seleção de itens. Ao contrário da IF, que fornece uma informação local em torno de θ, a Informação KL fornece uma informação global em torno de θ. A escolha de itens de qualidade no início do teste pode reduzir a quantidade total de itens no teste adaptativo. Em relação ao TAI, a informação global deve ser utilizada quando não há informações suficientes sobre essa localização da proficiência.

A medida de informação KL para o ML3 da TRI pode ser expressa por:

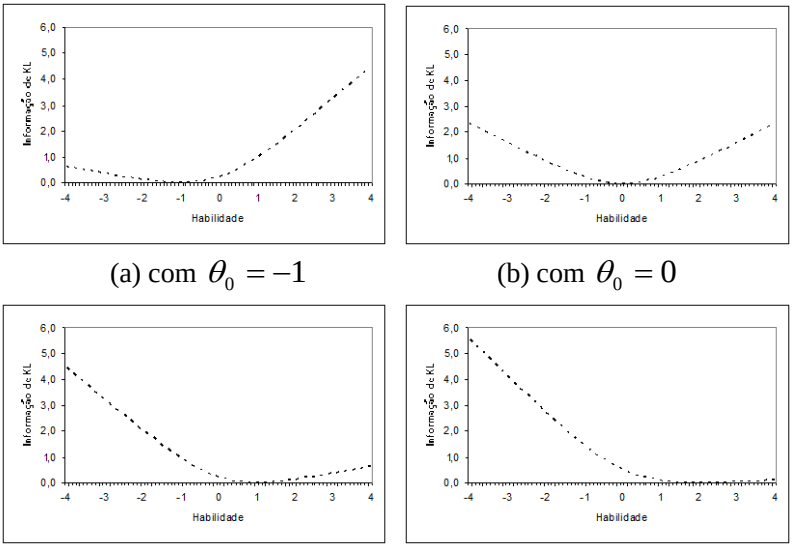
$$K_i(\theta \parallel \theta_0) = P_i(\theta_0)\log\left[\frac{P_i(\theta_0)}{P_i(\theta)}\right] + \left[1 - P_i(\theta_0)\right]\log\left[\frac{1 - P_i(\theta_0)}{1 - P_i(\theta)}\right], \quad (16)$$

onde

θ_0 é o valor verdadeiro da proficiência.

K é uma superfície de informação e representa o poder discriminatório de um item nos dois níveis θ e θ_0, resumindo a informação contida no item com respeito a uma amplo intervalo de θ. Se θ_0 varia ao longo da escala, *K* se torna uma superfície de informação global num espaço tridimensional.

A Figura 11 apresenta a informação KL para um item com parâmetros: a = 1,5, b = 0 e c = 0. Observa-se pela equação 16, que a função K_i é uma superfície de informação, pois também depende do valor de θ_0. Dessa forma, estabeleceram-se quatro valores para θ_0 com o objetivo de verificar o comportamento da informação KL. Observa-se que quando $\theta_0 = 0$, a informação cresce para a direita e para a esquerda da curva de forma simétrica em torno de 0. Quanto maior o valor de θ_0, mais informação existe para os valores menores de habilidade em relação a θ_0 e menor informação para os valores maiores de habilidade em relação à θ_0.

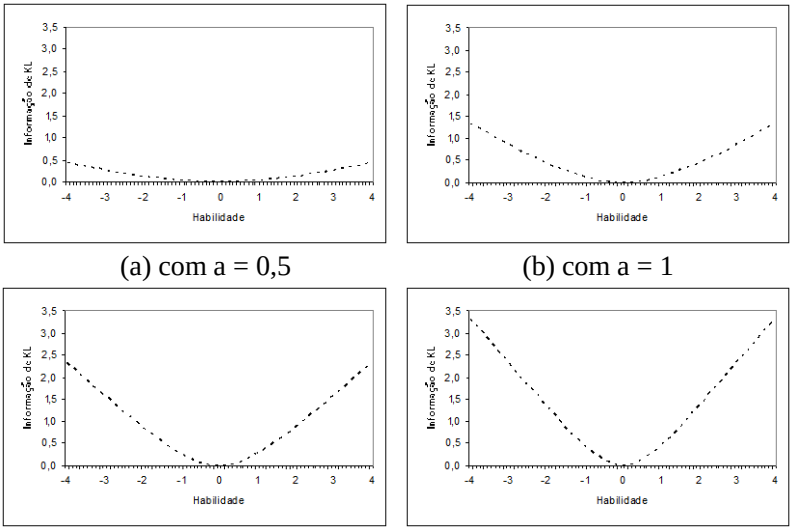


(a) com $\theta_0 = -1$ (b) com $\theta_0 = 0$

(c) com $\theta_0 = 1$ (d) com $\theta_0 = 2$

Figura 11. Informação de KL para um item com parâmetros: a = 1,5, b = 0 e c = 0.

Na Figura 12, optou-se por fixar $\theta_0 = 0$, e os parâmetros dos itens b = 0 e c = 0, variando o valor de a (0,5; 1; 1,5 e 2) para examinar a influência da discriminação. Nesse caso, trata-se de 4 itens diferentes. Observa-se que quanto maior a discriminação do item, maior é a informação KL dele para valores acima e abaixo de 0. Entretanto, quando $\theta_0 = 1$ (maior que *b*), o comportamento não é simétrico (Figura 11). Nesse caso, a Figura 12 não mostra, mas quanto maior a discriminação do item, maior é a informação para valores de habilidade abaixo de $\theta_0 = 1$. Porém, para valores de habilidade acima de $\theta_0 = 1$, a informação sofre uma sensível diferença, aumentando ou diminuindo segundo o valor de *a*. Por outro lado, quando $\theta_0 > b$, o comportamento da informação é inverso.



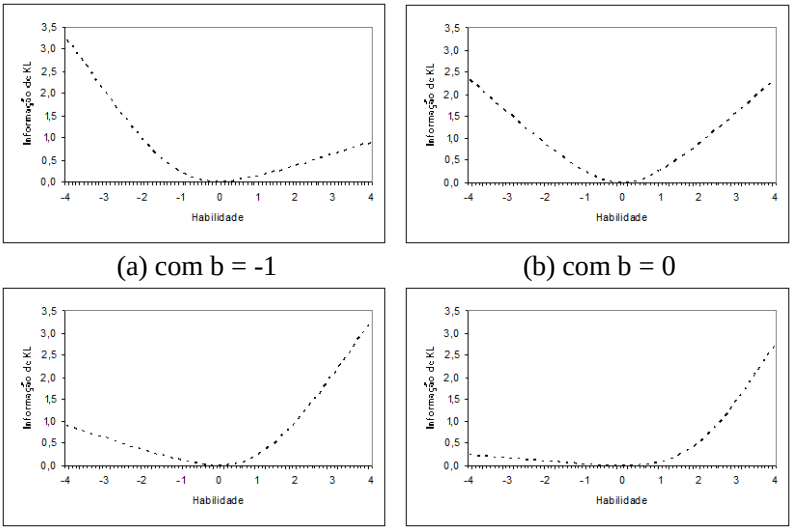
(a) com a = 0,5 (b) com a = 1

(c) com a = 1,5 (d) com a = 2

Figura 12. Informação de KL para itens com parâmetros: a = (0,5; 1; 1,5 e 2), b = 0 e c = 0 e com $\theta_0 = 0$.

Na Figura 13, optou-se por fixar $\theta_0 = 0$, e os parâmetros dos itens a = 1,5 e c = 0, variando o valor de b (-1; 0; 1 e 2) para examinar a influência da dificuldade. Nesse caso, também trata-se de 4 itens diferentes. Observa-se que quando a dificuldade do item é 1 (c) em

relação a (b) a informação aumenta para habilidades maiores que 0 e diminui para as habilidades menores. Entretanto, quando a dificuldade do item é 2 (d) em relação a (b) a informação também aumenta para habilidades maiores que 0 e diminui para as habilidades menores, mas em relação a (c), a informação diminui para todos os valores. Observa-se que, para valores de habilidade menores que 0 (a), o comportamento da informação é simétrico, pois c = 0. Porém, para outros valores de θ_0, o comportamento da informação KL é diferente e não deve se basear nessa análise.



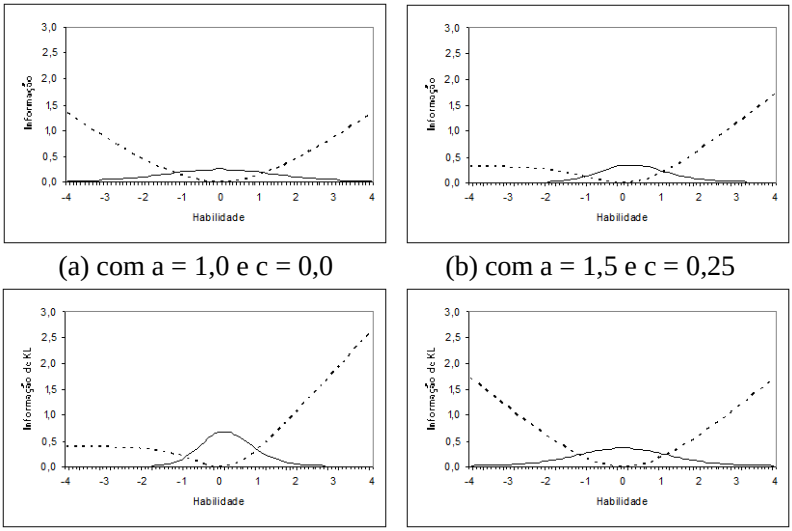
(a) com b = -1 (b) com b = 0

(c) com b = 1 (d) com b = 2

Figura 13. Informação de KL para um item com parâmetros: a = 1,5, b = (-1; 0; 1 e 2) e c = 0 e com $\theta_0 = 0$.

A Figura 14 apresenta uma comparação entre a IF (linha contínua) e a informação KL (linha tracejada) para quatro itens com a mesma dificuldade (b = 0) e diferentes valores para os parâmetros a e c, considerando que $\theta_0 = 0$. Segundo o critério MI, o item selecionado para esse nível de habilidade seria o (c). Por outro lado, pelo critério MIG, o item (d) possui maior informação quando se considera todo o espaço paramétrico (no caso, $-4 < \theta < 4$). Da mesma forma, se a decisão fosse entre os itens (a) e (b), o critério MI selecionaria o item (b)

enquanto que o critério (MIG) selecionaria o item (a). Essas interpretações são válidas considerando $\theta_0 = 0$. Nota-se que, para outros valores de θ_0, a IF não muda, somente a informação KL.



(a) com a = 1,0 e c = 0,0 (b) com a = 1,5 e c = 0,25

(c) com a = 2 e c = 0,2 (d) com a = 1,2 e c = 0

Figura 14. Comparação entre as Informações de KL e a IF para quatro itens.

Uma dificuldade para se selecionar itens pela medida de Informação de KL está no fato de que para um dado θ_0, K é uma função no espaço paramétrico enquanto a IF é um único valor. Chang e Ying (1996) desenvolveram uma forma para obter um único valor para K através do índice médio de Informação KL, definido como:

$$K_i\left(\hat{\theta}\right) = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} K_i\left(\theta \,\|\, \hat{\theta}\right) d\theta, \qquad (17)$$

onde δ determina o tamanho do intervalo sobre o qual a média será computada.

Para um pequeno valor de δ, este índice é essencialmente determinado pela curvatura de $K_i(\theta \,\|\, \hat{\theta})$ em $\hat{\theta}$, mas para grandes valores de δ, a área é muito influenciada pelos extremos de $K_i(\theta \,\|\, \hat{\theta})$, o que reflete a idéia da abordagem da informação global.

Assim como a IF colabora para fornecer informação total do teste (FIT), a informação KL também fornece uma informação total para o teste, a chamada informação KL do teste, que consiste na soma de todas as informações individuais dos itens e, para um teste com k itens, denota-se por:

$$K(\theta \parallel \theta_0) = \sum_{i}^{k} K_i(\theta \parallel \theta_0). \quad (18)$$

3.3.4.3. Critério da Máxima Informação Esperada (MIE)

Segundo Costa (2009), o critério MIE, também conhecido como método de Owen (1975), é um dos procedimentos Bayesianos mais empregados em TAI para seleção de itens. Todos os critérios Bayesianos para seleção de itens no TAI envolvem alguma forma de ponderação baseada na distribuição a posteriori de θ. A diferença básica entre os critérios já mencionados é que este faz uso de uma distribuição a priori. O mais fácil seria estabelecer uma distribuição a priori com base nas estimações das habilidades obtidas de estudos empíricos. Outra alternativa seria estabelecer uma distribuição a priori a partir das estimativas pontuais de habilidade mediante uma equação de regressão.

Em TAI, deseja-se prever a resposta aos itens ainda não administrados no teste, depois de $k-1$ respostas e, então, escolher o próximo item de acordo com as atualizações de uma quantidade a posteriori para essas respostas. Quando um item *k* qualquer é selecionado, o examinando responde e a resposta observada irá atualizar as seguintes quantidades: a distribuição a posteriori de θ; a estimativa pontual do valor da proficiência do examinando e a variância a posteriori de θ. Segundo Owen (1975), esse caso, não se busca o item que mais contribui com a precisão da estimação de um nível de habilidade, mas o mais apropriado para toda uma distribuição de estimações.

O objetivo do critério MIE consiste em maximizar a medida de Informação Observada (medida bayesiana) sobre as respostas preditas ao *k*-ésimo item. A escolha do próximo item que será administrado no TAI pelo critério MIE levará em conta a medida de Informação Observada (IO) dos itens no ponto $\hat{\theta}$. Nota-se que, diferente da IF e da Informação KL, a IO não é um valor fixo, mas é um valor que é atualizado de acordo com os itens que forem sendo administrados.

Esse critério é muito exigente computacionalmente, o que tem levado a utilização de procedimentos mistos, os quais proporcionam

maior precisão e eficiência do que qualquer procedimento aplicado individualmente, necessitando de menos itens para alcançar um determinado nível de precisão (SEGALL; MORENO, 1999).

3.3.4.4. Outros Critérios Utilizados

O Método Progressivo, proposto por Revuelta e Ponsoda (1998a), atribui um peso a cada item que indica a sua utilidade para estimar o nível de capacidade do sujeito. O item de maior resultado é selecionado e administrado ao indivíduo. O peso de um item é resultado da soma ponderada de um valor aleatório e da informação do item para a última habilidade estimada. A informação é ponderada pelo número de itens administrados no TAI, enquanto que o valor aleatório é ponderado pelo número de itens que não foram administrados no teste. Dessa forma, no início do TAI o componente aleatório tem um peso maior e os itens mais informativos não são selecionados. A medida que os itens vão sendo selecionados e aplicados, a informação adquire maior peso, o que contribui para selecionar itens mais informativos quando o nível de habilidade estimado está mais próximo do verdadeiro parâmetro.

Eggen (2001) propõe uma generalização do método progressivo (Método Progressivo Generalizado) com o objetivo de ajustar com mais precisão a importância dos componentes, o que melhora a precisão de medida do TAI, através da inclusão de um parâmetro pré-fixado pelo administrador do teste.

Garcia et al. (2000) sugeriram a utilização do Critério da Mínima Entropia. Conforme Shannon (1949), a entropia é uma quantificação imparcial da benignidade de um item, baseada nas características das CCIs, sem levar em conta o nível anterior de habilidade estimada. Esse indicador mede o poder de discriminação do item e oferece valores pequenos quando as CCIs são estreitas e separadas (itens com boa discriminação e posicionados em diferentes lugares ao longo da escala). Por outro lado, o indicador aumenta quando a distribuição de probabilidades começa a ser mais homogênea.

O Critério *Maximum Posterior Precision* (MPP) seleciona itens que têm o maior decréscimo na variância da distribuição a posteriori da habilidade. O item selecionado pode não ser o mais informativo daquele nível, porém ele mede com eficiência a região de maior densidade da distribuição a posteriori. O critério MPP não pode utilizar tabelas de informação (vide Seção 3.4.1.1) e é muito mais complexo computacionalmente. Maiores detalhes sobre esse critério são obtidos em Parshall et al. (2002).

O Critério da Informação Ponderada (*Weighted Information – WI*) é um procedimento que seleciona um item no qual os pesos da atual distribuição a posteriori de habilidade do indivíduo são aplicados na tabela de informação. Durante a seleção do item, a informação obtida por cada item em cada nível de habilidade são multiplicadas pelos seus pesos e somadas. O item com a maior informação somada é então selecionado. Maiores detalhes sobre esse critério são obtidos em Parshall et al. (2002).

Outro método utilizado é o Método 5-4-3-2-1 de McBride e Martin (1983), que consiste em selecionar aleatoriamente o primeiro item entre os cinco mais informativos no nível atual de habilidade estimada, o segundo entre os quatro mais informativos no nível atual de habilidade novamente estimada, e assim sucessivamente até o quinto item, onde passa a utilizar o critério MI.

Kingsbury e Zara (1989) desenvolveram o Método Randomesque, que consiste sempre em selecionar aleatoriamente entre os cinco itens mais informativos. Esse método pode ser utilizado tanto na seleção dos primeiros itens no início do teste quanto no teste inteiro (THISSEN; MISLEVY, 2000).

Abad et al. (2004) utilizaram o item mais informativo para a última habilidade estimada. Entretanto, para os 5 primeiros itens, utilizou-se o método de McBride e Martin (1983) e não foram aplicados itens com parâmetro *a* maior que 1,06. Após os 5 primeiros itens, os itens com parâmetro *a* menor ou igual a 1,06 deixam de ser aplicados.
O algoritmo de Olea et al. (2004) elege o item mais informativo, ainda não selecionado no teste, para a última habilidade estimada. Entretanto, para os 5 primeiros itens, utiliza o método de McBride e Martin (1983) no qual não foram aplicados itens com parâmetro *a* maior que 1. Após os 5 primeiros itens, os itens com parâmetro *a* menor ou igual a 1 deixam de ser aplicados.

Dood et al. (1995) salientam que para os TAIs politômicos existem outros critérios para a seleção de itens, além dos descritos anteriormente. Entre eles, destacam-se o critério da máxima informação da categoria, o critério da máxima proximidade entre o parâmetro de localização e o nível de habilidade estimado e o critério da máxima informação para um determinado intervalo de valores de habilidade.

Nos testes que possuem o objetivo de classificar o examinando, os métodos de seleção de itens podem ser utilizados para maximizar a informação na localização do ponto de corte e não na localização da estimativa provisória da proficiência do examinando (THOMPSON; PROMETRIC, 2007). Três métodos baseados no ponto de corte têm

sido sugeridos (LIN; SPRAY, 2000): critério da Máxima Informação (vide Seção 3.3.4.1), critério da Máxima Informação Global (vide Seção 3.3.4.2) e uma razão log-odds, proposta por Lin e Spray (2000), baseada na *Sequential Probability Ratio Test* (SPRT). Outro critério utilizado é a Informação Mútua (WEISSMAN, 2004), apropriado para o início do teste ou quando existem muitos pontos de corte.

Num TAI multidimensional, os itens são selecionados para maximizar a informação em várias dimensões simultaneamente (SEGALL, 2005). Reckase (2009) apresenta vários critérios de seleção de itens para um TAI multidimensional, sendo que alguns são generalizações do TAI unidimensional. Um TAI multidimensional pode apresentar de 30% a 50% menos itens na sua aplicação em comparação a um TAI unidimensional (FREY; SEITZ, 2009).

3.3.4.5. Alguns Estudos Comparativos

Garcia et al. (2000) fizeram um estudo para comparar a eficiência de três algoritmos de seleção de itens através de simulações de TAIs utilizando um banco com 28 itens de resposta gradual (modelo de Samejima), tomando como resposta ao TAI a resposta original dos sujeitos que foram utilizados para a calibração do banco. Os algoritmos analisados foram: a) baseado na máxima informação, b) baseado na mínima entropia, c) misto baseado na mínima entropia para os itens iniciais e na máxima informação para os demais. Foram realizadas simulações sobre as respostas reais, e a estimação de θ de cada um dos sujeitos foi feita pelo método da máxima verossimilhança. No caso da máxima Informação, o procedimento do TAI começou com uma estimação aleatória de θ entre -1 e 1, e quando houve respostas extremas que impediram esse método, se implementou um procedimento *step-size*. No caso da mínima entropia, o procedimento TAI utilizou como critério para a seleção de itens a minimização da função de entropia. Para cada método, se comparou a correlação entre o valor de θ estimado com todos os 28 itens e o valor de θ estimado por máxima verossimilhança em cada etapa do processo TAI. Os resultados mostraram que o critério misto apresentava maior correlação entre as habilidades estimadas e as verdadeiras do que os demais critérios. O algoritmo de mínima entropia não produz um processo adaptativo, uma vez que não utiliza nenhuma informação de θ. Nesse sentido, ele deve ser estudado no contexto misto (em qual momento do TAI a informação estimada de θ é suficientemente precisa para trocar o método para o de máxima informação) ou deve incorporar a informação sobre a resposta

do sujeito (isso restringirá o intervalo de θ possível e o tornará adaptativo). Apesar do algoritmo de entropia proporcionar uma solução satisfatória para a eleição do primeiro item do TAI, isso causa um problema na exposição desse item, uma vez que esse algoritmo seleciona como primeiro item o melhor item do banco, que é sempre o mesmo.

Chen et al. (2000) compararam o impacto de cinco regras de seleção de itens no desvio, na raiz quadrada do erro quadrado médio, e no erro padrão de medida estimado nas fases iniciais do TAI. Os métodos comparados foram: a) Máxima Informação (IF), b) Máxima Informação no intervalo proposto por Veerkamp e Berger (1997), c) Máxima Informação com distribuição a posteriori (FIP), d) a Função de Informação KL, e e) a Função de Informação KL com distribuição a posteriori (KLP). Os resultados mostraram que não houve diferenças em testes com mais de 10 itens.

García e Revuelta (2003) compararam os métodos Progressivo Generalizado, MI e Aleatório. Os métodos foram testados através de simulações, utilizando um banco de 197 itens de gramática inglesa tomados de um TAI real que havia sido calibrado com o ML3. Para cada cenário foram aplicados 5.000 TAIs simulados de tamanho fixo 20. Em cada simulação foram registrados: a habilidade estimada, o erro de estimação e os itens aplicados. Quanto à precisão da medida, os resultados mostraram que: o método MI foi o mais preciso, atendendo a todos os indicadores; o método progressivo resultou mais preciso quando o valor do parâmetro *a* era alto; e o método aleatório teve o pior desempenho em todos os indicadores. Quanto à taxa de exposição, o pior resultado, tanto de superexposição quanto subexposição dos itens, foi alcançado pelo MI. A proporção média de itens comuns entre dois testes foi maior no MI e menor no Aleatório. Os resultados elevados dessa proporção se devem ao reduzido banco de itens e ao tamanho do teste.

Revuelta e Ponsoda (1998a) compararam a eficiência de diferentes métodos (MI, Restrito, Progressivo, 5-4-3-2-1, Randomesque, algoritmo SH – *Sympson-Hetter*, entre outros). Os métodos Restrito e SH utilizaram uma taxa de exposição máxima de 0,4. Duas condições diferentes de parada de teste foram testadas: tamanho fixo do teste (35 itens) e critério misto (erro padrão inferior a 0,22 ou aplicação de 50 itens). Os resultados mostraram que: a) nenhum dos métodos foi completamente satisfatório, b) os métodos 5-4-3-2-1, Randomesque tiveram melhores resultados para o caso em que os respondentes não recebem itens similares no começo do teste, c) os métodos Restrito e SH

foram os que conseguiram reduzir em maior grau a superexposição de alguns itens, d) o método progressivo é o que consegue aumentar em maior grau a taxa de exposição de itens com baixa frequência de utilização, e) a combinação dos métodos Progressivo e Restrito obteve os melhores resultados globais em relação à precisão, redução das taxas de exposição e redução do número de itens utilizados no teste.

Barrada et al. (2009a) compararam três métodos de controle para a taxa de exposição do item: o método SH, o método Restrito, e o método de Elegibilidade do item. Os resultados mostraram que o método Restrito foi o melhor no controle da taxa de exposição, seguido pelo método de Elegibilidade do item. Entretanto, os autores recomendam o método de Elegibilidade do item, pois o método Restrito apresentou alguns problemas.

3.3.5. Estimação da Habilidade

Toda vez que um item é selecionado e aplicado num teste, a habilidade do examinando é reestimada juntamente com o seu erro padrão. Os principais métodos utilizados na estimação da habilidade foram mencionados na Seção 2.5. Entretanto, existem diversas adaptações, alterações ou combinações desses métodos no contexto dos TAIs, além da criação de novos métodos (ABAD et al., 2002; AL-A'ALI, 2007; BERGSTROM; LUNZ; GERSHON, 1992; BLAIS; RAÎCHE, 2002; 2010; BOCK; MISLEVY, 1982; CHEN; HOU; DODD, 1998; CHEN et al., 1997; CHENG; LIOU, 2000; 2003; DE AYALA, 2008; DE AYALA; SCHAFER; SAVA-BOLESTA, 1995; GLAS; WAINER; BRADLOW, 2000; GORIN et al., 2005; GREEN, 1997; HERRANDO, 1989; HONTANGAS; PONSODA; OLEA, 1999; HOU et al., 1996; LINACRE, 1995, 1998; MCBRIDE, 1977; MEIJER; NERING, 1999; MISLEVY; WU, 1988; NICEWANDER; THOMASSON, 1999; PENFIELD, 2007; PONSODA et al., 2004; RAÎCHE; BLAIS, 2002; 2009; REVUELTA, 2004; REVUELTA; PONSODA, 1997; RULISON; LOKEN , 2009; STOCKING, 1996; TSUTAKAWA; JOHNSON, 1990; VAN DER LINDEN, 1998; 1999a; 1999b; VAN DER LINDEN; PASHLEY, 2003; VEERKAMP, 2000; WANG; HANSON; LAU, 1999; WANG; VISPOEL, 1998; WANG; WANG, 2001; WARM, 1989; WEISS, 1973; 1974; 1982; WEISS; MCBRIDE, 1984; WISE, 1999a; YANG; POGGIO; GLASNAPP, 2006; YI; NERING, 1998). Por exemplo, Abad et al. (2004) utilizaram a seguinte estratégia para estimar a habilidade: se ocorre um padrão inicial de resposta constante (até o 5º item), utiliza a média entre a última habilidade estimada e 2 (se acerta) ou – 2 (se erra). Após o 5º item

aplica-se o procedimento de Herrando (1989) se o padrão se mantém constante, caso contrário, utiliza-se o método da máxima verossimilhança. É comum utilizar um método no início do teste, quando o erro padrão da estimativa da habilidade ainda é grande e pode ocorrer um padrão de resposta constante, e outro método durante o teste, quando o erro padrão é menor.

No contexto dos TAIs, o Método MV apresenta, em relação aos Métodos Bayesianos: maior erro padrão (especialmente para valores baixos e altas da habilidade), menor viés, menor fidelidade (correlações entre valores estimados e parâmetros), menor eficiência (precisa de mais itens para alcançar a mesma precisão), e maior tempo para os cálculos computacionais (WANG; VISPOEL, 1998). Wise e Kingsbury (2000) consideram mais adequado utilizar o método MV, pelo fato da estimativa do traço latente não ser afetada por qualquer outra coisa que não seja o desempenho no teste atual.

Num TAI, as propriedades do estimador MV dependem de fatores como a distribuição dos itens no banco e o critério utilizado para a seleção de itens. A questão da combinação entre critérios de seleção de itens e métodos de estimação da habilidade é muito complicada para ser tratada analiticamente. Uma solução usual é realizar estudos de simulação para comparar as diferentes combinações de métodos (VAN DER LINDEN; PASHLEY, 2010).

Nos TAIs, as estimativas Bayesianas tendem a ter a vantagem de erros padrão condicionais menores, mas possuem a desvantagem de ter viés da estimativa da habilidade condicional maior, especialmente para os níveis extremos de θ. Assim, a escolha do método de estimação deve levar em conta tanto a variância pequena (das estimativas bayesianas) quanto o viés pequeno (das estimativas por MV). Os procedimentos Bayesianos oferecem um menor erro quadrado médio (que é uma função de ambos variância e viés condicionais) do que o Método MV. Isto sugere que as estimativas Bayesianas podem fornecer uma classificação mais precisa da ordenação dos examinandos ao longo da escala do traço latente. Estudiosos que estão preocupados com os efeitos do viés ou que não têm informações sobre a distribuição do traço latente tendem a utilizar a abordagem MV. Por outro lado, estudiosos cujo principal objetivo é minimizar o erro padrão médio ou a variância condicional tendem a utilizar abordagens Bayesiana (SEGALL, 2005).

3.3.6. Critério de Parada

O critério de parada é uma regra usada para determinar quando o teste deve terminar e está relacionada com o objetivo do teste, as

características do banco de itens, e com as restrições operacionais. (SEGALL, 2005).

Existem diversos critérios que podem ser utilizados para finalizar o teste que podem ser utilizados individualmente ou combinados (BLAIS; RAÎCHE, 2002; 2010; CHOI; GRADY; DODD, 2011; DODD; KOCH; DE AYALA, 1993; LINACRE, 2000; 2006; PARSHALL et al., 2002; RILEY et al., 2007; SEGALL, 2005; THISSEN, 1986; TIAN et al., 2007). O critério de parada adotado influencia diretamente o erro padrão da medida (EP), o qual possui relação direta com a precisão (GREEN et al., 1984).

3.3.6.1. Testes para Fins de Estimação da Habilidade

Normalmente, dois critérios de parada são utilizados, para fins de estimação da habilidade: tamanho fixo ou tamanho variável (SEGALL, 2005):

Nos testes de tamanho fixo, todos os indivíduos respondem a mesma quantidade de itens e, consequentemente, a precisão da habilidade estimada não vai ser a mesma para todos os indivíduos, motivo pelo qual não é muito recomendado. Nesse caso, segundo Revuelta e Ponsoda (1998a), a tendência é que as estimativas de indivíduos com níveis de habilidade extremos tenham menor precisão, podendo estar abaixo do nível desejado. Os problemas serão maiores se o banco tiver poucos itens informativos nesses níveis de habilidade.

A quantidade de itens pode ser determinada por meio de estudos de simulação, mas não deve ser maior do que a versão "papel e lápis" do teste, caso exista. Segundo Fernandes (2009), normalmente utiliza-se entre 30 e 70 itens. Abad et al. (2004) adotaram 30 itens, mas a quantidade pode ser ainda menor, por exemplo, Garcia e Revuelta (2003) e Olea et al. (2000) que utilizaram 20.

Nos testes de tamanho variado, os itens vão sendo aplicados até que seja alcançada uma determinada precisão para a habilidade estimada. Essa precisão normalmente está relacionada com a informação do teste, no caso da estimação por MV, ou com a variância da posteriori, no caso da estimação bayesiana. Conforme Weiss e Kingsbury (1984), nesse caso, é desejável mensurar cada examinando para um nível de precisão fixo, ou seja, um nível pré-determinado do erro-padrão é fixado. Isso resultará em um conjunto de medidas em que todos os examinandos terão escores com equivalentes precisões, assim as interpretações e decisões serão feitas com a mesma precisão para todos os indivíduos. Para implementar essa medida, o TAI permite que seja especificado o nível da medida de erro-padrão desejável para cada

examinando. Assumindo que o banco de itens tem um número suficiente de itens distribuídos em toda escala do traço latente e que o tamanho do teste seja o suficiente para cada examinando, o teste é finalizado quando o nível do erro-padrão da medida for satisfeito.

Se o banco de itens não for grande o suficiente para alcançar esse critério de parada (EP) ou se outro critério for adotado, a medida do erro irá depender do nível de habilidade (GREEN et al., 1984). O valor do erro padrão considerado pequeno depende também da metodologia escolhida para o teste e da quantidade de questões no banco de questões, mas ele normalmente é considerado pequeno se for menor que 0,32 na escala (0, 1) (FLIEGE et al., 2005).

3.3.6.2. Testes para Fins de Classificação

Para fins de classificação de um indivíduo (aprovado/reprovado, classificado/desclassificado, satisfeito/insatisfeito, etc.), a proficiência de um examinando geralmente é comparada com algum valor de corte expresso na mesma escala do traço latente (BARTROFF, 2008; BERGSTROM; LUNZ, 1992; CHANG, 2005; EGGEN, 1999; 2010; EGGEN; STRAETMANS, 2000; FINKELMAN, 2008; GLAS; VOS, 2006; 2010; KALOHN; SPRAY, 1999; KINGSBURY; WEISS, 1983; LAU; WANG, 1999; LIN, 2011; LIN; SPRAY, 2000; PARSHALL et al., 2002; RECKASE, 1983; SEGALL, 1997; SPRAY; RECKASE, 1994; 1996; THOMPSON, 2006; 2009a; 2011; THOMPSON; PROMETRIC, 2007; VOS, 2000; VOS; GLAS, 2000; 2010; WANG; LIU, 2011; WEISS, 1982; WEISS; KINGSBURY, 1984; WEISSMAN, 2007; WIBERG, 2003; XIAO, 1999; YANG; POGGIO; GLASNAPP, 2006).

Segundo Lin e Spray (2000), esse tipo de TAI é chamado de Teste de Classificação Informatizado (*Computerized Classification Testing - CCT*). Porém outras nomenclaturas são usadas, como Teste Adaptativo de Maestria (*Adaptive Mastery Testing – AMT*), conforme GLAS e VOS (2010), cujo objetivo é classificar o examinando como "mestre" ou "não mestre". Thompson e Prometric (2007) acreditam que a falta de nomenclatura padronizada nos componentes e tipos de CCT tem causado confusão.

Três critérios de terminação são utilizados em nos CCT (THOMPSON; PROMETRIC, 2007): intervalos de confiança, o teste SPRT e a teoria da decisão bayesiana. A adequação de cada critério e a maximização de seus benefícios dependem de vários outros fatores, tais como o modelo de resposta ao item e o algoritmo de seleção de item. Por exemplo, o uso de intervalos confiança exige um grande banco de

itens calibrados, que, por sua vez, requer uma amostra grande para a calibração.

A finalização poderá ocorrer quando o nível de habilidade estimado estiver seguramente longe do ponto de corte, com uma distância de pelo menos dois erros padrão ou quando não houver mais itens suficientes para o indivíduo alcançar o ponto de corte, segundo Linacre (2000). Weiss e Kingsbury (1984) indicam que tanto a estimativa da proficiência como o erro-padrão da medida associado devem ser usados. No caso da estimação das proficiências pelo método EAP, o PSD (*Posterior Standard Deviation*) é o erro-padrão associado à medida ou o nível de precisão. A confiabilidade das estimativas das proficiências será a mesma para todos os examinandos. Dessa maneira, uma medida de precisão constante para todos os indivíduos assegura que os erros cometidos na classificação serão uniformes para todas as decisões. Esse tipo de eqüidade nos procedimentos de seleção e classificação dos indivíduos não pode ser alcançado nos testes convencionais de tamanho fixo. Um indivíduo pode ser classificado como sendo acima do valor de corte se a estimativa da sua proficiência e seu intervalo de 95% de confiança estiver acima do escore de corte, ou classificado abaixo, caso contrário (EGGEN; STRAETMANS, 2000; KINGSBURY; WEISS, 1983). Dessa forma, o teste termina quando o nível de habilidade estimado se afasta significativamente do ponto de corte estabelecido. Na utilização da abordagem do intervalo de confiança, os procedimentos que selecionam itens mais próximos da habilidade estimada são mais adequados, pois diminuem a largura do intervalo de confiança (EGGEN; STRAETMANS, 2000).

Outro procedimento também utilizado, baseado na SPRT é o teste de hipótese, onde é testado se a habilidade do indivíduo é igual ou maior que um valor, usualmente o ponto de corte (RECKASE, 1983; WEITZMAN, 1982). Esse procedimento também pode ser estendido para o caso onde existem dois ou mais pontos de corte utilizados para classificar os examinandos em três ou mais grupos (EGGEN; STRAETMANS, 2000; RUDNER, 2002). Na utilização da abordagem do teste de hipótese baseado na SPRT, os procedimentos que selecionam itens mais próximos do ponto de corte são mais adequados (SPRAY; RECKASE, 1994).

A teoria da decisão bayesiana é fundamentada na TCC e tem sido tradicionalmente utilizada em testes de classificação baseados na TCC (THOMPSON; PROMETRIC, 2007). No entanto, a TRI e técnicas de seleção adaptativa de itens foram recentemente incorporados utilizando este critério de terminação (GLAS; VOS, 2006).

Vários pontos de discórdia permanecem nos CCTs, sendo que o mais importante é em relação ao critério de terminação mais eficaz (THOMPSON; PROMETRIC, 2007). Spray e Reckase (1996) consideram que o desempenho do teste de hipótese baseado na SPRT é superior ao desempenho dos intervalos de confiança, porém Chang (2005) considera o contrário. Além disso, esses métodos não foram comparados com a teoria da decisão bayesiana.

Os procedimentos utilizados em testes de classificação geralmente resultarão em testes mais curtos para indivíduos que possuem habilidade localizada longe do ponto de corte, e, por outro lado, em testes mais longos para indivíduos que possuem habilidade próxima do ponto de corte. Um guia para a elaboração de testes adaptativos para fins de classificação foi elaborado por Thompson e Prometric (2007). Esses autores não consideram que o CCT seja um TAI, pois o item não é selecionado segundo a estimativa da proficiência do indivíduo, mas segundo a localização do ponto de corte. Entretanto, o CCT possui várias características de um TAI, pois é um teste onde nem todos os examinandos recebem a mesma quantidade de itens, e que utiliza uma estrutura semelhante a dos TAIs, com um modelo psicométrico, um banco de itens, uma estimativa inicial da proficiência, um algoritmo de seleção de itens e um critério de parada. Além disso, o item nem sempre é selecionado segundo a localização do ponto de corte, podendo ser selecionado segundo a estimativa da proficiência do indivíduo (EGGEN; STRAETMANS, 2000), como no caso dos intervalos de confiança.

Nos testes tradicionais com ponto de corte que foram transformados em testes adaptativos, uma questão relevante é como estabelecer o novo ponto de corte na escala criada pelo TAI (VAN DER LINDEN; PASHLEY, 2010). Alguns procedimentos para estabelecer esse ponto de corte equivalente ao teste tradicional são apresentados em Lord (1980), Olsen, Maynes e Slawson (1986), Segall (1993; 1997) e Stocking (1996), Van der Linden (2001; 2006b, 2010a), por exemplo.

3.3.6.3. Outros Critérios Considerados

Em TAIs com banco de itens pequeno, o esgotamento do banco de itens (quando todos os itens forem aplicados) poderá encerrar o teste (LINACRE, 2000; TIAN et al., 2007).

O algoritmo também pode ser programado para encerrar o teste quando for reconhecido um comportamento inadequado do examinado, por exemplo, ele pode estar respondendo a mesma alternativa em todos

os itens, ou respondendo muito rápido ou muito devagar (LINACRE, 2000; TIAN et al., 2007).

Outro critério que pode ser considerado para finalizar o teste é estabelecer um tempo limite de teste. Segundo Revuelta e Ponsoda (1998a), esse critério não é recomendado, pois além do erro padrão diferente, cada indivíduo poderá ter respondido uma quantidade diferente ou até mesmo insuficiente de itens, podendo até mesmo deixar de responder alguns itens, comprometendo seu desempenho no teste. Alguns grupos étnicos e classes sociais necessitam de mais tempo para realizar o teste (WISE; KINGSBURY, 2000). Entretanto, em muitas situações práticas é necessário impor um limite de tempo. Nesses casos, normalmente o tempo limite é suficientemente grande para os examinandos realizarem o teste (SEGALL, 2005), não limitando significativamente o desempenho dos alunos e mantendo a sessão de testes razoavelmente curta (WISE; KINGSBURY, 2000).

Outro fator que pode ser considerado é a definição da uma quantidade máxima de itens aplicados (LINACRE, 2000), para que o teste não seja muito demorado, caso a precisão determinada demore a ser alcançada. Também pode ser definida uma quantidade mínima de itens aplicados, pois, mesmo que o nível de precisão seja alcançado, alguns testes, por exemplo, exigem que seja aplicada uma quantidade mínima de itens de cada conteúdo do teste.

Segundo Fernandes (2009), existem avaliações clássicas que permitem que o examinando termine a avaliação sem ter respondido todas as questões. Dessa forma, a nota dele é calculada apenas com base nas questões respondidas, ou seja, as questões não respondidas não são consideradas como erradas, mas são consideradas como se não existissem. No TAI, esse tipo de procedimento pode prejudicar o desempenho do teste se o critério de parada não for alcançado. No caso de itens que não foram respondidos, existem diferentes alternativas para tratar essa situação: a) pontuar de modo que se penalizem esses itens, b) considerar esses itens errados, c) supor que eles sejam acertados por acaso, ou d) assumir que eles sejam acertados ou não, conforme a predição do modelo. Uma forma de solucionar esse problema foi proposta por Mills e Stocking (1996), onde, a partir do último item respondido pelo examinando, as respostas dos próximos itens administrados são consideradas incorretas até que o critério de parada seja atingido. Esse procedimento penaliza o candidato diminuindo o valor da estimativa da habilidade, o que desestimula o examinando a terminar a avaliação antes do tempo.

3.4. MÉTODOS DE CONTROLE DA EXPOSIÇÃO DE ITENS

Um problema comum nos métodos de seleção de itens é que eles muitas vezes apresentam os itens mais discriminativos no teste, tornando esses itens superexpostos e conhecidos, o que pode por em risco a segurança e a validade dos testes onde os itens devem ser sigilosos. Se um examinando conhece de antemão parte do conteúdo de um TAI, sua habilidade estimada terá um viés positivo (BARRADA; OLEA; ABAD, 2008). Além disso, os itens mais expostos são mais utilizados do que os outros, sendo que alguns podem nunca ser utilizados em um teste.. Wainer (2000a) afirma que de 15% a 20% dos itens de um banco correspondem a mais de 50% dos itens administrados em um teste. Essa situação exige o estabelecimento de métodos para controlar a exposição de itens, o que é, atualmente, um requisito de todo TAI que é aplicado muitas vezes. É importante salientar que todos os métodos utilizados para controlar a exposição dos itens têm alguma repercussão na precisão da estimação dos itens, já que tendem a não seguir estritamente o método MI.

O controle da exposição de itens tem como principais objetivos:

- Reduzir a exposição de itens muito expostos: Alguns critérios adotados para a seleção dos itens no TAI (por exemplo, o MI) tendem a administrar diversas vezes os itens com maior poder discriminativo. Com isso, esses itens podem acabar sendo conhecidos por indivíduos que irão fazer o teste, adicionando um erro na estimativa da sua proficiência que fica superestimada e, conseqüentemente, prejudicando a validade do teste (COSTA, 2009).
- Aumentar a exposição de itens poucos expostos: Existem itens que são pouco utilizados, com taxas de exposição próximas ou iguais a zero. Esta é uma propriedade indesejada nos TAIs, porque se um item é considerado inútil, deve ser eliminado do banco. Os itens que possuem os requisitos necessários para permanecerem no banco deveriam ser utilizados para se conseguir a maior variedade possível de itens entre as distintas aplicações de TAIs (GARCÍA; REVUELTA, 2003).
- Controlar o conteúdo: Os critérios adotados para a seleção dos itens podem interferir no controle de conteúdo de um teste. Dessa forma, o teste de um indivíduo pode conter muitas questões relacionadas a um assunto e poucas sobre outro assunto.

Diversos métodos de controle da exposição dos itens têm sido estudados (ARIEL; VELDKAMP; VAN DER LINDEN, 2004;

BARRADA; ABAD; OLEA, 2011; BARRADA; ABAD; VELDKAMP, 2009; BARRADA; MAZUELA; OLEA, 2006; BARRADA; OLEA; ABAD, 2008; BARRADA; OLEA; PONSODA, 2007; BARRADA; VELDKAMP; OLEA, 2009; BARRADA et al., 2008; 2009; 2011; BOYD; DODD; FITZPATRICK, 2003; BREITHAUPT; HARE, 2007; BURT; DAVIS; DODD, 2003; DOONG, 2009; CHANG; ANSLEY, 2003; CHANG; ANSLEY; LIN, 2000; CHANG; HARRIS, 2002; CHANG; QIAN; YING, 2001; CHANG; TWU, 1998; 2001; CHANG; VAN DER LINDEN, 2003; CHANG; YING, 1999; 2008; CHANG; ZHANG, 2002; CHEN, 2010; CHEN; ANKENMANN; CHANG, 2000; CHEN; ANKENMANN; SPRAY, 2003; CHEN; DOONG, 2008; CHEN; LEI, 2005; CHEN; LEI; LIAO, 2008; CHENG; CHANG, 2009; CHENG; CHANG; YI, 2007; CHENG et al., 2009; CHENG; LIOU, 2003; DAVEY; FAN, 2000; DAVEY; NERING, 2002; DAVEY; PARSHALL, 1995; DAVIS, 2004; DAVIS; DODD, 2001; 2003; 2008; DAVIS et al., 2003; DENG; CHANG, 2001; DIAO; VAN DER LINDEN; YEN, 2001; EDWARDS; THISSEN, 2007; EGGEN, 2001; FINKELMAN; WEISS; KIM-KANG, 2010; FRENCH; THOMPSON, 2003; GARCÍA; REVUELTA, 2003; GEORGIADOU; TRIANTAFILLOU; ECONOMIDES, 2007; GU; RECKASE, 2007; HALKITIS, 1998; HAN; HAMBLETON, 2004; HAU; CHANG, 2001; HETTER; SYMPSON, 1997; HONTANGAS et al., 2000b; IRAMANEERAT; STAHL, 2007; KENG et al., 2008; KINGSBURY; HOUSER, 1988; KINGSBURY; ZARA, 1989; LEE; IP; FUH, 2008; LEUNG; CHANG; HAU, 1999; 2002a; LEWIS, 2007; LI; SCHAFER, 2004; 2005a; LU; HAMBLETON, 2004; LUECHT, 2003; LUECHT; HADADI; NUNGESTER, 1996; LUNZ; STAHL, 1998; MCLEOD; LEWIS; THISSEN, 2003; MEIJER; NERING, 1999; MORRISON; SUBHIYAH; NUNGESTER, 1995; MUCKLE et al., 2005; MULDER; VAN DER LINDEN, 2009; PARSHALL; DAVEY; NERING, 1998; PARSHALL; HARMES; KROMREY, 2000; PARSHALL; HOGARTY; KROMREY, 1999; PARSHALL et al., 2001; 2002; PASTOR et al., 1999; PASTOR; DODD; CHANG, 2002; REVUELTA; PONSODA, 1996; 1998a; REVUELTA; PONSODA; OLEA, 1998; RILEY; DENNIS; CONRAD, 2010; SEGALL, 2003; 2004; STOCKING, 1993; STOCKING; LEWIS, 1995; 1998; 2000; SWANSON; STOCKING, 1993a; 1993b; SYMPSON; HETTER, 1985; THOMASSON, 1995; 1998; TRUELL; ZHAO; ALEXANDER, 2005; VAN DER LINDEN, 1995; 2000; 2003; 2005a; 2005b; 2010a; VAN DER LINDEN; VELDKAMP, 2004; 2007; VEERKAMP; BERGER, 1997; VEERKAMP; GLAS, 2000; VELDKAMP; VAN DER LINDEN,

2002; 2008; VELDKAMP; VERSCHOOR; EGGEN, 2010 WEN; CHANG; HAU, 2000; YI, 2002). Uma relação de métodos de controle da exposição de itens desenvolvidos entre 1983 e 2005 pode ser encontrada em Georgiadou; Triantafillou; Economides, (2007) e outra em Van der Linden (2010a). Dentre esses métodos, destacam-se aqueles que se referem ao controle da freqüência da exposição dos itens e ao balanceamento do conteúdo.

3.4.1. Controle da Taxa de Exposição do Item

A taxa de exposição de um item é a probabilidade dele ser administrado em um TAI. Na prática, essa taxa é estimada mediante a proporção de aplicações do item. Os critérios adotados para seleção de itens podem acabar dividindo o banco de itens em duas partes, segundo a informação de cada um deles: os superexpostos e os subexpostos. Isso acontece, por exemplo, se o TAI for delineado de forma que os indivíduos iniciam o teste com a mesma estimativa provisória da proficiência (COSTA, 2009) ou quando muitos indivíduos possuem o mesmo nível de habilidade (SEGALL, 2005), sob o critério de seleção dos itens pela MI. Dessa forma, o item mais informativo será o mesmo para todos os indivíduos, o segundo item será um entre as duas escolhas (no caso dos modelos dicotômicos): será um item se a resposta for correta ou outro após uma resposta incorreta, e assim por diante. Conseqüentemente, a seqüência da administração dos itens será previsível e os itens iniciais serão usados com mais freqüência. Os itens que possuem uma alta probabilidade de utilização podem passar a ser de domínio público, tornando superestimados os escores do teste e prejudicando a sua validade. Por outro lado, não faz sentido ter no banco itens que possuem uma baixa (ou nenhuma) probabilidade de utilização. Segundo Hornke (2000), existem bancos onde mais de 80% dos itens não são selecionados para os testes. O controle da taxa de exposição do item consiste em limitar a frequência de exposição dos itens, principalmente dos mais informativos, a fim de melhorar a segurança do teste (SEGALL, 2005). Entre os diversos procedimentos para o controle da exposição de itens, destacam-se: procedimentos probabilísticos e os métodos de estratificação do banco de itens.

3.4.1.1. Procedimentos Probabilísticos

Hetter e Sympson (1997) desenvolveram o procedimento *Sympson-Hetter* (SH) de seleção condicional dos itens, que calcula parâmetros de exposição do item para controlar probabilisticamente a freqüência com a qual o item é selecionado. O procedimento SH tem

sido o mais utilizado atualmente (COSTA, 2009). Esse procedimento consiste em especificar o valor máximo esperado para a taxa de exposição do item para o teste e construir uma tabela de Informação (*infotable*) que consiste em uma lista das informações dos itens por proficiência. Os passos para a execução desse procedimento são descritos em Costa (2009).

Quanto ao valor para a taxa de exposição de itens, Way (1998) sugere uma exposição máxima de 15%, enquanto que Olea et al. (2004) utilizaram 25% e desenvolvedores de testes com alto risco utilizam 10% (SEGALL, 2005). Van der Linden e Glas (2000b) ressaltam que a taxa de exposição do item nunca deve ser menor que a razão entre o tamanho do teste e a quantidade de itens no banco. Como na prática, um banco de itens é tipicamente 7 a 10 vezes o tamanho do teste, os valores mínimos para as taxas de exposição seriam de 0,10 a 0,14. Geralmente, os valores de *r* mais utilizados estão entre 0; 20 e 0; 30.

A restrição no controle dos itens com tendência à superexposição poderá causar uma redução na informação do teste para a estimação da proficiência. Porém, a quantidade de informação perdida não será grande se os itens do banco foram de alta qualidade e se o algoritmo conseguir administrar itens com informação levemente menor que o item mais informativo para a estimativa da proficiência.

Revuelta e Ponsoda (1998a) propuseram o Método Restrito que consiste em fixar para cada item uma taxa máxima de exposição que, quando atingida, não permite que o item seja administrado, até que a sua taxa de exposição fique abaixo da taxa máxima.

Barrada et al. (2009b) desenvolveram um método denominado Taxa de Exposição Máxima Múltipla, no qual cada item do banco tem uma taxa máxima de exposição. Comparado com o tradicional método SH que utiliza uma única taxa de exposição máxima, esse método proporciona um uso mais balanceado do banco de itens e retarda a possível distorção de estimação relacionada com a segurança do banco.

Chang e Ying (1999) advertem que o procedimento probabilístico possui duas limitações: (1) os itens que não foram selecionados não podem ser administrados, o que faz com que os itens que possuem baixa probabilidade de serem selecionados continuem apresentando baixas taxas de exposição; e (2) os parâmetros de controle da exposição necessitam de atualização através de um número grande de complicadas simulações a cada alteração do banco de itens ou se a distribuição das proficiência da população de interesse for modificada. Para tentar contornar tal situação, alguns métodos de estratificação do banco de itens foram propostos.

3.4.1.2. Métodos de estratificação do banco de itens

Um dos primeiros métodos de estratificação do banco de itens foi proposto por Chang e Ying (1999), onde o banco de itens é divido em diferentes estratos baseado nos valores dos parâmetros dos itens e o teste adaptativo é dividido em estágios. No método de estratificação pelo parâmetro *a*, por exemplo, divide-se os itens do banco em diversos estratos em ordem ascendente dos valores de *a*. Cada teste consiste em um número idêntico de estágios e estratos. O primeiro estágio consiste em administrar itens com menores parâmetros *a* selecionados do primeiro estrato. Os estágios subseqüentes selecionarão itens mais discriminativos que pertencem aos diferentes estratos. Isso porque nos estágios iniciais do teste o ganho na informação usada pelos itens mais discriminativos não é adequado já que a estimação da proficiência ainda é relativamente imprecisa. Conseqüentemente, os itens com valores mais altos de *a* devem ser usados nas fases finais do teste.

Chang, Qian e Ying (2001) mencionam que, ao se estratificar o banco de itens pelo parâmetro *a*, alguns bancos de itens podem não possuir itens suficientes com baixos valores do parâmetro de dificuldade no último estrato do teste. Dessa forma, eles desenvolveram outro método de estratificação do banco pelo parâmetro *a*, onde o banco de itens é dividido em pequenos níveis baseados nos parâmetros de dificuldade dos itens. Em relação a cada nível, itens são classificados na ordem ascendente dos valores de *a*. Em seguida, itens com menores valores de discriminação de cada nível são agrupados no primeiro estrato, itens com os segundos menores valores de a são agrupados no segundo estrado e assim por diante. Assim, o último estrato conterá os itens mais discriminativos de cada nível do parâmetro da dificuldade.

Uma estratégia alternativa para evitar a superexposição de itens consiste em dividir o grande banco de itens em vários sub-bancos e rotar o emprego deles no teste. Barrada et al. (2008) compararam esse método com o método da taxa máxima de exposição de SH e verificaram que o método de rotar ofereceu resultados ligeiramente melhores.

3.4.2. Balanceamento de Conteúdo

Os algoritmos de seleção de itens que maximizam a precisão podem não administrar de forma harmoniosa os diferentes conteúdos do teste, resultando em resultados que têm validade questionável (SEGALL, 2005). Em muitas situações, o teste precisa levar em consideração a administração de itens de todos os assuntos do teste, ou seja, deve haver um balanceamento de conteúdo (BELOV;

ARMSTRONG; WEISSMAN, 2008; CHANG; YING, 2008; CHENG et al., 2009; CHENG; CHANG; YI, 2007; DAVIS et al.,2003; KINGSBURY; ZARA, 1989; LEUNG; CHANG; HAU, 1999; 2001; 2002a; 2003; LEUNG et al., 2003; LUECHT; DECHAMPLAIN; NUNGESTER, 1997; LUECHT; HADADI; NUNGESTER, 1996; MORRISON; SUBHIYAH; NUNGESTER, 1995; OLEA; PONSODA; PRIETO, 1999; WAINER; KIELY, 1987). Esse balanceamento pode ser inserido no algoritmo do teste, por exemplo, a seleção de itens pelo critério MI pode intercalar um item de cada conteúdo até atingir a quantidade mínima de itens de cada conteúdo.

O método proposto por Kingsbury e Zara (1989) é uma modificação do procedimento de seleção do item pela Máxima Informação levando também em conta a categoria do conteúdo de cada item no processo de seleção. Uma vez que o item é selecionado pela MI para o atual respondente, se o item selecionado representa um assunto da área do conhecimento que ainda não foi representado no teste, o item é administrado. Caso contrário, o item que oferece a próxima maior informação é avaliado em relação aos assuntos estabelecidos e o processo é repetido até que todos os assuntos sejam abrangidos conforme o que foi especificado.

Stocking e Swanson (1993) propuseram um procedimento para formular matematicamente as restrições estabelecidas (por exemplo, um limite mínimo e máximo de itens para cada categoria de conteúdo) junto a outras restrições estatísticas. Na seleção de um item, se considera o valor que cada item disponível no banco proporciona em uma função matemática onde se pondera de maneira diferente os desvios a respeito dos limites estabelecidos para cada restrição.

Existem situações onde o procedimento de seleção de itens individuais não é adequado. Para esses casos, Wainer e Kiely (1987) estabeleceram unidades de análises alternativas aos itens que consistiam em pequenos grupos de itens, chamados de *testlets*, referentes a um mesmo conteúdo. São criados vários *testlets*, com diferentes níveis de dificuldade, de um determinado conteúdo. Assim, é a seleção de *testlets* é que vai se adaptar ao nível do indivíduo, e não a seleção de itens individuais. Os *testlets* são formados por grupos pequenos de itens pertencentes a um mesmo conteúdo e, portanto, ao mesmo traço latente, e que se aplicam juntos, ou seja, todos os itens do *testlet* selecionado, de acordo com a habilidade atual estimada, são aplicados ao indivíduo. Nesse caso, são criados vários *testlets* que abordam diferentes conteúdos de um mesmo traço latente (caso unidimensional) ou de diferentes traços latentes (caso multidimensional).

Os estudos têm se direcionado em buscar a melhor maneira de agrupar os itens em *testlets* e organizá-los em estruturas hierarquizadas (ARIEL; VELDKAMP, 2006; ARMSTRONG et al., 2004; BELOV; ARMSTRONG, 2008; BOYD; DODD; FITZPATRICK, 2003; BRADLOW; WAINER; WANG, 1999; BREITHAUPT; HARE, 2007; DAVIS; DODD, 2001; 2003; GLAS; WAINER; BRADLOW, 2000; KENG et al., 2008; LI; LI; WANG, 2010; LUECHT; BRUMFIELD; BREITHAUPT, 2002; MURPHY; DODD; VAUGHN, 2010; REESE; SCHNIPKE; LUEBKE, 1997; SCHNIPKE; GREEN, 1995; SHEEHAN; LEWIS, 1992; THISSEN; STERNBERG; MOONEY, 1989; THOMPSON; DAVEY, 1999; VOS; GLAS, 2000; 2010; WAINER; BRADLOW; DU, 2000; WAINER; BRADLOW; WANG, 2007; WAINER; KAPLAN; LEWIS, 1992; WAINER; KIELY, 1987; WAINER; WANG, 2000; WAINER et al., 1991).

3.4.3. Outros Métodos

Outras formas de controlar a exposição dos itens têm sido realizadas com a utilização da programação linear (ADEMA, 1990; ADEMA; BOEKKOOI-TIMMINGA; VAN DER LINDEN, 1991; THEUNISSEN, 1985; 1986; VAN DER LINDEN, 2000; 2010a; VAN DER LINDEN; ADEMA, 1998; VAN DER LINDEN; BOEKKOOI-TIMMINGA, 1989; VAN DER LINDEN; REESE, 1998). Nesse procedimento, a decisão sobre os itens que vão compor o teste se resolve maximizando a informação que o teste proporciona, levando em conta um conjunto de restrições.

Um algoritmo ótimo deve selecionar seus itens sequencialmente de modo a permitir a escolha dos itens que otimizam o teste e a realizar todas as restrições simultaneamente. As restrições podem ser as quantidades máxima e mínima de itens no teste e de cada conteúdo, itens que não podem ser aplicados num mesmo teste (*enemy itens*) (WEISS, 2011), taxa de exposição do item, etc. Para isso, Van der Linden e Reese (1998) propuseram a utilização do chamado teste *shadow*. Um teste *shadow* é um teste que reúne todas as restrições do teste, contém todos os itens já administrados ao examinando e fornece a informação máxima para a habilidade atual estimada (DIAO; VAN DER LINDEN, 2011; VAN DER LINDEN, 2000; 2005b; 2008a; 2010a; VAN DER LINDEN; ARIEL; VELDKAMP, 2006; VAN DER LINDEN; CHANG, 2003; VAN DER LINDEN; VELDKAMP, 2004). Os passos para utilizar o teste *shadow* são: a) definição da habilidade inicial, b) construção um teste *shadow* que cumpra as restrições e seja o mais informativo para a habilidade atual, c) aplicação do item mais

informativo para o nível atual de habilidade, d) reestimação da habilidade, e) atualizar o modelo do teste incluindo o item administrado no próximo teste *shadow*, f) retornar todos os itens não utilizados para o banco de itens; g) repetir-se os passos b), c), d) e) e f) até a administração de *n* itens.

Outra estratégia utilizada é a criação de listas de itens que não devem ser administrados conjuntamente (WAY; DAVIS; FITZPATRICK, 2006). Existem itens que podem fornecer informações que podem ajudar o indivíduo a encontrar a resposta correta de outros itens (OLIVEIRA, 2002). Esse procedimento ajuda a melhorar a validade de conteúdo do teste, porém pode prejudicar a eficiência da estimação da habilidade se essas listas forem muito extensas.

Outra situação que pode ocorrer é um indivíduo realizar o teste mais de uma vez e ser submetido a um item já administrado a ele (LUNZ; O'NEILL, 1998;) ou semelhante a algum item administrado a ele. Uma solução possível é criar um registro, relacionado com a identificação do indivíduo, que possa identificar a quantidade de vezes que ele fez a avaliação e os itens que já foram administrados a ele (KINGSBURY; ZARA, 1989).

Um método alternativo é a criação de múltiplos bancos de itens que vão sendo utilizados de forma rotatória, especialmente útil nos TAIs que se aplicam muitas vezes (ARIEL; VELDKAMP; VAN DER LINDEN, 2004; MILLS; STEFFEN, 2000; STOCKING SWANSON, 1998). Estudos mostraram que esse procedimento pode ter resultados melhores do que utilizar métodos para controlar a exposição dos itens (BARRADA; OLEA; ABAD, 2008).

Nos testes para fins de classificação, uma restrição adicional que tem sido sugerida são as regras de truncamento (FINKELMAN, 2008). O objetivo deste é abordar a situação em que um examinando tem um nível de habilidade muito próximo do ponto de corte e os testes adaptativos de classificação podem continuar até que o banco de itens está esgotado, sem conseguir tomar uma decisão de classificação. A utilização dessa restrição permite que o algoritmo do teste reconheça esta situação e determine quando os demais itens no banco não têm informações suficientes para tomar uma decisão de classificação, mesmo que o examinando responda todos os itens restantes no banco de itens. Esse critério é mais adequado do que, por exemplo, estabelecer uma quantidade limite de itens para ser aplicado.

3.5. VALIDADE E PRECISÃO DE UM TAI

Antes de um TAI ser definitivamente implantado e utilizado, é necessário verificar algumas propriedades psicométricas, através de simulações ou estudo empírico. Green et al. (1983) aborda nove dimensões diferentes nas análises dos TAIs, relativas ao conteúdo, dimensionalidade, confiabilidade, validade, estimação de parâmetros, ancoragem, características dos banco de itens, seleção de itens e fatores humanos. Vários desses aspectos já foram tratados nas seções anteriores desse trabalho e outros possuem a mesma forma de tratamento dos testes tradicionais da TCT (GARCÍA JIMÉNEZ; GIL; RODRÍGUEZ GÓMEZ, 1998). Nessa seção serão abordadas as questões relacionadas com a a precisão, a validade e o viés nos TAIs.

3.5.1. Precisão

Um algoritmo de teste é considerado mais eficiente do que outro se ele fornece uma estimativa do traço latente mais precisa com a mesma quantidade de itens aplicados ou se ele fornece a mesma precisão dessa estimativa com uma menor quantidade de itens aplicados (SEGALL, 2005). Uma das vantagens fundamentais da TRI está no fato que proporciona medidas de precisão condicionadas aos diferentes níveis de habilidade, ou seja, diferentes medidas de precisão para os distintos respondente. Dessa forma, a eficiência de um TAI pode ser verificada através de estudos empíricos ou simulação, em relação ao seguintes aspectos (MUÑIZ; HAMBLETON, 1999):

- Erro Padrão Médio (EPM): Obtém-se através da média do erro padrão das habilidades estimadas, onde quanto menor for o valor, melhor é a precisão. O erro padrão da habilidade de um indivíduo é a raiz quadrada do inverso da soma das informações fornecidas por cada item administrado a ele.
- Raiz Quadrada do Erro Quadrado Médio (RQEQM): Calcula-se o quadrado da diferença entre a habilidade estimada e o valor do seu parâmetro (geralmente obtido por simulação), soma-se todos os resultados, divide-se pela quantidade de habilidades estimadas e extrai-se a raiz quadrada. Quanto menor for esse valor, melhor será a precisão.
- Desvio Empírico Médio (DEM): Calcula-se a média das diferenças entre a habilidade estimada e o valor do seu parâmetro, onde quanto mais próximo de zero for o valor, melhor será a precisão. Um viés positivo indica que as habilidades estão superestimadas, enquanto

que um viés negativo indica que as habilidades estão superestimadas subestimadas.

- Eficiência (EF): Define-se pela quantidade média de itens necessários para alcançar um erro padrão pré-determinado utilizado como critério de parada em testes de tamanho variável. O teste mais eficiente é aquele que consegue atingir o erro padrão pré-determinado com a menor quantidade de itens aplicados.
- Correlação linear (CL): Calcula-se a correlação entre as estimativas das proficiências e o seu parâmetro. Quanto maior for o valor, maior será a precisão.
- Procedimentos da TCT: Outros procedimentos da TCT podem ser utilizados, por exemplo, os coeficientes Alfa de Crombach, teste-reteste, formas paralelas, entre outros (CROMBACH; GLESER, 1957; COHEN, 1960; GULLIKSEN, 1950).

3.5.2. Validade

Como qualquer outro teste, o TAI também precisa submeter-se aos exames de validade, além de considerar algumas situações particulares (MUÑIZ; HAMBLETON, 1999):

- Validade de Conteúdo: as restrições implementadas nos algoritmos de seleção de itens podem ajudar a amostra de itens a ser representativa de todos os conteúdos estabelecidos pelos especialistas na elaboração do banco de itens. Assim, o conteúdo de um TAI é validado pelo procedimento de balanceamento de conteúdo, discutido na seção 3.4.2.
- Validade Preditiva: Dependendo do contexto onde o TAI é aplicado, pode ser de interesse correlacionar seus resultados com medidas externas. Por exemplo, no ASVAB as proficiências estimadas são correlacionadas com as qualificações que os recrutas obtêm em cursos após o treinamento militar.
- Validade de Construto: A verificação sobre a dimensionalidade do banco de itens, discutida na Seção 2.1, pode ser considerada o primeiro estudo sobre a validade de construto. Outro estudo pode ser feito, no caso de testes que antigamente eram realizados via "papel e lápis", a fim de verificar a equivalência entre as duas versões e os ganhos da utilização do TAI em termos eficiência e de economia de tempo e de recursos financeiros.

3.5.3. Viés

Outra questão importante para verificar num TAI está relacionada com a presença de um viés (ou vício). O viés refere-se ao desvio das estimativas do seu verdadeiro valor e pode ser detectado pelo desvio empírico médio (DEM), definido na Seção 3.5.1. Foi visto anteriormente, na Seção 2.5, que a utilização do Método MV produz um viés na estimação de valores altos e baixos da habilidade, onde valores altos são superestimados e valores baixos são subestimados, e que a utilização do Método Bayesiano produz um viés contrário, ou seja, valores altos são subestimados e valores baixos são superestimados (KIM; NICEWANDER, 1993), embora nem sempre aconteça dessa forma dependendo do critério de parada do teste (YI; WANG; BAN, 2000). Wang e Wang (2001) afirmam que todos os procedimentos produzem viés quando se utilizam bancos de itens reais, não simulados. Entretanto, quando o nível de dificuldade dos itens se ajusta ao nível de habilidade do examinando, o viés é mínimo (WANG; VISPOEL, 1998).

Para corrigir o viés, tem-se adotado métodos corretivos e preventivos (LORD, 1983; 1986; SAMEJIMA, 1998; VISPOEL; WANG; BLEILER, 1997; WANG; VISPOEL, 1998; WANG, 1997; WANG; HANSON; LAU, 1999; WANG; WANG, 2001; WARM, 1989; YI; WANG; BAN, 2000). Os métodos corretivos atuam após a obtenção da estimação através de uma fórmula corretiva, enquanto que os métodos preventivos atuam antes da estimação através da modificação da função de maximização utilizando distribuições a priori não informativas (HONTANGAS et al., 2000a).

3.6. CONSIDERAÇÕES FINAIS SOBRE TAIS

Nesse capítulo, buscou-se levantar uma grande quantidade de referências sobre os Testes Adaptativos Informatizados, onde foram priorizadas as publicações mais recentes em periódicos. Outra extensa lista de referências sobre TAIs, atualizada periodicamente, está disponível em University of Minnesota (2010), página da Web elaborada e mantida por David J. Weiss. Além disso, essa página oferece várias informações atualizadas sobre TAIs, tais como, programas de avaliações que utilizam TAIs, livros, softwares e congressos sobre TAIs.

Congressos especificamente sobre TAIs têm sido realizados, dentre eles, destacam-se: *Computerized Adaptive Testing Conference* em 1977 e 1979, *Item Response Theory and Computerized Adaptive Testing Conference* em 1982, GMAC (*Graduate Management Admission Council*) *Conference on Computerized Adaptive Testing*, em

2007 e 2009, IACAT (*International Association for Computerized Adaptive Testing*) *Conference on Computerized Adaptive Testing* em 2010 e 2011 (UNIVERSITY OF MINNESOTA, 2010). O interesse em TAIs também é evidente em outros congressos, como os elaborados pelo *National Council on Measurement in Education* (NCME) e pela *American Educational Research Association* (AERA), onde as contribuições relacionadas com TAI crescem a cada edição (MEIJER; NERING, 1999; PONSODA, 2000).

Em alguns indivíduos em que se aplica um TAI, ocorre um padrão de resposta inapropriado (responde a mesma alternativa em todos os itens ou de forma aleatória, ou responde muito rápido ou muito devagar), onde o resultado da habilidade estimada é incorreta e não representa a habilidade real. Isso pode ocorrer devido à "cola", descuido, ou conhecimento prévio dos itens. Técnicas para detectar padrões de respostas inconsistentes nos TAIs estão sendo desenvolvidas (BELOV; ARMSTRONG, 2010; CHANG et al., 2011; EGEBERINK, 2010; KINGSBURY; HOUSER, 2008; MEIJER; VAN KRIMPEN-STOOP, 2010; PONSODA et al., 2004; RILEY; DENNIS; CONRAD, 2010; VAN DER LINDEN; GUO, 2008; VAN KRIMPEN-STOOP; MEIJER, 2000, 2002).

Além dos TAIs tradicionais, algumas variações do teste têm sido desenvolvidas. Uma delas são os Testes Sequenciais Adaptativos Informatizados (*Computer Adaptive Sequential Testing - CAST*), onde o teste inicia com um *testlet* de dificuldade média e, com base na pontuação do aluno sobre esse *testlet*, é selecionado um próximo *testlet* e assim sucessivamente (LUECHT; NUNGESTER, 1998; 2000).

Lord (1980) desenvolveu os testes multietápicos informatizados, com *k* níveis, de tal forma que nos últimos níveis se situam testes onde cada um é apropriado para o seu nível de habilidade.

Outra variação dos TAIs é o teste autoadaptativo informatizado (*Self-adapted testing*) (HONTANGAS; OLEA; PONSONDA, 1998; HONTANGAS et al., 2000b; 2004; OLEA; PONSODA, 1996; PITKIN; VISPOEL, 2001; PONSODA et al., 1997; REVUELTA, 2004; 2010; ROCKLIN; O´DONNELL, 1987; ROCKLIN; O´DONNELL; HOLST, 1995; ROOS; WISE; PLAKE, 1992; 1997; VISPOEL, 1998b; VISPOEL; ROCKLIN; WANG, 1994; WISE, 1999b; WISE; PONSODA; OLEA, 2002; WISE et al., 1994), no qual o examinando tem o controle sobre o teste escolhendo o nível de dificuldade do teste ou dos itens, antes do início do mesmo. Estudos mostram que o fato do examinando possuir o controle sobre o nível de dificuldade dos itens diminui sua ansiedade e melhora seu desempenho (ROCKLIN;

O´DONNELL; HOLST, 1995). No entanto, algumas investigações teóricas mostraram que os métodos de estimação de capacidade podem apresentar inconvenientes e o teste autoadaptativo parece mais adequado para medir características de personalidade do que para medir a capacidade (REVUELTA, 2010).

O teste de múltiplos estágios (do inglês, *Multiple Stage Adaptive test* - *MST*) (ADEMA, 1990; BELOV; ARMSTRONG, 2008; BREITHAUPT; ARIEL; HARE, 2010; BREITHAUPT; HARE, 2007; EDMONDS; ARMSTRONG, 2009; LUECHT; NUNGESTER, 1998; MELICAN; BREITHAUPT; ZHANG, 2010; VAN DER LINDEN; ADEMA, 1998; ZENISKY; HAMBLETON; LUECHT, 2010) é outra variação de TAI. Nesses testes, os examinandos realizam uma sequência de subtestes, movendo-se para um teste mais difícil ou mais fácil, conforme os resultados obtidos.

Häusler (2006) propôs um teste alternativo ao teste adaptativo informatizado clássico, chamado de Controle de Sucesso Adaptativo (*Adaptive Success Control*) que propõe minimizar a duração do teste ao invés de minimizar a quantidade de itens aplicados. A idéia é que indivíduos que respondem itens mais fáceis se sentem mais seguras e conseguem responder o item em menos tempo do que se respondesse um item aplicado por um TAI clássico.

Outros estudos têm se dedicado em avaliar o efeito do *feedback* do item (*item feedback*) nos testes adaptativos (ROOS; WISE; PLAKE, 1992; 1997; WEISSMAN, 2006). O *feedback* fornece ao indivíduo informações sobre o seu desempenho nos itens respondidos durante a realização do teste adaptativo.

Os TAIs também são utilizados em bateria de testes (COSTA, 2009; VAN DER LINDEN, 2010b; VAN DER LINDEN; GLAS, 2010). Uma bateria de teste consiste na aplicação sequencial de vários testes diferentes compostos com diferentes bancos de itens.

Almond e Mislevy (1999) abordaram o TAI baseado na TRI sob a perspectiva da Modelagem Gráfica (*Graphical Modeling* - *GM*). Segundo os autores, a Modelagem Gráfica fornece métodos para fazer inferências sobre as habilidades multifacetadas e conhecimento, e para extrair dados de performances complexas.

Com o avanço da tecnologia, os testes adaptativos também têm sido desenvolvidos para serem implantados em dispositivos móveis, tais como telefones celulares, com a denominação de CAT-MD (*Computerized Adaptive Testing on Mobile Devices*) (TRIANTAFILLOU; GEORGIADOU; ECONOMIDES, 2008).

Diversos estudos citados nesse Capítulo comprovam a eficiência do desempenho de um TAI baseado na TRI. Entretanto, é possível encontrar outros tipos de TAIs não baseados na TRI, que utilizam, por exemplo, *Sequential Probability Ratio Test – SPRT* (RECKASE, 1983), redes bayesianas (COLLINS; GREER; HUANG, 1996; DESCOVI, 2009; GROENWALD; RUIZ, 2006; VOMLEL, 2004), *Measurement Decision Theory* (RUDNER, 2002), Lógica Nebulosa (*Fuzzy Logic*) (SUÁREZ, 2003), Modelo de Diagnóstico Cognitivo (*cognitive diagnosis with computer adaptive testing* – CD-CAT) (CHENG, 2009; 2010; GIERL; ZHOU, 2008; WANG; CHANG, 2011b; ZHANG, 2008), o método Zinnes-Griggs (STARK; CHERNYSHENKO; GUENOLE, 2011), *Number Right Elimination Testing* – NRET (HOE et al., 2009) e a Teoria Clássica dos Testes (DALPIAZ, 2007, p. 42; DAMANDO, 2003, p. 76). Contudo, a TRI é a metodologia preferida para o desenvolvimento de TAIs (THISSEN; MISLEVY, 2000), sendo que um dos principais motivos é o fato de colocar os respondentes e os itens na mesma métrica o que beneficia o procedimento de seleção dos itens (WISE; KINGSBURY, 1984).

Esse Capítulo teve como objetivo apresentar os princípios básicos dos TAIs e levantar as principais literaturas existentes sobre o assunto. Entretanto, existem muitas outras literaturas sobre TAI que não foram mencionadas aqui, pois isso seria uma tarefa quase que inesgotável, uma vez que as pesquisas sobre TAIs crescem a cada dia. A teoria e a prática da seleção de itens e da estimação da habilidade para TAIs ainda estão em evolução (VAN DER LINDEN; PASHLEY, 2010).

Conforme Thompson e Prometric (2007), existem várias opções para cada componente de um TAI, o que cria uma infinidade de possíveis projetos. Diante da complexidade de TAIs e da variedade de abordagens disponíveis, como deve-se proceder? Van der Linden e Pashley (2010) sugerem definir alguns arranjos viáveis e realizar um estudo de simulação a fim de identificar uma solução adequada, mas não necessariamente a solução ótima. Segundo Segall (2005), os estudos de simulações são muito utilizados nos TAIs pois permitem comparar métodos distintos, validar novos métodos, procedimentos ou algoritmos, verificar tamanho de amostra necessário para atingir determinada precisão, etc. Por meio dos resultados das simulações, é possível identificar situações onde o teste não está tendo um bom desempenho e corrigi-las, por exemplo, aumentando o número de itens, melhorando a qualidade dos itens, modificando o algoritmo de seleção de itens ou alterando as restrições.

4. METODOLOGIA

Segundo Richardson (2010), metodologia são procedimentos e regras utilizadas por determinado método. Para Pacheco Junior et al. (2007), método é o modo com se chega a determinado objetivo, porém, para Marconi e Lakatos (2007), método é a ordem imposta aos diferentes processos necessários para atingir um determinado objetivo. Esse Capítulo apresenta a caracterização metodológica da pesquisa e os procedimentos metodológicos utilizados para o desenvolvimento desse trabalho, tanto da sistemática proposta quanto do estudo de caso.

4.1. CARACTERIZAÇÃO DA PESQUISA

A pesquisa científica pode ser caracterizada em relação à natureza, à abordagem, ao objetivo e aos procedimentos técnicos.

Em relação à natureza, essa pesquisa é do tipo teórico-aplicada, pois visa propor, aplicar e validar uma sistemática para a implantação de TAIs. Gil (2010) considera que o objetivo da pesquisa aplicada é gerar conhecimentos com o propósito de aplicá-los em uma determinada situação.

Em relação à abordagem, essa pesquisa é do tipo quantitativa, pois utiliza perguntas baseadas em instrumento, dados de desempenho e análise estatística para o tratamento dos dados (CRESWELL, 2007). Richardson (2010) destaca que o método quantitativo é amplamente utilizado na condução das pesquisas, representando a intenção de garantir a precisão dos resultados e evitar distorções de análise e interpretação, possibilitando uma margem de segurança quanto às inferências.

Em relação ao objetivo, essa pesquisa pode ser classificada como exploratória, segundo Gil (2010). A pesquisa é exploratória porque utiliza pesquisa bibliográfica e estudo de caso com objetivo de proporcionar uma familiarização do tema estudado. Nesse trabalho, a pesquisa bibliográfica sobre TAIs é utilizada para a elaboração de uma sistemática para a implantação de TAIs e o estudo de caso é utilizado para a aplicação dessa sistemática.

Em relação aos procedimentos técnicos, essa pesquisa pode ser classificada como bibliográfica, experimental e estudo de caso, segundo Gil (2010). Essa pesquisa compreende os procedimentos técnicos de pesquisa bibliográfica, pelo exaustivo levantamento de referências relacionadas aos TAIs, utilizadas para a elaboração da sistemática proposta. Também são utilizados procedimentos experimentais, pela utilização de simulações e análises de dados. A pesquisa também se

caracteriza por ser estudo de caso, pois se aplica para o caso da avaliação teórica do DETRAN-SC.

4.2. ETAPAS DA PESQUISA

Essa pesquisa foi basicamente constituída por uma parte teórica caracterizada pelo levantamento de referências bibliográficas e pela elaboração da sistemática proposta, e por uma etapa prática caracterizada pela utilização da sistemática proposta em um estudo de caso.

Na parte teórica, primeiramente foi feito um breve levantamento bibliográfico sobre a TRI (teoria que dá suporte aos TAIs), disponibilizado no Capítulo 2, com ênfase no ML3. Em seguida, foi realizado um exaustivo levantamento bibliográfico sobre TAIs, disponibilizado no Capítulo 3, incluindo testes, relatórios técnicos, artigos de periódico e congressos de várias épocas. Porém, foram priorizadas as publicações de periódicos científicos, principalmente as mais recentes. Na sequência, foi elaborada a sistemática proposta, apresentada no Capítulo 4.3, com base no levantamento bibliográfico.

Na parte prática, foi aplicada a sistemática desenvolvida em um estudo de caso para a avaliação teórica do DETRAN-SC para a obtenção da carteira de habilitação para motoristas, onde os resultados estão apresentados no Capítulo 5.

Nesse estudo de caso, a população compreende todas as pessoas do estado de Santa Catarina com potencial para serem motoristas de qualquer tipo de veículo que necessita da utilização da CNH (Carteira Nacional de Habilitação) para ser conduzido. A amostra é com reposição e se constitui 221.933 provas com 40 itens, respondidas por 178.828 candidatos, e aplicadas no ano de 2008. Esses dados foram obtidos no CIASC-SC (Centro de Informática e Automação do Estado de Santa Catarina), com a devida autorização e permissão do DETRAN-SC, para fins de estudos acadêmicos. Dessa forma, não foi feita nenhuma elaboração de itens e nem levantamento de dados, porém foram utilizados os 462 itens e as respostas dos candidatos existentes no banco de dados fornecido.

Na análise do estudo de caso, foram utilizados os softwares TESTFACT (BOCK et al., 2003) para a análise da dimensionalidade, BILOG-MG (TOIT, 2003) para a calibração dos itens (fase 2 do Software), e CATSim (WEISS; GUYER, 2010) para as simulações das respostas e aplicação dos algoritmos dos TAIs.

4.3. DEFINIÇÃO DA SISTEMÁTICA PARA A IMPLANTAÇÃO DE TESTES ADAPTATIVOS INFORMATIZADOS – SITAI

Este capítulo descreve a Sistemática para a Implantação de Testes Adaptativos Informatizados (SITAI) proposta nesse trabalho e desenvolvida a partir do referencial teórico discutido no Capítulo 3. A sistemática consiste em 8 etapas, conforme a Figura 15, que engloba todo o processo para a implantação de um TAI desde a sua definição até a sua efetiva aplicação e manutenção.

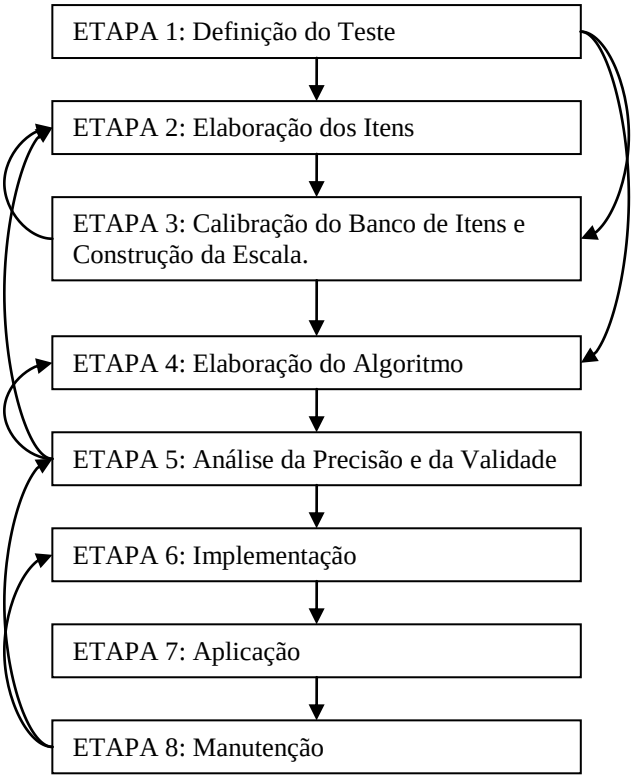


Figura 15. Sistemática para a Implantação de TAIs

Teoricamente, o TAI é uma idéia relativamente simples, ou seja, consiste na elaboração de um algoritmo que seleciona itens adequados ao examinandos de forma adaptativa à sua habilidade ou proficiência. Entretanto, a realidade do planejamento, implementação e manutenção

de um programa de TAI é substancialmente mais complexa (WISE; KINGSBURY, 2000). O desenvolvimento de um teste consistente requer uma grande equipe multidisciplinar, composta por profissionais de várias áreas, além de recursos de infra-estrutura e computacionais com hardware e software apropriados. O desenvolvimento do software de um TAI requer implementação de algoritmos complicados que utilizam métodos estatísticos complexos, combinados com regras de seleção de itens e restrições.

No desenvolvimento da SITAI, optou-se por criar uma metodologia dividida em 8 etapas. No caso da criação de um teste desde seu começo, todas as etapas deverão ser percorridas, sendo que pode ser necessário refazer alguma etapa, em algumas situações, conforme mostra a Figura 15. No caso da adaptação de um teste convencional, além da possibilidade de refazer alguma etapa, as Etapas 2 e 3 poderão ser "puladas".

A Etapa 1 (definição do teste) consiste em verificar a existência prévia do teste e definir a dimensão, o traço latente e o objetivo do teste. A Etapa 2 (elaboração dos itens) consiste em utilizar uma equipe multidisciplinar de especialistas para elaborar, definir tipos e quantidade de itens. A Etapa 3 (calibração do banco de itens) consiste em escolher um modelo de resposta ao item adequado, um método de calibração e construir e interpretar a escala. A Etapa 4 (elaboração do algoritmo) consiste em definir: critérios para a habilidade inicial, procedimento de seleção dos itens, método de estimação da habilidade, critério de parada e restrições (se existirem). A Etapa 5 (análise da precisão e da validade) consiste em fazer simulações e utilizar critérios para analisar a precisão e a validade de um ou mais testes planejados nas etapas anteriores. A Etapa 6 (implementação) consiste na implementação prática do teste, verificando os recursos e as condições necessárias. A Etapa 7 (aplicação) consiste em executar o teste e constituir um banco de dados com as respostas dos examinandos. A Etapa 8 (manutenção) consiste em realizar manutenções periódicas e fazer as modificações necessárias implementando-as na Etapa 6. As próximas Seções desse Capítulo apresentam os detalhes das 8 etapas que constituem a SITAI.

4.3.1. Etapa 1: Definição do Teste

A primeira etapa compreende a formulação e planejamento do teste, abrangendo os seguintes aspectos: verificar a existência prévia do teste e definir a dimensão, o traço latente e os objetivos do teste. A Figura 16 apresenta os detalhes da primeira etapa da sistemática.

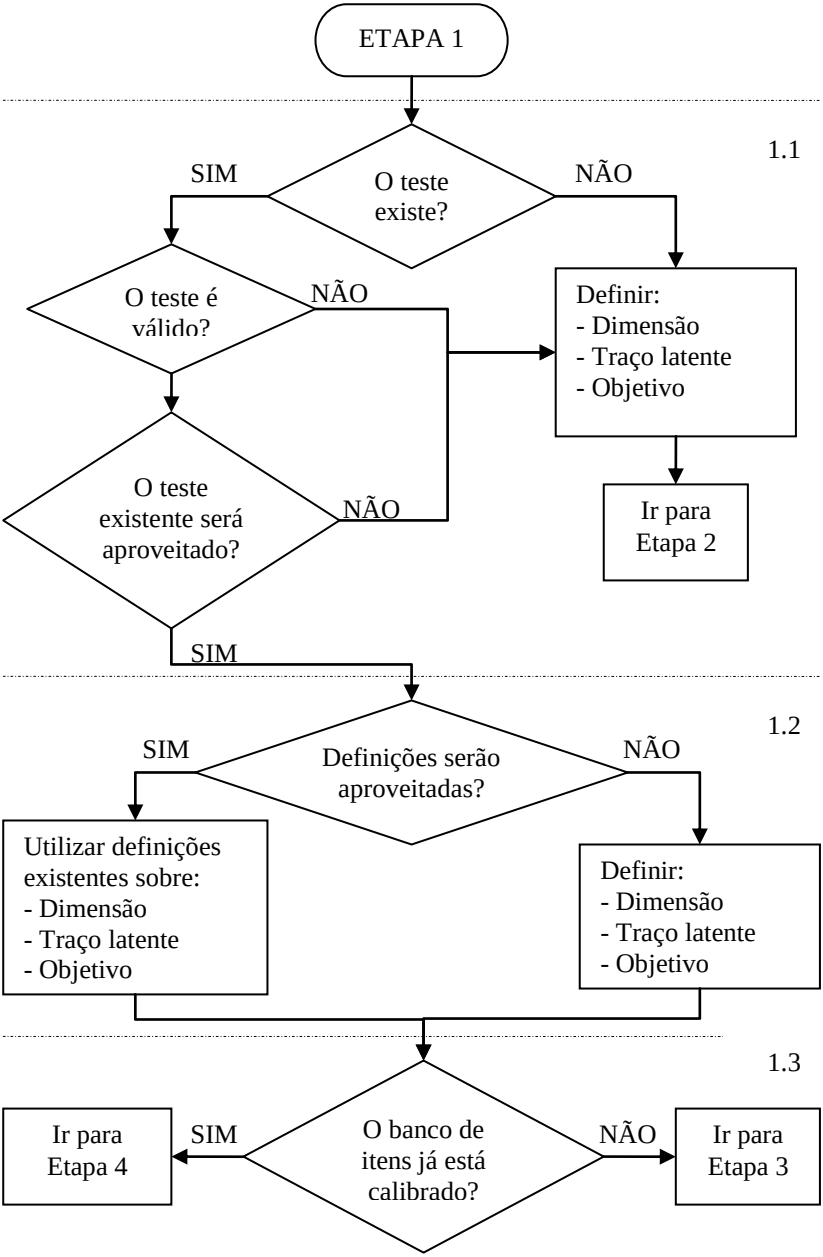


Figura 16. Etapa 1: Definição do Teste

A primeira questão a ser verificada, conforme a Subetapa 1.1, é se o teste a ser construído já existe ou não em uma versão não adaptativa. Se o teste não existe, será construído desde o princípio. O primeiro passo para a implantação de um teste novo é a definição de três aspectos: qual a dimensão do teste (quantidade de traços latentes avaliados), qual(is) o(s) traço(s) latente(s) que o teste procura medir (conhecimento, atitude, etc.) e qual o objetivo ou finalidade do teste (estimar uma nota ou classificar os indivíduos).

A quantidade de dimensões abordada pelo teste é fundamental para a sua elaboração, pois interfere diretamente na construção dos itens, no *design* do banco de itens (BI) e nas escolhas do modelo de resposta ao item e do método de seleção de itens. Se apenas uma dimensão for abordada (por exemplo, proficiência em matemática), então o teste será do tipo unidimensional. Se mais de uma dimensão for abordada (por exemplo, proficiência em ciências exatas, envolvendo as áreas de matemática, química e física), então o teste será multidimensional. Há situações em que o mesmo teste poderá ser tratado tanto da forma unidimensional quanto multidimensional. Por exemplo, um teste para medir a proficiência em matemática poderá ser tratado como unidimensional (proficiência em matemática) ou multidimensional, considerando as subáreas da matemática desse teste (álgebra, geometria e trigonometria, por exemplo). Nesses casos, o desenvolvedor do teste deverá decidir como irá tratar a questão da dimensão e poderá fazer simulações para comparar o desempenho das duas abordagens. Wise e Kingsbury (2000) sugerem a utilização de um modelo unidimensional, sempre que possível, por ser mais parcimonioso.

A definição do traço latente está diretamente relacionada com a sua dimensão, ou seja, se o teste for unidimensional, abordará apenas um traço latente, caso contrário, abordará mais de um. O traço latente é a característica a ser estudada, ou seja, é aquilo que o teste pretende medir: qualidade, satisfação, desempenho, proficiência, conhecimento, habilidade, atitude, de quem (indivíduos, objetos, serviços, sistemas, etc.) e referente a quê (uma área do conhecimento, uma característica psicológica, serviços, etc.). Essa definição ira influenciar no tipo de item a ser elaborado (dicotômico, politômico nominal ou gradual) na Etapa 2. Wise e Kingsbury (2000) salientam que o trabalho de identificar e medir traços latentes não é uma tarefa fácil.

O objetivo ou a finalidade do teste interfere diretamente no *design* do banco de itens e nas escolhas do método de seleção dos itens e do critério de parada do teste. Nesse sentido, deve-se verificar se o teste pretende estimar um valor para o traço latente (uma nota ou escore) na

escala que será criada ou se pretende classificar os respondentes em dois ou mais grupos. Em um teste novo (que não é uma versão adaptativa de um teste existente), após a definição desses três aspectos (dimensão, traço latente e objetivo), deve-se começar a elaborar os itens (Etapa 2).

Se o TAI a ser elaborado já existe em uma versão tradicional não adaptativa, deve-se verificar se o teste tem validade psicométrica. Caso não tenha, recomenda-se que ele não seja utilizado e que um novo teste seja elaborado. Se o teste for válido, deve-se verificar se será feita uma adaptação para um TAI a partir dos dados históricos existentes (banco de itens e de respostas dos examinandos) ou se será desconsiderado e construído desde o princípio, aproveitando ou não alguma parte do banco de itens existente. Se o teste existente não for aproveitado, o teste deverá ser reconstruído desde o princípio, assim como um teste novo, onde, após a definição da dimensão, do traço latente e do objetivo, deve-se elaborar os itens (Etapa 2). Usualmente, as versões adaptativas de testes tradicionais aproveitam os dados históricos (banco de itens e de respostas), o que permite um ganho substancial na elaboração de um TAI em relação à economia de recursos financeiros e de tempo com a elaboração dos itens e a coleta de dados.

Se o teste existente for aproveitado, deve-se verificar se a definição dos três aspectos (dimensão, traço latente e objetivo) também existe e será aproveitada (Subetapa 1.2). Caso negativo, esses aspectos deverão ser definidos, conforme discutido anteriormente antes de prosseguir. Após a decisão em utilizar ou definir os três aspectos, o próximo passo (Subetapa 1.3) será verificar se o banco de itens já foi calibrado, ou seja, se o teste na sua versão tradicional já utilizava a TRI. Se o teste já utilizava a TRI, então os itens já estarão calibrados e a próxima etapa será a construção do algoritmo do TAI (Etapa 4). Se o teste não utilizava a TRI, ou seja, a pontuação era baseada nos escores da TCT, então os itens deverão ser calibrados (Etapa 3). Deve-se atentar que a utilização de um banco de itens pré-existente, não projetado para um TAI, dependerá da sua qualidade (verificada na Etapa 3).

Caso o desenvolvedor do TAI não seja o mesmo do teste já existente, Renom e Doval (1999) recomendam que deve-se verificar os aspectos legais envolvidos (ex. direitos autorais e autorização para utilizar dados históricos e desenvolver a versão adaptativa).

4.3.2. Etapa 2: Elaboração dos Itens

A segunda etapa (Figura 17) compreende todos os processos envolvidos na elaboração dos itens, tais como, a definição dos conteúdos, dos tipos de item e da quantidade de itens.

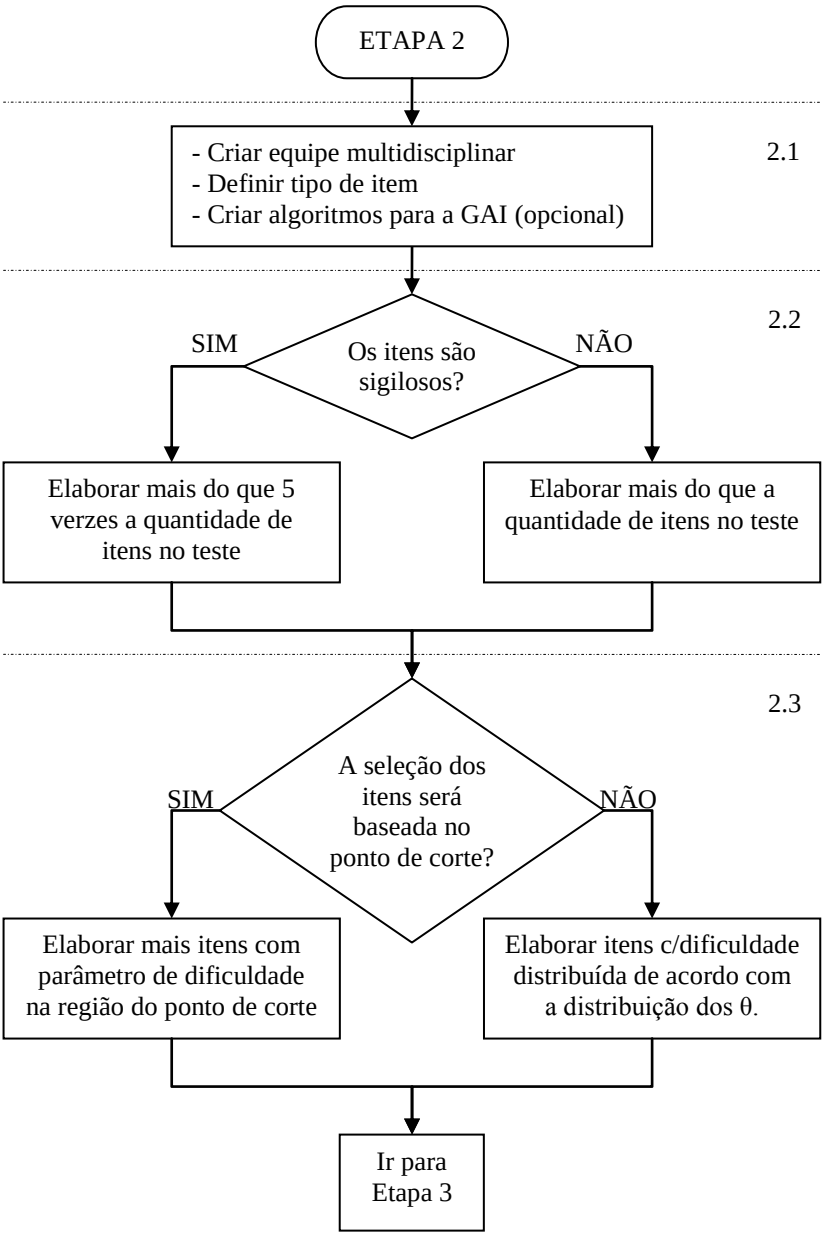


Figura 17. Etapa 2: Elaboração dos Itens

Essa etapa pode ser considerada uma das mais trabalhosas do processo, pois envolve todas as mesmas fases que compõem a elaboração dos itens utilizadas na TCT e observadas por Pasquali (1998). Como mencionado na Seção 1.4, esse processo não será avaliado nesse trabalho, entretanto é importante ressaltar que é necessária a constituição de uma equipe multidisciplinar composta por profissionais que sejam especialistas em relação ao traço latente estudado, além de estatísticos e psicometristas, para que sejam elaborados itens adequados (Subetapa 2.1). Várias literaturas fornecem subsídios para a elaboração dos itens, dentre as quais, pode-se mencionar Anastasi (1977), Fayers e Machin (2007), Garret (1979), Gullinksen, (1950), Miguel (1974), Osterlind (1997), Pasquali (1996; 1998), Prieto e Delgado (1996) e Wilson (2005). Dentro do contexto dos TAIs, alguns cuidados na elaboração dos itens devem ser considerados e são abordados nas Seções 3.3.2.1 e 3.3.2.2.

Na Subetapa 2.1 também será definido o tipo (ou formato) de item discutido na Seção 3.3.2.1, que influenciará no Modelo da TRI a ser escolhido na calibração (Etapa 3). O formato do item depende do objetivo do teste, por exemplo, para medir a proficiência, recomenda-se utilizar um formato de resposta de escolha múltipla, para medir satisfação, recomenda-se um formato de categorias ordenadas. Vale salientar que também poderão ser utilizados itens com recursos multimídias (sons, figuras em movimentos, cores piscantes, uso do teclado e do mouse, etc.), já que o teste será realizado via computador. Outro recurso disponível que poderá ser utilizado nessa etapa é a geração automática de itens – GAI (BEJAR et al., 2003; TEJADA, 2001), discutido nas Seções 2.3, 3.3.2.1 e 3.3.2.2, embora não possa ser utilizado em qualquer tipo de teste. A GAI poderá ser utilizada em testes onde os itens e suas alternativas possam ser elaboradas segundo um conjunto de regras que seguem uma determinada lógica (BEJAR, 1993; HORNKE; HABON, 1986; IRVINE; KYLLONEN, 2002), como, por exemplo, em testes de proficiência ou avaliações psicológicas de raciocínio.

Num TAI, a quantidade de itens a ser elaborado (Subetapa 2.2) é uma questão relevante. Se os itens forem sigilosos (por exemplo, teste de proficiência), será adotada uma taxa para controlar a exposição deles e uma grande quantidade de itens deverá ser elaborada. Como discutido na Seção 3.3.2.1, não existe entre os autores um consenso na quantidade de itens que devem compor o banco, mas recomenda-se que seja no mínimo 5 vezes mais do que a quantidade de itens que o teste terá. Além disso, a quantidade de itens no banco deve levar em consideração outros

fatores, como a quantidade de restrições no algoritmo, a taxa de exposição dos itens e a frequência de aplicações do TAI (WISE; KINGSBURY, 2000). Além do mais, recomenda-se elaborar uma quantidade de itens maior do que a necessária para um TAI, já que alguns itens provavelmente serão eliminados durante o processo de calibração. Obviamente, quanto maior a quantidade de itens no banco, melhor será o teste, pois haverá itens mais adequados para um determinado nível de habilidade, além de contribuir para o sigilo dos itens, já que diminui a taxa de exposição dos itens. Por outro lado, se os itens não forem sigilosos (por exemplo, avaliação do nível de satisfação e avaliações psicológicas), não há a exigência de um tamanho mínimo para a quantidade de itens no banco e não há a necessidade de elaborar uma quantidade muito grande de itens. Nesse caso, a quantidade de itens pode ser um pouco maior do que a quantidade que será utilizada no teste e, consequentemente, o algoritmo do teste será bem menos complexo.

Outro aspecto relacionado com a quantidade de itens é a dificuldade dos itens. Embora o parâmetro de dificuldade seja estimado na etapa da calibração, os especialistas podem desenvolver itens com diferentes graus de dificuldades provisoriamente estipulados. A distribuição da dificuldade dos itens (*design* do banco de itens) pode influenciar na escolha do método de seleção dos itens, a ser definido na Etapa 4. Se o método de seleção dos itens for baseado na estimativa provisória da proficiência (Subetapa 2.3), sugere-se que a distribuição de dificuldade dos itens seja semelhante a distribuição das proficiências na população (BERGSTROM; LUNZ, 1999), que usualmente é considerada uma Distribuição Normal. Porém, se o teste tiver o objetivo de classificar indivíduos e o método de seleção for baseado no ponto de corte, sugere-se que boa parte dos itens tenha o parâmetro de dificuldade situado na região do ponto de corte (THOMPSON; PROMETRIC, 2007). Entretanto, essa recomendação pode nem sempre ser adequada na prática, pois podem existir dificuldades em elaborar certa quantidade de itens com determinado grau de dificuldade.

4.3.3. Etapa 3: Calibração do Banco de Itens e Construção da Escala

A terceira etapa compreende o processo de calibração do banco de itens (BI) por meio da Teoria da Resposta ao Item (TRI), com a aplicação dos itens, a coleta dos dados, e a escolha do modelo de resposta e do método de calibração, além da construção e interpretação da escala. A Figura 18 apresenta os detalhes da terceira etapa da sistemática.

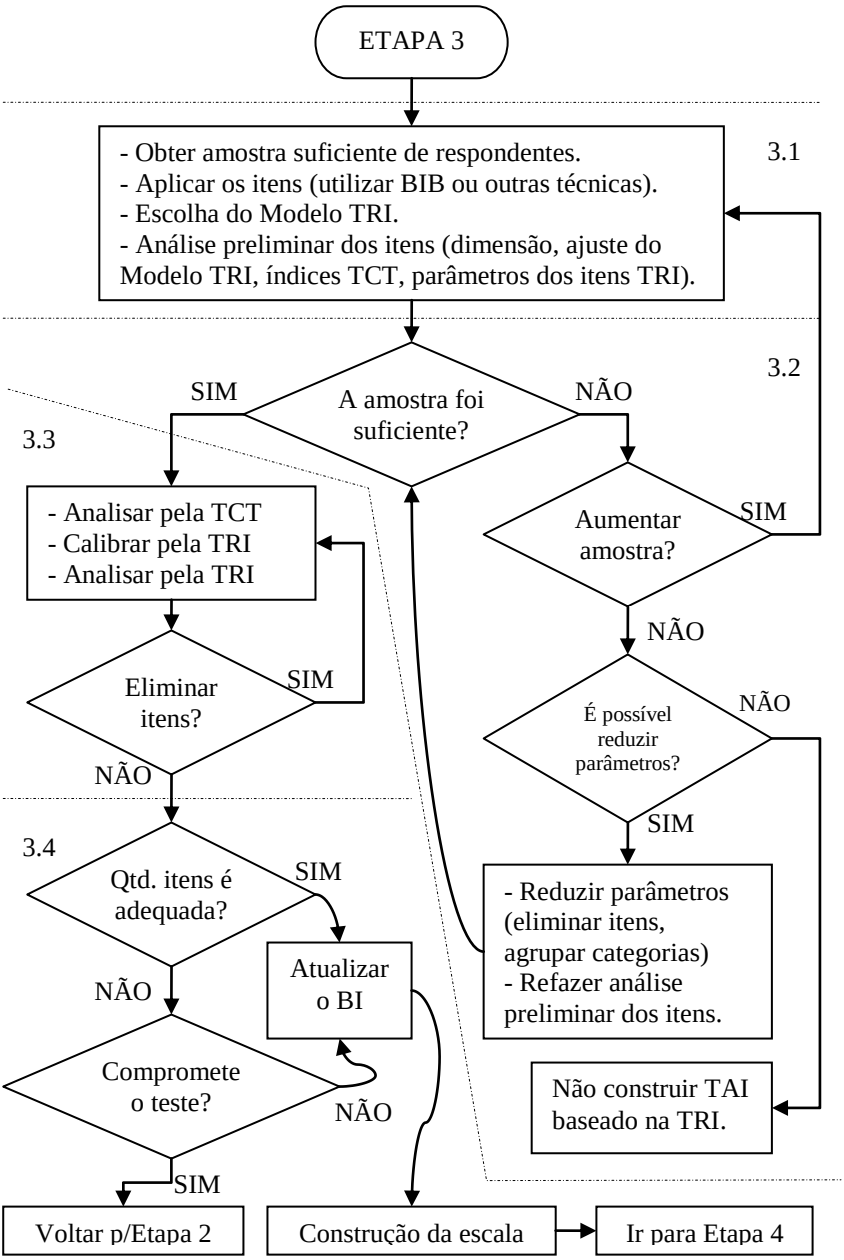


Figura 18. Etapa 3: Calibração do Banco de Itens

Após a elaboração dos itens, a próxima etapa é o processo de calibração deles. Para isso, é necessário ter uma amostra suficiente de respondentes (Subetapa 3.1). Como discutido na Seção 2.3, o tamanho da amostra necessário para calibração depende da quantidade de itens do banco (quanto mais itens, maior deve ser a amostra), da quantidade de parâmetros do modelo da TRI a ser utilizado (quanto maior a quantidade de categorias do item, maior deve ser a amostra) e do padrão de respostas da própria amostra (é necessário que todas as categorias de respostas tenham uma quantidade de resposta suficiente para a estimação dos parâmetros dos itens). Nas literaturas sobre TRI apresentadas no Capítulo 2, pode-se encontrar informações e estudos sobre o tamanho da amostra necessário para a estimação dos parâmetros do modelo. Nessa etapa, para a obtenção da amostra, o teste provavelmente será realizado no formato tradicional "papel e lápis" (numa versão informatizada, caso tenha itens multimídia). Geralmente, o BI é muito grande para que todos os indivíduos respondam todos os itens. Uma forma de resolver isso é utilizar técnicas como a do bloco incompleto balanceado (BIB) (ANDRADE; TAVARES; VALLE, 2000; EGGEN; VERHELST, 2011) e métodos de equalização (KOLEN; BRENNAM, 1995; NAVAS, 1996), conforme discutido na Seção 2.3.

Na prática, nem sempre é possível obter uma quantidade suficiente de amostra para calibrar os itens. Após coletada a amostra, deve-se fazer uma análise preliminar dos itens por meio dos critérios da TCT e da TRI (Subetapa 3.1), conforme abordado na Seção 3.3.2.3. Nesse processo, será escolhido e utilizado um modelo matemático adequado (vide Seção 3.3.1), dentre os vários modelos utilizados na TRI (ANDRADE; TAVARES; VALLE, 2000; DE AYALA, 2008; OSTINI; NERING; 2006; REVUELTA; PONSODA; ABAD, 2006; VAN DER LINDEN; HAMBLETON, 1997), segundo o tipo de traço latente (unidimensional ou multidimensional) e o tipo de item (dicotômico, politômico, escala gradual, etc.). Também é possível utilizar mais de um modelo no mesmo teste, se houverem diferentes tipos de itens no teste. Quanto ao traço latente, se o mesmo for unidimensional, utiliza-se um modelo unidimensional, mas caso seja multidimensional, pode-se optar por utilizar um modelo multidimensional ou unidimensional através da elaboração de mini-tais (SCHNIPKE; GREEN, 1995), vistos na Seção 3.3.1 ou de *testlets* (WAINER; KIELY, 1987), vistos na Seção 3.4.2.

Na análise preliminar da Subetapa 3.1 deve-se verificar se a dimensionalidade dos dados, o ajuste do modelo da TRI, os índices da TCT e os parâmetros dos itens da TRI. Quanto à dimensionalidade (vide Seção 2.1), deve-se verificar se é a mesma que foi determinada na

definição do(s) traço(s) latente(s) na Etapa 1 da SITAI. Se houver divergência na dimensionalidade do teste, pode-se: (1) adotar um modelo de resposta que considere essa dimensionalidade, (2) eliminar os itens que contribuem para a existência de dimensões indesejáveis ou (3) elaborar mais itens que contemplem a(s) dimensão(ões) desejada(s) e coletar mais amostras para calibrar esses novos itens. Quanto ao Modelo da TRI (vide Seção 2.1), deve-se verificar se o ajuste foi adequado e, se preciso, alterar o Modelo ajustado, pois um modelo mal-ajustado não fornecerá parâmetros invariantes para os itens e para as habilidades. Ainda na análise preliminar dessa Subetapa, se as estimativas dos parâmetros dos itens da TRI estiverem inconsistentes, apresentando valores absurdos ou erro padrão elevado, talvez seja devido ao tamanho inadequado da amostra. Se a amostra obtida não foi suficiente (Subetapa 3.2), deve-se aumentar o tamanho da amostra e repetir o processo da Subetapa 3.1. Se a amostra não puder ser aumentada (Subetapa 3.2), algumas modificações podem ser necessárias no sentido de reduzir a quantidade de parâmetros, por exemplo, eliminando os itens com desempenho mais fraco, reduzindo a quantidade de categorias dos itens ou dicotomizando itens politômicos. Após essas modificações, deve-se refazer a análise preliminar e, se a amostra for suficiente, conforme a consistência dos resultados, prossegue-se para a Subetapa 3.3. Se, após todos os esforços necessários para aproveitar essa amostra, o resultado não fornecer uma calibração adequada, a TRI não poderá ser utilizada na calibração e um TAI baseado na TRI não poderá ser elaborado. Nesse caso, uma alternativa menos sofisticada e que não possui os benefícios de um TAI baseado na TRI é utilizar um TAI baseado na TCT, como eram os primeiros testes adaptativos, assim como fizeram Dalpiaz (2007, p. 42) e Damando (2003, p. 76).

Se a amostra obtida for suficiente (Subetapa 3.2), prossegue-se para a Subetapa 3.3, onde deve-se fazer uma análise dos itens por meio dos critérios da TCT e a calibração e análise dos itens por meio da TRI. As análises por meio dos índices da TCT (discriminação, dificuldade, correlação bisserial) ajudarão na eliminação dos itens inadequados (vide Seção 3.3.2.3). O processo de calibração dos itens por meio da TRI, discutido nas Seções 2.3, 2.4 e 3.3.2.3 e a análise dos parâmetros estimados (vide Capítulo 2 e Seção 3.3.2.3) são feitos de forma consecutiva, ou seja, os itens são calibrados e, em seguida, analisados. Essa análise ajudará na decisão da eliminação de itens inadequados. Após a eliminação desses itens, as análises por meio dos indicadores da TCT e da dimensionalidade devem ser refeitas para verificar se não foram afetadas pela eliminação dos itens, e o processo de calibração

deve ser refeito para verificar se os itens restantes estão adequados e não foram afetados pela eliminação dos itens. Geralmente, os itens são calibrados na escala (0,1), com média zero e desvio padrão 1, o que pode ser mudado posteriormente por meio de uma relação linear.

Após a calibração final dos itens (Subetapa 3.4), deve-se verificar se a quantidade de itens que permaneceram no BI é suficiente para a aplicação do TAI, se abrangem todos os conteúdos do teste e se estão distribuídos e fornecem informação adequada em toda extensão do traço latente avaliado (itens fáceis, medianos e difíceis), conforme o objetivo do teste e o critério de seleção dos itens. Caso seja necessária a elaboração de novos itens, deve-se verificar se é possível começar a aplicação do TAI com o BI atual sem que haja um comprometimento significativo do teste e, assim, esses itens podem ser posteriormente adicionados e calibrados gradativamente conforme a manutenção na Etapa 8. Se a quantidade de itens que permaneceram compromete a validade do teste, é necessário elaborar mais itens (voltar para a Etapa 2) e realizar a pré-testagem e a calibração desses (pode-se utilizar métodos de equalização, conforme Seção 2.3) e adicioná-los aos itens já existentes e calibrados. Ainda que a quantidade teoricamente necessária de itens para o TAI não tenha sido alcançada, estudos de simulação considerando o algoritmo e as restrições do TAI podem ser feitos para verificar a eficiência e a validade do BI e possivelmente adequá-lo para o teste, por exemplo, diminuindo a quantidade de restrições e de categorias e a complexidade do algoritmo (SEGALL, 2005). Se a quantidade de itens que permaneceram não compromete a validade do teste, as informações psicométricas obtidas no processo de calibração (parâmetros dos itens) devem ser incorporadas ao (BI), assim como qualquer outra informação que pode ser considerada importante, por exemplo, os índices obtidos pela TCT. Os itens eliminados nessa etapa devem ser excluídos do BI. Finalizado o processo de calibração e atualizado o BI, procede-se para a construção e interpretação da escala, conforme posicionamento dos itens âncoras, que foi discutido na Seção 2.6. Finalizada essa parte, procede-se para a Etapa 4.

4.3.4. Etapa 4: Elaboração do Algoritmo

A quarta etapa compreende todos os processos envolvidos na elaboração do algoritmo adaptativo do teste: definição da habilidade inicial, do método de seleção dos itens e do critério de parada e a imposição das restrições existentes (controle do conteúdo e da exposição dos itens). A Figura 19 apresenta os detalhes da quarta etapa da sistemática.

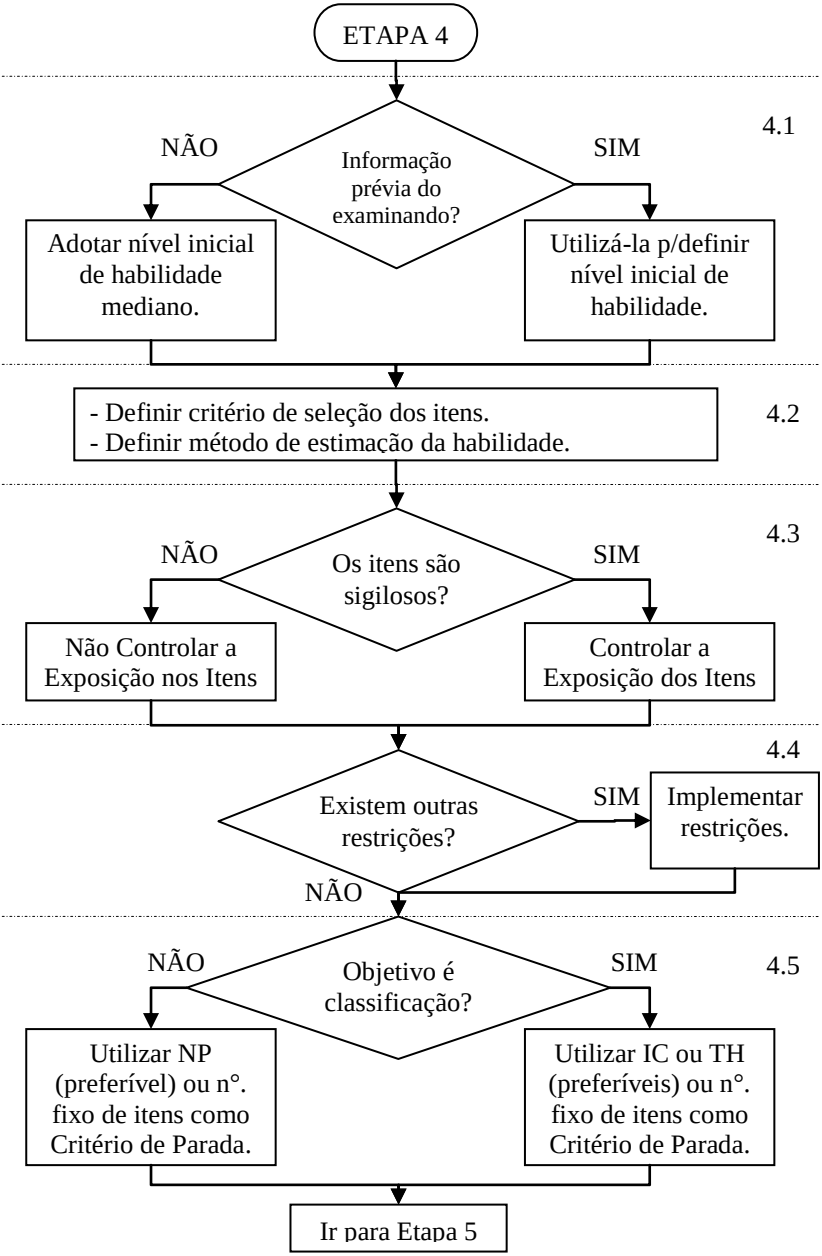


Figura 19. Etapa 4: Elaboração do Algoritmo

Após a finalização da calibração dos itens (Etapa 3), deve-se construir o algoritmo do teste, que envolve a definição da habilidade inicial dos examinandos, os critérios de seleção dos itens no início e durante o teste, o método de estimação da habilidade no início e durante o teste e a definição do critério de parada do teste, conforme mostra a Figura 9 da Seção 3.3.

Na Subetapa 4.1, deve-se verificar se existe alguma informação prévia do examinando. Para que o teste possa selecionar o primeiro item a ser apresentado ao examinando, é necessária a existência de um critério de seleção adequado para essa situação, como discutido na Seção 3.3.3. Em geral, verifica-se a existência de alguma informação prévia dos indivíduos que possa ser utilizada para estimar o nível inicial da sua habilidade (WISE; KINGSBURY, 2000; YANG; POGGIO; GLASNAPP, 2006). Por exemplo, pode-se utilizar a habilidade estimada em aplicações anteriores do teste, resultados de outros tipos de testes, ou outros tipos de variáveis que podem estar relacionadas com a característica medida (nível escolar, classe econômica, etc.). Na ausência dessas informações, geralmente adota-se um nível de habilidade inicial mediano, que pode ser fixo, centrado na média, ou um valor aleatório dentro de um intervalo mediano. No caso do valor fixo, o primeiro item será sempre o mesmo para todos os examinandos, salvo se houver controle da taxa de exposição. A seleção de um valor aleatório mediano contribui para o controle da exposição dos itens, mas poderá diminuir um pouco a eficiência do teste.

Na Subetapa 4.2, serão definidos o critério de seleção dos itens e o método de estimação da habilidade. O critério de seleção dos itens iniciais pode ser diferente do critério de seleção dos demais itens, por exemplo, podem-se utilizar itens menos informativos no início do teste já que a habilidade estimada no início ainda é pouco precisa (VAN DER LINDEN; GLAS, 2010). Outros critérios seleção dos itens iniciais podem ajudar no controle da taxa de exposição dos itens (WEISS; GUYER, 2010). Em testes com objetivo de classificação, recomenda-se que os primeiros itens selecionados sejam localizados próximos do ponto de corte (BERGSTROM; LUNZ, 1999). O nível de habilidade inicial, os critérios de seleção do(s) primeiro(s) item(ns) e o método de estimação da habilidade no início do teste são discutidos na Seção 3.3.3.

Diversos critérios de seleção dos itens podem ser utilizados nos TAIs (vide Seção 3.3.4), onde os mais comuns são baseadas no procedimento da MV que selecionam itens procurando maximizar a informação na estimativa atual da habilidade e nos procedimentos

Bayesianos que selecionam itens procurando minimizar a variância da posteriori (SEGALL, 2005). Os critérios apresentados na Seção 3.3.4 são destinados preferencialmente ao Modelo Logístico Unidimensional. Para a determinação do critério de seleção dos itens, deve-se levar em conta vários fatores, tais como, a dimensionalidade, o modelo de resposta, o *design* do banco de itens, as restrições do algoritmo, o método de estimação da habilidade, o objetivo do teste e o critério de parada (vide Seção 3.3.4.4). Para testes com os Modelos Politômicos e Multidimensionais, sugere-se utilizar a literatura específica mencionada na Seção 3.3.1.

Juntamente com o critério de seleção dos itens, deve-se definir um método a ser utilizado na estimação da habilidade durante o teste. O método de estimação da habilidade no início do teste pode ser diferente do método de estimação da habilidade durante o teste. Por exemplo, no caso da estimação por MV, que não possui um máximo definido quando existe um padrão de resposta (todas respostas corretas ou todas incorretas) podem-se utilizar outros métodos até que esse padrão de resposta seja "quebrado". Esses métodos são discutidos nas Seções 2.5 e 3.3.5.

Na Subetapa 4.3, deve-se verificar se existe a necessidade de controlar a taxa de exposição dos itens. Se os itens do teste são sigilosos (por exemplo, testes de proficiência), existe a necessidade de controlar a exposição dos itens para que eles não se tornem conhecidos e comprometa a confiabilidade do teste (Subetapa 4.3). O controle da exposição dos itens, assim como a utilização do balanceamento de conteúdo ou qualquer outro tipo de restrição, fará com que o teste não alcance o seu melhor desempenho. Inevitavelmente, em algum momento do teste, o item escolhido não será o melhor (o de maior informação para o determinado nível de habilidade) devido às restrições necessárias definidas no algoritmo do teste. Disso resultará uma perda na precisão do teste, uma vez que a precisão está relacionada com a informação do item, ou seja, quanto maior a informação que o item fornece, maior a precisão na estimativa da habilidade. Para que essa perda seja praticamente imperceptível, é necessário que o banco de itens possua um boa quantidade de itens de qualidade, isto é, itens com uma boa quantidade de informação, distribuídos ao longo da escala. Além disso, quanto menor for o valor fixado para a taxa de exposição do item, maior deverá ser a quantidade de itens no banco. Quanto ao valor para a taxa de exposição de itens, Way (1998) sugere uma exposição máxima de 15%, enquanto que Olea et al. (2004) utilizaram 25% e desenvolvedores

de testes com alto risco utilizam 10% (SEGALL, 2005). Métodos para o controle da taxa de exposição dos itens são discutidos na Seção 3.4.1.

Se existirem outras restrições (por exemplo, balanceamento de conteúdo, itens que não podem ser aplicados num mesmo teste, quantidade mínima e/ou máxima de itens num teste), elas deverão ser implementadas (Subetapa 4.4). Como mencionado anteriormente, qualquer tipo de restrição, fará com que o teste não alcance o seu melhor desempenho, pois, em algum momento do teste, o item escolhido inevitavelmente não será o melhor. Essas restrições são discutidas na Seção 3.4.

Na Subetapa 4.5 deve-se definir o critério de parada, o qual depende do objetivo do teste. Se o objetivo for estimar um valor para o traço latente, recomenda-se que a habilidade estimada deverá alcançar um nível mínimo de precisão (WEISS; KINGSBURY, 1984), ou seja, um determinado erro padrão (EP) mínimo. Se o objetivo for classificar ou dividir os respondentes em dois ou mais grupos (aprovado/reprovado, classificado/desclassificado, satisfeito/insatisfeito, alto/médio/baixo, etc.), recomenda-se comparar a habilidade estimada com um ou mais pontos de corte pré-determinados (EGGEN; STRAETMANS, 2000; SPRAY; RECKASE, 1996; XIAO, 1999), utilizando-se, para isso, o intervalo de confiança (IC) para a habilidade ou o teste de hipótese (TH). Por exemplo, se o intervalo estiver acima do ponto de corte, o indivíduo é classificado acima, caso contrário, é classificado abaixo. Embora essas condições de critério de parada não sejam uma regra obrigatória, o TAI produz melhores resultados, em termos de precisão, se elas forem utilizadas. Outro critério muito utilizado, independente do objetivo do teste é estabelecer uma quantidade fixa de itens no teste. Esse critério, entretanto, não é muito recomendado, pois, se essa quantidade de itens for grande, em muitas situações, poucos itens podem ser o suficiente para estimar o traço latente com precisão ou classificar o examinando acima ou abaixo de um ponto de corte. Por outro lado, se essa quantidade for pequena, pode não ser suficiente para estimar o traço latente com precisão ou classificar corretamente um examinando. Esses e outros critérios de parada são discutidos na Seção 3.3.6.

4.3.5. Etapa 5: Análise da Precisão e da Validade

A quinta etapa compreende a utilização de simulações para a análise da precisão e da validade do teste por meio de critérios estatísticos adequados. A Figura 20 apresenta os detalhes da quinta etapa da sistemática.

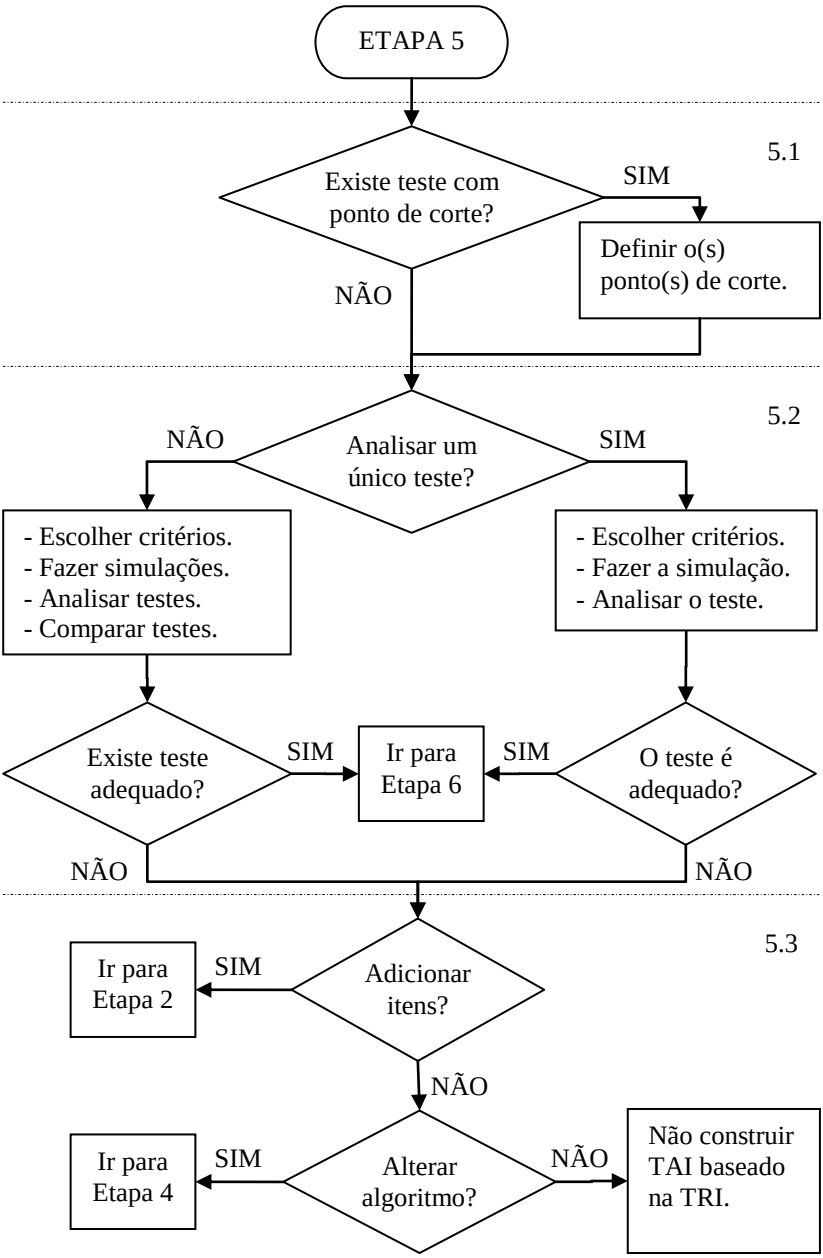


Figura 20. Etapa 5: Análise da Precisão e da Validade

Na Subetapa 5.1, verifica-se se, entre os testes a serem testados, existe algum que possui ponto de corte, ou seja, se existe algum teste com objetivo de classificação. Caso exista, deve-se definir os pontos de cortes necessários. Se o teste construído foi uma adaptação de um teste tradicional, é comum definir o ponto de corte de tal forma que se mantenha a mesma taxa de aprovação ou classificação do teste tradicional. Esse e outros procedimentos são discutidos brevemente na Seção 3.3.6.2.

Uma vez definido o algoritmo do TAI, é necessário avaliá-lo mediante algum controle psicométrico de qualidade para verificar a precisão e a validade do teste, através de dados empíricos ou simulações. Como visto na Etapa 4, vários algoritmos para um mesmo TAI podem ser projetados, combinando diferentes bancos de itens, critérios de seleção de itens, métodos de estimação do traço latente, critérios de parada e restrições. Por exemplo, pode-se formar um banco de itens com itens com parâmetros aceitáveis e outro banco mais rígido selecionando apenas itens com desempenho acima do aceitável. Nos casos onde a quantidade de itens é baixa, pode-se montar bancos com diferentes tamanhos e testá-los por meio de simulações para verificar até que ponto os itens menos adequados interferem na qualidade do teste. Na Subetapa 5.2, os algoritmos projetados serão testados e avaliados. Nessa subetapa, pode ser analisado o desempenho de um único teste ou comparado o desempenho de várias opções de testes definido nas etapas anteriores. Na verificação do desempenho de um único teste, deve-se escolher os critérios de avaliação, realizar a simulação e analisar o teste. Se as análises mostrarem que o teste é adequado, ele está pronto para ser implementado (Etapa 6) e utilizado (Etapa 7), caso contrário, o teste deverá ser reformulado, conforme Subetapa 5.3. Na verificação do desempenho de um único teste, deve-se escolher os critérios de avaliação, realizar as simulações, analisar os testes e comparar os desempenhos dos testes. Após a comparação, deve-se verificar quantos testes tiveram desempenho adequado e se existe algum deles com desempenho significativamente melhor do que os demais. Se as análises mostrarem que existe um teste adequado com desempenho superior aos demais, ele estará pronto para ser implementado (Etapa 6) e utilizado (Etapa 7), caso contrário, o teste deverá ser reformulado, conforme Subetapa 5.3.

Muñiz e Hambleton (1999) comentam sobre os seguintes critérios para a avaliação de um TAI: a) em relação à precisão, tem-se o erro padrão médio de estimação (EPM), a raiz quadrada do erro quadrado

médio (RQEQM), o desvio empírico médio (DEM), a eficiência (EF), as correlações entre a habilidade simulada e a habilidade estimada (CL), e outros procedimentos provenientes da TCT, e b) em relação à validade (de conteúdo, de construto e de predição), os quais são discutidos na seção 3.5. Se o objetivo do teste for estimação do traço latente utilizando um nível de precisão como critério de parada, recomenda-se utilizar os critérios RQEQM, DEM, EF e CL. Se o objetivo do teste for estimação do traço latente utilizando uma quantidade fixa de itens como critério de parada, recomenda-se utilizar os critérios EPM, RQEQM, DEM e CL. Se o objetivo do teste for classificação, é mais apropriado verificar as proporções de acerto e erro na classificação dos examinandos do que utilizar os critérios de precisão mencionados por Muñiz e Hambleton (1999). A verificação desses critérios usualmente é feita por meio de simulação (simula-se ou define-se o valor do traço latente para vários examinandos e simulam-se as respostas para cada item selecionado pelo algoritmo). Pode-se programar um software para realizar as simulações ou utilizar algum software disponível para tal. Dentre os softwares listados no Capítulo 3, alguns fazem simulações em TAI.

Os resultados são analisados verificando o quanto o algoritmo testado conseguiu recuperar da habilidade simulada, ou seja, se as habilidades estimadas são as mesmas ou próximas das habilidades simuladas. Os resultados são analisados segundo os critérios que foram selecionados anteriormente.

As comparações entre os algoritmos são feitas buscando verificar qual dos algoritmos obteve o melhor desempenho e são baseadas nos critérios que foram selecionados anteriormente.

A Subetapa 5.3 é aplicada na remota hipótese de nenhum teste ter desempenho adequado para um TAI. Se for necessário e possível, deve-se voltar à Etapa 2 e elaborar e incorporar mais itens e realizar a calibração desses (pode-se utilizar métodos de equalização, conforme Seção 2.3). Se for possível alterar o algoritmo (mudar critérios, alterar restrições, etc.), deve-se voltar à Etapa 4. Se não for possível adicionar mais itens nem alterar o algoritmo, não será possível construir o TAI.

4.3.6. Etapa 6: Implementação

A sexta etapa compreende a implementação prática do teste, onde deve-se considerar vários recursos materiais além de uma equipe multidisciplinar. A Figura 21 apresenta os detalhes da sexta etapa da sistemática.

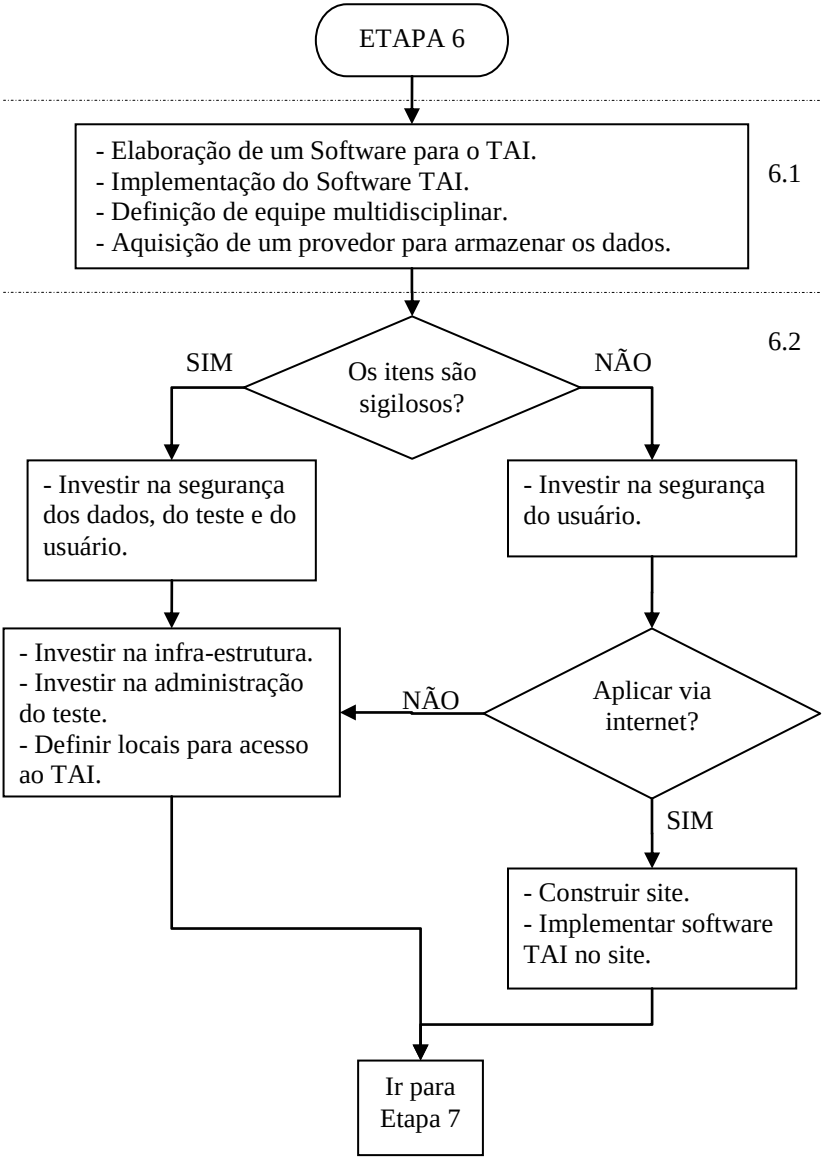


Figura 21. Etapa 6: Implementação

Se algum teste for considerado adequado, segundo as análises da Etapa 5, então ele está pronto para ser implementado. Para implementar o teste, diversos fatores deverão ser observados, entre eles: a linguagem de programação, a segurança do banco de itens (vide Seção 3.3.2.6) e de respostas (quando o TAI for aplicado), os recursos materiais, local dos terminais de acesso para a realização dos testes.

Na Subetapa 6.1, são mencionados os aspectos básicos para a implementação de um TAI. Para a aplicação do TAI, pode-se programar um software ou utilizar algum software disponível para tal. Alguns softwares disponíveis para a implementação de TAIs foram listados no Capítulo 3. No software, deve-se cuidar alguns aspectos como o *design*, a forma de exibição dos itens na tela, a utilização de recursos de mídia e outros aspectos observados por Hambleton, Zaal e Pieters (1991) e descritos no Capítulo 3.

Depois de elaborado o software, ele deverá ser implementado em computadores, servidores ou na internet, conforme o propósito do teste. Deve-se cuidar aspectos, tais como, compatibilidade entre hardware e software, requisitos do sistema operacional, compatibilidade com navegadores de internet, configurações do computador (processador utilizado, memória necessária), etc.

Por envolver muitas áreas diferentes, o processo de implementação do teste também necessita de uma equipe multidisciplinar (profissionais de programação, técnicos de informática, de ergonomia e *design*, gestores). Mais detalhes sobre o processo de implementação podem ser vistos em Fernandes (2009) e Piton-Gonçalves (2004).

Também será necessária a aquisição de um provedor para armazenar os dados coletados. Se o teste for aplicado em larga escala, maior será a quantidade de informação a ser armazenada, e maior será, consequentemente, o investimento com provedores. Os dados coletados serão utilizados posteriormente na manutenção do TAI (Etapa 8).

A Subetapa 6.2 verifica se os itens são ou não sigilosos e como se deve proceder em ambos os casos. No caso da necessidade de sigilo tanto do banco de itens quanto do banco de respostas e dados dos examinandos (quando o TAI for aplicado), deve-se investir na segurança dos dados. Os dados (informações dos itens e respostas e dados dos examinados) deverão ser armazenados em um provedor central que deverá estar interligado com todos os terminais de acesso ao teste. Nesses casos, é comum o uso de criptografia e senhas, principalmente em testes que podem ser acessados por rede ou pela internet. Testes com itens sigilosos (por exemplo, teste de proficiência) deverão ser feitos em

ambientes onde os candidatos devem dispor de uma infra-estrutura adequada e não tenham contato com outras pessoas ou fontes de informação durante o teste, possivelmente acompanhados por fiscais e instrutores.

Testes com itens não sigilosos (por exemplo, avaliação de satisfação ou atitude) não precisam necessariamente de investimento na segurança dos itens, que podem ser conhecidos antecipadamente pelo examinando sem que isso interfira na avaliação do traço latente. Entretanto, o sistema deve garantir a segurança e o sigilo de dados dos usuários que venham a ser solicitados no teste, por exemplo, dados de identificação, tais como nome, número de documentos ou endereço residencial. Esses testes poderiam ser administrados via internet e acessados por qualquer indivíduo que esteja interessado em fazê-los. Nesse caso, deverá ser criado um site para a aplicação do teste e o software TAI deverá ser implementado nele. Um exemplo prático desse tipo é o *Assessment Center*^{SM} (CELLA; GERSHON, 2010), cujo site oferece diversos testes adaptativos relacionados com saúde emocional, física e social, onde qualquer usuário pode participar.

Nos testes que não são aplicados via internet, deve-se investir na infra-estrutura e na administração do teste, ou seja, garantir lugares adequados, computadores para todos os examinandos e alguém para gerenciar e fiscalizar os testes. Os computadores devem estar em bom estado de conservação e funcionamento (mouse, teclado, monitor, boa velocidade de conexão com o servidor que contém o banco de itens) e não devem permitir o acesso do candidato à internet durante o teste. Também deve-se definir os locais dos terminais de acesso para a realização dos testes e interligação desses terminais com o provedor. Se o teste é considerado de larga escala, ele deverá ser realizado várias vezes durante o ano e possivelmente em diversos lugares do país ou do mundo. Deverá haver terminais em várias localidades e muitas datas disponíveis para a realização do teste. Se o teste for destinado a uma população pequena, pode ser que não haja a necessidade de muitos terminais nem de várias datas disponíveis para a realização do teste.

4.3.7. Etapa 7: Aplicação

A sétima etapa compreende a aplicação efetiva do teste, fornecendo um retorno ao examinando sobre o seu desempenho, e com a coleta de dados para constituir um banco com as respostas dos examinandos para posterior a análise e manutenção do teste. A Figura 22 apresenta os detalhes da sétima etapa da sistemática.

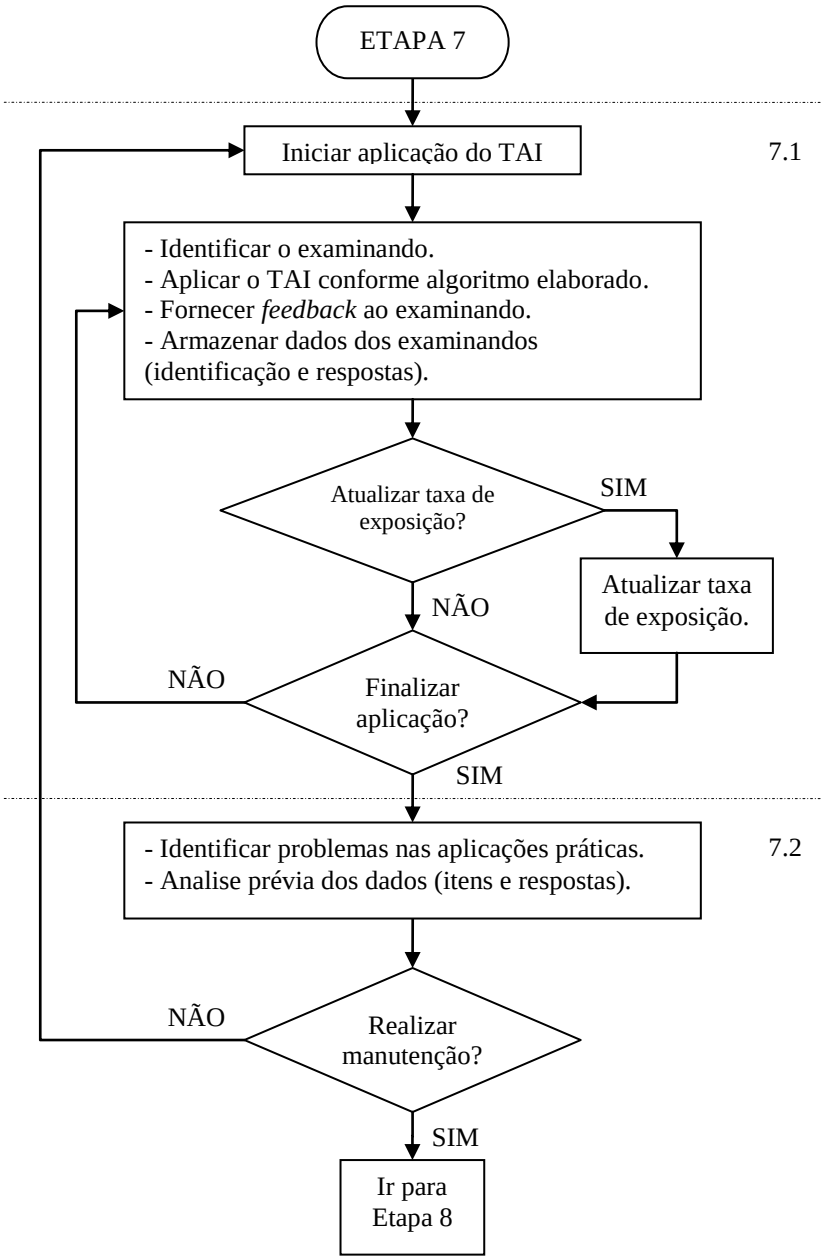


Figura 22. Etapa 7: Aplicação

Após a implementação do TAI, o teste está pronto para ser aplicado. A Subetapa 7.1 detalha o processo de aplicação do teste. Nessa subetapa, os examinandos vão realizar o teste simultaneamente ou não, dentro de um período pré-estabelecido de tempo. Durante esse período, para cada teste aplicado, serão cumpridas as seguintes etapas:

- Identificar o examinando: Essa identificação tem por objetivo cadastrar as informações de identificação do indivíduo para obter um controle, por exemplo, em relação às outras vezes que esse indivíduo já fez o teste, o que poderá influenciar na seleção de itens adequados (que não foram aplicados a ele anteriormente e que estejam de acordo com a habilidade estimada no seu último teste). Existem tipos de testes onde não há o interesse ou necessidade de identificar o examinando.
- Aplicar o TAI conforme algoritmo elaborado: Aplicação completa do teste desde a seleção do primeiro item até o último.
- Fornecer *feedback* ao examinando: Após a finalização do teste, fornecer ao examinando a sua nota ou classificação com um relatório resumido do seu desempenho, identificando os conteúdos que ele domina e os que ele não domina.
- Armazenar dados dos examinandos (identificação e respostas): Armazenar no servidor os dados dos examinados. A identificação poderá ser utilizada caso o examinando tenha que fazer o teste novamente. As respostas serão utilizadas posteriormente para a análise e manutenção do teste.

Após a aplicação de cada teste, deve-se verificar se é necessário (e se é possível) atualizar a taxa de exposição dos itens nos testes onde os itens são sigilosos. Se algum item está sendo muito aplicado, a sua taxa deve ser atualizada no algoritmo. Existem outras opções para o algoritmo controlar a exposição do item, tais como a seleção aleatória da habilidade inicial ou a seleção aleatória dos itens iniciais (WEISS; GUYER, 2010), como mencionado na Etapa 4. Essas alternativas podem contribuir para que a taxa de exposição dos itens não tenha que ser atualizada toda vez que um teste seja aplicado, já que, em algumas situações não é possível atualizar a taxa de exposição durante a aplicação do teste, porém somente na manutenção (Etapa 8).

Se o teste já foi aplicado uma quantidade suficiente para que se possa fazer uma análise prévia dos dados, deve-se encerrar temporariamente essa aplicação, para essa verificação. Caso a

quantidade de aplicações ainda não seja suficiente, os teste continuam sendo aplicados.

Na subetapa 7.2 deve-se verificar se durante a aplicação prática dos testes foi observado algum problema, tanto operacional quanto de infra-estrutura. Nessa subetapa, também é feita uma análise prévia dos dados. Os dados que foram coletados são utilizados para verificar:

- Se a taxa de exposição dos itens precisa ser atualizada;
- Se os parâmetros dos itens estimados com base nas respostas coletadas durante a aplicação dos TAIs não diferem significativamente dos parâmetros dos itens estimados na Etapa 3 ou das estimativas atuais dos parâmetros;
- Se existem itens que devem ser removidos;
- Se existem itens que devem ser acrescentados;
- Se existem itens que foram acrescentados recentemente e precisam ter seus parâmetros estimados;
- Se foram observados problemas nas respostas dos examinandos (identificar padrões inesperados de respostas, existência de DIF, etc.);
- Se o algoritmo parece não estar funcionando adequadamente e precisa ser modificado;
- Se houver sido diagnosticado qualquer outro problema que precisa ser resolvido.

Caso nenhuma manutenção seja necessária, retorna-se à Subetapa 7.1 e o teste pode continuar sendo aplicado. Se houver alguma modificação a ser realizada, avança-se para Etapa 8, que é a manutenção do teste.

4.3.8. Etapa 8: Manutenção

A oitava etapa compreende a manutenção do teste, que deverá ser periódica e, se necessário, poderá resultar em alterações no teste, evidenciadas por meio da análise do desempenho do TAI com os dados reais. A Figura 23 apresenta os detalhes da oitava etapa da sistemática.

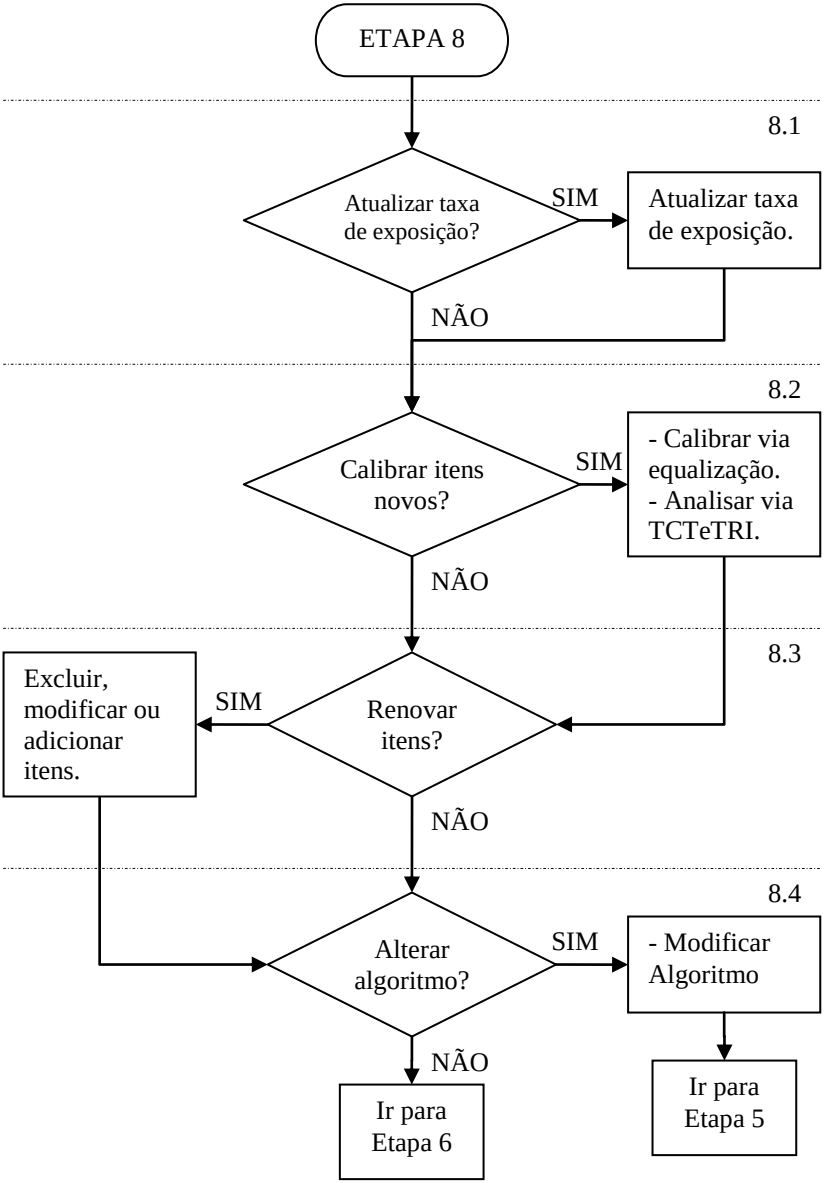


Figura 23. Etapa 8: Manutenção

Quando necessário e evidenciado na Subetapa 7.2, deverá se fazer uma manutenção periódica no banco de itens, discutida na Seção 3.3.2.5.

A Subetapa 8.1 verifica se há a necessidade de atualizar a taxa de exposição dos itens administrados. Se o teste não for projetado para fazer a atualização dessa taxa durante a aplicação dos testes, então essa taxa deverá ser atualizada nessa Subetapa.

A Subetapa 8.2 verifica se existe algum item novo, adicionado na última manutenção, que já tenha uma quantidade suficiente de respostas para ser calibrado mediante um método de equalização adequado, que utilize itens em comum, e colocados na escala existente. Esses novos itens devem ser analisados com base nos critérios da TCT e da TRI, já mencionados anteriormente.

A Subetapa 8.3 trata da renovação de itens no banco (exclusão ou adição). Os parâmetros dos itens devem ser reestimados com base nas respostas coletadas durante a aplicação dos TAIs e comparados com os parâmetros estimados no banco de itens para verificar se o item deve permanecer no banco ou ser excluído. Se as estimativas forem semelhantes, o item permanece no banco, mas se forem observadas alterações (por exemplo, uma queda significativa da estimativa do parâmetro *a*), o item deverá ser eliminado. As análises dos itens novos servirão para verificar se eles podem ou não ser adicionados no banco de itens. Nessa etapa, a idéia não é repetir toda a análise da Etapa 3, porém utilizar os critérios da TCT e da TRI discutidos naquela etapa, para tomar as decisões de exclusão, modificação ou adição de itens.

Itens deficientes (detectados pelos critérios da TCT e da TRI, ou pelas respostas inesperadas ou pela presença de DIF) ou obsoletos (que não atendem mais às especificações e objetivos da avaliação) deverão ser eliminados, ou modificados (nas suas alternativas ou parte do enunciado), ou substituídos por itens atualizados e equivalentes. Tejada (2001) recomenda que os itens existentes sejam periodicamente substituídos por itens novos, para evitar a superexposição de itens no caso de itens sigilosos. Uma alternativa para que os itens não sejam superexpostos é a utilização de bancos de itens rotativos (vide Seção 3.3.2.6), o que contribui com a segurança do banco de itens e ajuda no controle da exposição dos itens.

Novos itens podem ser adicionados gradualmente e calibrados na próxima manutenção com as respostas obtidas. Esses novos itens serão aplicados no TAI para coleta de dados para posterior calibração, não sendo utilizados para avaliar o examinando. Os novos itens deverão ser elaborados observando-se as considerações da Etapa 2, o que aumenta o

custo de implementação e operacionalização. No entanto, os benefícios de um TAI superam os custos envolvidos no seu desenvolvimento, na sua implementação e na sua manutenção (FETZER et al., 2008; HORNKE, 1999).

A Subetapa 8.4 verifica se o algoritmo do teste precisa ser modificado. A modificação do banco de itens (recalibração, eliminação e adição de itens) resultará em diferenças na estrutura do banco de itens. Por exemplo, se o banco havia sido projetado para ter mais itens na região do ponto de corte ou itens distribuídos em toda extensão do traço latente considerado, de acordo com a distribuição de θ na população, deve-se verificar se houve mudanças. Tais modificações podem afetar o desempenho do algoritmo do TAI, sendo necessário modificá-lo. O desempenho pode ser verificado por meio de simulações, conforme os indicadores mencionados na Etapa 5. Caso o algoritmo precise ser modificado, novas simulações deverão ser feitas e os dados analisados segundo os critérios da Etapa 5. Após a validação do teste, as alterações feitas deverão ser atualizadas no software na Etapa 6, juntamente com qualquer outra manutenção que tenha ido feita.

Qualquer outro problema que tenha sido diagnosticado, tanto operacional quanto de infra-estrutura, deverá ser resolvido nessa manutenção. As alterações feitas durante a manutenção deverão ser consideradas novamente na implementação do TAI (Etapa 6) somente referente ao que foi modificado.

5. ESTUDO DE CASO

O DETRAN-SC (Departamento Estadual de Trânsito de Santa Catarina) é uma empresa estatal vinculada ao DENATRAN (Departamento Nacional de Trânsito), sendo um dos seus atributos estabelecer procedimentos sobre a aprendizagem e habilitação de condutores de veículos, e expedir a Carteira Nacional de Habilitação (CNH) (BRASIL, 2005). Em outras palavras, um dos objetivos do DETRAN-SC é "produzir motoristas" habilitados para dirigir veículos.

O termo "habilitado" significa capacitado, apto, qualificado ou competente para alguma coisa. Dentro desse contexto, um indivíduo que consegue receber a CNH está teoricamente capacitado a dirigir o veículo segundo a categoria para qual se habilitou. Entretanto, não é isso que se pensa quando se observam as estatísticas sobre acidentes de trânsito. No Brasil, o número de vítimas no trânsito, entre mortos e feridos graves ultrapassa 150 mil e os custos relacionados são estimados em torno de R$ 28 bilhões ao ano (BACCHIERI; BARROS, 2011). Esses números, muitas vezes considerados alarmantes, colocam em dúvida se os motoristas realmente estão capacitados a dirigir. Geralmente, a imprudência dos motoristas é considerada a principal causa dos acidentes de trânsito. Logo, uma ação preventiva para evitar acidentes de trânsito é conscientizar e a educar motoristas. Atuar com maior eficácia para a conscientização e para a educação dos motoristas certamente poderá contribuir para reduzir os números dos acidentes de trânsito.

A educação dos motoristas depende, dentre outros fatores, da qualidade das aulas teóricas sobre o trânsito e da forma como os motoristas são avaliados. No DETRAN-SC, duas avaliações são realizadas com os motoristas: uma teórica e outra prática. A avaliação teórica consiste na aplicação de um teste com 40 questões, enquanto que a avaliação prática é um teste de direção assistido por um avaliador.

O objetivo desse Capítulo é elaborar um Teste Adaptativo Informatizado para a avaliação teórica do DETRAN-SC para a obtenção da carteira de habilitação para motoristas utilizando como base a Sistemática para a Implantação de Testes Adaptativos Informatizados (SITAI) desenvolvida no Capítulo 0 dessa tese. Entretanto, o TAI desenvolvido não será efetivamente implantado nesse momento. Dessa forma, as Etapas 6, 7 e 8 da SITAI não serão desenvolvidas, mas algumas considerações serão feitas.

5.1. ETAPA 1: DEFINIÇÃO DO TESTE

Conforme definido na Seção 4.3.1, a primeira etapa da SITAI compreende a formulação e planejamento do teste, abrangendo os seguintes aspectos: verificar a existência prévia do teste e definir a dimensão, o traço latente e os objetivos do teste.

A primeira questão a ser verificada (Subetapa 1.1) é se o teste a ser construído já existe ou não em uma versão não adaptativa. Como mencionado anteriormente, a avaliação teórica do DETRAN-SC já existe em uma versão não adaptativa, sendo realizada por meio da aplicação de um teste convencional do tipo "papel e lápis", que pode ser aplicado também na versão informatizada, constituído por 40 questões de múltipla escolha com quatro alternativas, sendo que apenas uma delas é correta. Para que o candidato seja aprovado no exame, ele deve responder corretamente pelo menos 70% das questões, ou seja, ele deve acertar a resposta de, no mínimo, 28 questões quaisquer da prova. A prova é elaborada com a seleção de 40 questões de um banco de itens existentes. Quando o elaborador executa a formação de um exame, ele seleciona a quantidade de exames a serem impressos e pode optar por gerar todos os exames com os mesmos 40 itens ou não. Entretanto, quando os exames não possuem os mesmos itens, nada garante que o nível de dificuldade dos diferentes exames seja equivalente e, provavelmente, não é. Isso tornaria o processo de seleção não imparcial, pois alguns candidatos podem receber uma prova mais fácil do que outros. Essa é uma limitação da TCT discutida no Capítulo 1, que será suprida aqui pela utilização da TRI, vista no Capítulo 2, a qual permite que o candidato seja corretamente avaliado independentemente do nível de dificuldade do teste.

O próximo passo é verificar se o teste existente é válido. O teste teórico do DETRAN-SC tem sido utilizado há muito tempo para a obtenção da CNH. Por ser um teste importante, acredita-se que todos os requisitos relacionados com as suas propriedades psicométricas tenham sido alcançados. Dessa forma, nesse estudo, não será verificada a validade do teste existente, porém será suposto que o teste é válido.

Considerando que o teste é válido, deve-se decidir se será feita uma adaptação do teste existente para um TAI a partir dos dados históricos existentes ou se será desconsiderado e construído desde o princípio. Usualmente, as versões adaptativas de testes tradicionais aproveitam os dados históricos, o que permite um ganho substancial na elaboração de um TAI em relação à economia de recursos financeiros e de tempo com a elaboração dos itens e a coleta de dados. Nesse estudo, serão utilizados os dados históricos (banco de itens e de respostas dos

examinandos) dos testes aplicados no ano de 2008 pelo DETRAN-SC. Acredita-se que o banco de itens existente deve ser adequado para esse estudo, pois provavelmente deve ter sido elaborado por uma equipe multidisciplinar especializada no assunto e, portanto, não existe a necessidade de elaborar novamente os itens. Todos esses dados foram obtidos com a devida autorização e permissão do DETRAN-SC para fins de estudos acadêmicos. Como os itens são sigilosos, assim como os testes de proficiência, eles não serão divulgados e serão identificados por um código. Em 2008 foram aplicadas 221.933 provas para 178.828 candidatos (alguns fizeram o exame mais de uma vez), sendo 31.168 provas distintas que agregaram um total de 462 questões.

Como o teste já existe e os dados históricos serão utilizados, deve-se verificar se as definições existentes sobre dimensão, traço latente e objetivo serão utilizadas (Subetapa 1.2). Nesse estudo, essas definições serão utilizadas. O traço latente analisado é o nível de conhecimento para conduzir veículos. Esse conhecimento envolve os seguintes conteúdos: legislação de trânsito, direção defensiva, funcionamento do veículo, meio ambiente e convívio social, e primeiros socorros (OLMA, 2008) o que poderia indicar uma possível avaliação multidimensional. Entretanto, nesse estudo, o traço latente será considerado e tratado como unidimensional, se possível. O objetivo do teste é classificar o candidato em habilitado ou não habilitado para conduzir veículos, embora seja apenas uma etapa do processo pois o candidato ainda precisa passar pela avaliação prática. Nesse tipo de objetivo, precisa-se definir um ponto de corte na escala de proficiência que será criada, o qual será definido mais adiante. No entanto, nessa análise, será construído também outro teste com o objetivo de estimar o traço latente, para fins de estudo.

Como o banco de itens existente não está calibrado pela TRI (Subetapa 1.3), a próxima etapa, segundo a SITAI, consiste na calibração do banco de itens (Etapa 3).

5.2. ETAPA 2: ELABORAÇÃO DOS ITENS

Embora a etapa 2 não seja explorada nesse estudo, vale a pena comentar sobre alguns aspectos relacionados a elaboração dos itens.

Se o banco histórico de itens não fosse utilizado, outros recursos poderiam ser aplicados na elaboração dos itens, como, por exemplo, a utilização de recursos multimídia e a técnica de geração automática de itens de itens (GAI), conforme a Subetapa 2.1 da SITAI.

Como o TAI é realizado no computador, os chamados “tipos inovadores de itens” (PARSHALL et al., 2010), ou seja, aqueles que se

beneficiam dos recursos computacionais, tais como, uso de sons, gráficos, animação, vídeo ou *mouse*, poderiam ser criados. Por exemplo, uma imagem em movimento poderia simular uma situação de trânsito e questionar ao candidato qual seria a atitude a ser tomada em tal situação. Ou ainda, poderia ser solicitado ao candidato que marcasse com o *mouse* o trajeto a ser percorrido em uma figura conforme as placas de trânsito existentes no percurso. Ou ainda, poderia ser solicitado ao candidato que marcasse determinado componente em uma figura com o motor do carro.

A técnica GAI (GLAS; VAN DER LINDEN; GEERLINGS, 2010), poderia ser utilizada, por exemplo, em questões que relacionem placas de trânsito e seus significados, onde dezenas de combinações entre enunciados e possibilidades de alternativas poderiam ser elaboradas automaticamente usando a mesma regra. Outras dezenas de combinações poderiam ser criadas por meio da GAI entre o tipo de infração cometida no trânsito e a sua penalização.

Quanto à quantidade de itens necessários para a construção do TAI, onde os itens são sigilosos (Subetapa 2.2), a quantidade de 462 itens, a princípio, é adequada, considerando que o banco de itens deve ter pelo menos 5 vezes a quantidade de itens que será aplicada no teste e que o teste adaptativo deve ter menos itens aplicados do que o teste convencional. Embora ainda não esteja definido o tamanho do teste, um TAI bem elaborado permite reduzir o tamanho de um teste convencional em aproximadamente 50% (WAINER, 2000b). Supondo que o teste venha a ser composto por 20 itens (metade do tamanho original de 40 itens), tem-se uma quantidade de aproximadamente 23 vezes mais itens do que o provável tamanho do teste, embora ainda esses precisem ser calibrados e avaliados.

5.3. ETAPA 3: CALIBRAÇÃO DO BANCO DE ITENS E CONSTRUÇÃO DA ESCALA

A terceira etapa da SITAI consiste na calibração do banco de itens, conforme definido na Seção 4.3.3. Primeiramente, foi feita uma análise preliminar dos itens, conforme a Subetapa 3.1. Como os dados já estavam coletados, não foi necessário obter uma nova amostra de respondentes. Primeiramente, foi feita uma "limpeza" no banco de itens, excluindo-se os dados problemáticos, por exemplo, itens que já haviam sido excluídos pelo DETRAN-SC, porém constavam na base de dados e respostas de candidatos duplicadas.

Com o tipo de item existente nesse estudo e considerando o traço latente unidimensional, dois possíveis modelos da TRI poderiam ser

ajustados: o Modelo Logístico de 3 parâmetros (ML3) de Birnbaum (1980), apresentado na Seção 2.2, e o Modelo de Resposta Nominal de Bock (1972). O modelo escolhido para essa análise foi o ML3, que permite estimar o parâmetro de acerto casual. As grandes aplicações em larga escala para avaliar proficiência têm preferido a utilização desse modelo para a calibração dos itens. Para a utilização desse modelo, as respostas dos itens (categorias A, B, C e D) foram dicotomizadas em duas categorias: (1) correta e (0) incorreta. Itens administrados que não foram respondidos ou que tiveram mais de uma alternativa assinalada foram considerados como resposta incorreta.

Para verificar o pressuposto da unidimensionalidade, foi realizada uma análise fatorial baseada nas correlações tetracóricas utilizando-se o software TESTFACT (BOCK et al., 2003). Esse software foi escolhido por utilizar a mesma formatação do banco de respostas que o BILOG-MG (TOIT, 2003). No programa da sintaxe do software, foram selecionadas as opções para analisar um único fator e para utilizar as estimativas do acerto casual (motivo pelo qual a análise da dimensionalidade foi feita depois da estimação dos parâmetros dos itens pelo BILOG-MG). O resultado obtido, após mais de 12 horas de processamento computacional, mostrou que um único fator foi responsável por explicar 82,82% da variabilidade geral dos dados, sendo um forte indicativo para aceitar a suposição de unidimensionalidade nos dados. Dessa forma, conclui-se que o teste está medindo um único traço latente, ou seja, o teste mede o nível de conhecimento para conduzir veículos.

Antes de prosseguir com o processo iterativo da calibração do banco de itens, foi feita uma análise preliminar dos 462 itens utilizando a TCT e a TRI. Os itens foram aplicados, em média, 19.215 vezes, entretanto, a maioria dos itens (60,6%) teve uma freqüência de administração menor do que 10.000 vezes. Dois itens foram aplicados em todos os exames realizados, ou seja, 221.933 vezes. O item menos administrado foi aplicado 3.155 vezes, o que significa que o processo de calibração provavelmente não será afetado por falta de amostra.

O processo de calibração do banco de itens foi realizado através da aplicação do software BILOG-MG[12] (TOIT, 2003), o qual apresenta três fases (etapas) de análise: (1) análise clássica (TCT), (2) estimação dos itens (fase da calibração), e (3) estimação da proficiência. No

[12] A análise clássica apresentada pelo BILOG-MG é muito precária. Existem outros softwares mais adequados para fazer uma análise baseada na teoria clássica, por exemplo, o ITEMAN (THOMPSON; GUYER, 2010).

processo da calibração, o interesse está na fase 1 e principalmente na fase 2. Quanto à proporção de acertos (índice de dificuldade da TCT), obtida por meio do Software BILOG-MG, os valores variaram entre 0,08 e 0,97, sendo que 92,2% dos itens tiveram mais de 50% de certo. O valor médio da proporção de acertos foi de 0,75, sugerindo que os itens em geral são fáceis. O valor médio do coeficiente de correlação bisserial, obtido também pelo BILOG-MG, foi de 0,521, variando entre -0,307 e 1,08. De acordo com Ceccato et al. (2008), esse coeficiente da TCT mede a correlação entre o resultado de um item particular do teste (certo ou errado) e o total de acertos de cada respondente, sendo que valores negativos ou positivos próximos de zero indicam que o item não possui consistência interna, sendo, portanto, inadequado. O BILOG-MG calcula um coeficiente bisserial modificado, eliminando do total os valores referentes ao item que está sendo correlacionado. O BILOG-MG não fornece os valores do coeficiente de discriminação da TCT. Embora essa breve análise utilizando indicadores da TCT tenha mostrado que alguns itens não são adequados, optou-se por iniciar o processo de calibração utilizando todos eles, já que a TRI também tem a capacidade de identificar itens inadequados.

Todo o banco de itens foi calibrado simultaneamente numa única etapa, como se fosse uma única grande prova, onde cada respondente apresenta uma resposta para 40 itens e não apresenta resposta para 422 itens, que são tratados como "não apresentados" e não interferem na estimativa da proficiência do candidato. Nesse estudo, os itens foram calibrados na escala (0, 1), ou seja, com média igual a zero e desvio padrão igual a um. A fase 1 do BILOG-MG apresentou os resultados da TCT mencionados anteriormente e também identificou um item que não pode ser calibrados devido ao coeficiente bisserial ser menor do que $-0{,}15$, restando, portanto, 461 itens. Na fase 2, optou-se por oito formas diferentes de calibração, a fim de avaliar o comportamento dos itens nessas diferentes situações: (1) sem a utilização de distribuições a priori nos parâmetros dos itens; (2) com priori no parâmetro *a*; (3) com priori no parâmetro *b*; (4) com priori no parâmetro *c*; (5) com priori nos parâmetros *a* e *b*; (6) com prioris nos parâmetros *a* e *c*; (7) com prioris nos parâmetros *b* e *c*; e (8) com prioris nos parâmetros *a*, *b* e *c*. A Tabela 2 apresenta os valores médios dos parâmetros de todos os itens e dos seus erros padrões (EP) segundo o método de estimação utilizado.

Segundo os resultados da Tabela 2, observou-se que nos métodos com priori no parâmetro *a* e com priori nos parâmetros *a* e *c*, os métodos numéricos utilizados pelo BILOG-MG no processo de estimação dos

itens não conseguiram estimar os parâmetros dos itens, o que mostra que esses métodos não são adequados para esse estudo. Os métodos sem priori, com priori no parâmetro *b* e com priori nos parâmetros *a* e *b* não se mostraram adequados porque não conseguiram estimar com eficiência o valor do parâmetro *c*, cujo valor médio foi praticamente zero. Dado que o parâmetro *c* é a probabilidade de acerto casual, a qual é maior que zero numa prova de múltipla escolha com quatro alternativas, concluiu-se que esses métodos não são adequados para esse estudo. Quanto aos métodos com priori no parâmetro *c*, com priori nos parâmetros *b* e *c* e com priori nos parâmetros *a, b* e *c*, observou-se que eles apresentaram praticamente os mesmos valores médios para os três parâmetros do modelo, mostrando-se adequados para esse estudo. Dessa forma, optou-se por continuar o processo de calibração com o método que utiliza priori nos parâmetros *a, b* e *c*, pois esse foi o que apresentou ou menores valores médios dos EP para os parâmetros *a* e *b*.

Tabela 2 Média dos parâmetros dos itens e dos seus respectivos EP

	Parâmetros e EP					
Método	*a*	*EP(a)*	*b*	*EP(b)*	*c*	*EP(c)*
Sem priori	1,29	0,0906	-1,13	0,5209	0,01	0,1440
Com priori no parâmetro *a*	-	-	-	-	-	-
Com priori no parâmetro *b*	1,29	0,0824	-1,13	0,3326	0,00	0,1180
Com priori no parâmetro *c*	1,17	0,0615	-1,10	0,1249	0,07	0,0262
Com priori nos parâmetros *a* e *b*	1,29	0,0809	-1,20	0,3022	0,00	0,1141
Com priori nos parâmetros *a* e *c*	-	-	-	-	-	-
Com priori nos parâmetros *b* e *c*	1,16	0,0604	-1,11	0,1130	0,07	0,0260
Com priori nos parâmetros *a, b* e *c*	1,16	0,0598	-1,16	0,1127	0,07	0,0262

Itens com valores baixos no parâmetro *a* também possuem correlação bisserial baixa e vice-versa (o coeficiente de correlação linear entre eles foi 0,91). A Figura 24, gerada pelo BILOG-MG, mostra a função de informação total (FIT) do banco de itens (linha contínua) e o erro padrão (EP) associado (linha tracejada), considerando todos os 461

itens na primeira rodada da calibração com o método que utiliza priori nos parâmetros *a*, *b* e *c*.

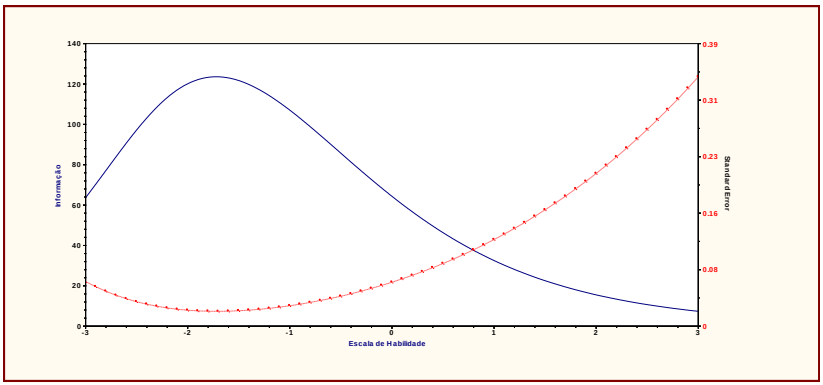


Figura 24. FIT após a primeira calibração

O gráfico da Figura 24 indica que uma grande quantidade de itens é considerada fácil, já que boa parte da informação se encontra abaixo do valor zero, onde estão posicionados 89,6% dos itens. As CCI de cada item da primeira calibração encontram-se no Apêndice A.

A análise preliminar dos itens nessa primeira rodada de calibração mostrou que o banco de itens, em geral parece adequado para a construção do TAI. O baixo erro padrão médio das estimativas dos parâmetros mostrou que a amostra foi suficiente para a análise (Subetapa 3.2). Dessa forma, prossegue-se com o processo iterativo da calibração da Subetapa 3.3.

O ideal seria eliminar um item por vez (por exemplo, com o menor valor para o parâmetro *a*) e rodar a calibração novamente para verificar o comportamento dos demais itens, que pode mudar. Pode ser conveniente fazer isso quando o banco de itens é pequeno, mas nesse caso onde o banco de itens é grande e onde cada rodada de calibração leva cerca de 30 minutos, outra estratégia deve ser adotada. Dessa forma, para a segunda rodada da calibração, optou-se por remover os itens com baixo poder de discriminação ($a < 0,5$), num total de 23 itens, restando para a análise 438 itens. Itens com valores baixos no parâmetro *a* também possuem correlação bisserial baixa e vice-versa (o coeficiente de correlação linear entre eles foi 0,90). Quanto à proporção de acertos (índice de dificuldade da TCT), os valores variaram entre 0,14 e 0,97, sendo que 93,4% dos itens tiveram mais de 50% de certo. O valor médio da proporção de acertos foi de 0,77, sugerindo que os itens em geral são

fáceis. O valor médio do coeficiente de correlação bisserial foi de 0,544, variando entre 0,121 e 1,084. Nessa etapa, houve um aumento no valor médio do parâmetro *a* para 1,21 e uma diminuição no valor médio do parâmetro *b* para -1,28. O parâmetro *c* permaneceu com média igual a 0,07. Assim como na primeira rodada de calibração, grande parte da informação se encontra abaixo do valor zero, onde estão posicionados 91,3% dos itens.

Para a terceira rodada da calibração, optou-se por remover os itens com poder de discriminação médio-baixo (a < 0,7), num total de 42 itens, restando para a análise 396 itens, com um parâmetros de discriminação aceitável. O coeficiente de correlação linear entre o parâmetro *a* e a correlação bisserial foi 0,88. Quanto à proporção de acertos (índice de dificuldade da TCT), os valores variaram entre 0,14 e 0,97, sendo que 93,7% dos itens tiveram mais de 50% de certo. O valor médio da proporção de acertos foi de 0,78, sugerindo que os itens em geral são fáceis. O valor médio do coeficiente de correlação bisserial foi de 0,569, variando entre 0,108 e 1,087. Nessa etapa, houve um aumento no valor médio do parâmetro *a* para 1,28 e uma diminuição no valor médio do parâmetro *b* para -1,31. O parâmetro *c* permaneceu com média igual a 0,07. Assim como nas rodadas anteriores de calibração, grande parte da informação se encontra abaixo do valor zero, onde estão posicionados 91,7% dos itens. Isso é um fator preocupante para o teste, já que a quantidade de itens posicionados na parte superior da escala é pouca, apenas 33 itens, sendo 4 itens entre os valores 1 e 2 na escala, 2 itens entre os valores 2 e 3, e 2 itens entre os valores 3 e 4.

Para a quarta rodada da calibração, optou-se por remover os itens com poder de discriminação médio-alto (a < 0,9), num total de 67 itens, restando para a análise 329 itens. O coeficiente de correlação linear entre o parâmetro *a* e a correlação bisserial foi 0,87. Quanto à proporção de acertos (índice de dificuldade da TCT), os valores variaram entre 0,14 e 0,97, sendo que 94,5% dos itens tiveram mais de 50% de certo. O valor médio da proporção de acertos foi de 0,80, sugerindo que os itens em geral são fáceis. O valor médio do coeficiente de correlação bisserial foi de 0,606, variando entre 0,085 e 1,09. Nessa etapa, houve um aumento no valor médio do parâmetro *a* para 1,39 e uma diminuição no valor médio do parâmetro *b* para -1,40. O parâmetro *c* permaneceu com média igual a 0,07. Assim como nas rodadas anteriores de calibração, grande parte da informação se encontra abaixo do valor zero., onde estão posicionados 93,6% dos itens

Para a quinta rodada da calibração, optou-se por manter apenas os itens com poder de discriminação alto (a > 1), sendo removidos 43 itens,

restando para a análise 286 itens. O coeficiente de correlação linear entre o parâmetro *a* e a correlação bisserial foi 0,84. Quanto à proporção de acertos (índice de dificuldade da TCT), os valores variaram entre 0,14 e 0,97, sendo que 95,5% dos itens tiveram mais de 50% de certo. O valor médio da proporção de acertos foi de 0,82, sugerindo que os itens em geral são fáceis. O valor médio do coeficiente de correlação bisserial foi de 0,631, variando entre 0,238 e 1,091. Nessa etapa, houve um aumento no valor médio do parâmetro *a* para 1,46 e uma diminuição no valor médio do parâmetro *b* para -1,46. O parâmetro *c* permaneceu com média igual a 0,07. A Figura 25 mostra a função de informação total (FIT) do banco de itens (linha contínua) e o erro padrão (EP) associado (linha tracejada), considerando todos os 286 itens da quinta rodada da calibração.

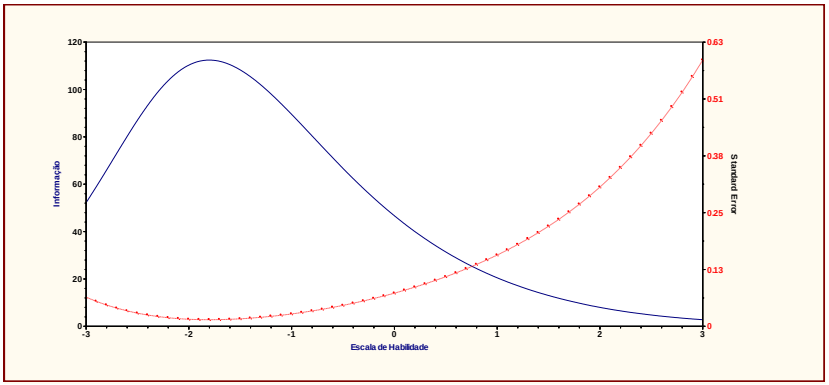


Figura 25. FIT após a quinta calibração

Assim como nas rodadas anteriores de calibração, grande parte da informação se encontra abaixo do valor zero, onde estão posicionados 94,4% dos itens. Isso é um fator preocupante para o teste, já que a quantidade de itens posicionados na parte superior da escala é muito pouca, apenas 16 itens, sendo 4 itens entre os valores 1 e 2 na escala e 1 itens entre os valores 2 e 3.

A quinta calibração resultou na permanência de itens com alto poder de discriminação, o que é bom para um TAI. Teoricamente, quanto mais discriminativos forem os itens, melhor será o desempenho do TAI. Por outro lado, a distribuição do parâmetro de dificuldade dos itens deveria ser semelhante à distribuição do traço latente na população, usualmente considerado com Distribuição Normal na maioria dos estudos. Entretanto, essa a assimetria da distribuição das dificuldades

dos itens manteve-se em todas as rodadas de calibração, ou seja, é uma característica desse conjunto de itens que não tem como ser alterada. Dessa forma, optou-se por manter o banco de itens calibrado após a quinta rodada de calibração, constituído por 286 itens. Finalizada a calibração dos itens, os parâmetros estimados foram incorporados ao banco de itens. As CCI de cada item após a quinta rodada de calibração encontram-se no Apêndice B.

Finalizada a calibração dos itens e a atualização do banco de itens, o próximo passo é a construção de a interpretação da escala. Entretanto, isso não será feito nesse estudo, por se tratar de itens sigilosos e por não haver a participação de nenhum especialista no construto envolvido nesse trabalho.

5.4. ETAPA 4: ELABORAÇÃO DO ALGORITMO

A quarta etapa da SITAI consiste na elaboração do algoritmo do teste, conforme definido na Seção 4.3.4. Para a construção do algoritmo, foram consideradas as opções disponíveis no software CATSim (WEISS; GUYER, 2010), que foi utilizado nas simulações da Etapa 5.

Primeiramente, deve-se verificar se existe alguma informação prévia do examinando (Subetapa 4.1) que possa ser utilizada para definir o nível inicial de habilidade. Embora exista a informação de avaliações anteriores de examinandos que não foram aprovados, não foi fornecida nenhuma informação prévia desse tipo para esse estudo. Dessa forma, o usual é adotar um nível de habilidade mediano igual a zero na escala (0,1). O CATSim também permite que a habilidade inicial seja um valor aleatório dentro de um determinado intervalo, por exemplo, entre -1 e 1, que é o *default* do CATSim. Essa opção permite reduzir a taxa de exposição dos primeiros itens do teste (WEISS; GUYER, 2010), o que pode ser útil nos testes onde os itens são sigilosos. Como nesse estudo os itens são sigilosos, essa opção também será considerada.

O próximo passo consiste em definir os métodos de seleção dos itens e de estimação da habilidade (Subetapa 4.2). O CATSim oferece três opções para a seleção dos itens. A primeira é o método da Máxima Informação (MI), que será considerado nesse estudo. A segunda é uma variação do método da Máxima Informação (MI) que permite selecionar aleatoriamente *m* itens com máxima informação dentre os primeiros *n* itens do teste. Essa opção permite é mais uma que reduzir a taxa de exposição dos primeiros itens do teste (WEISS; GUYER, 2010), o que pode ser útil nos testes onde os itens são sigilosos. Como nesse estudo os itens são sigilosos, serão consideradas duas opções: a primeira seleciona aleatoriamente 5 itens com máxima informação dentre os

primeiros 5 itens do teste, e a segunda seleciona aleatoriamente 10 itens com máxima informação dentre os primeiros 10 itens do teste, chamado de Máxima Informação Modificado I e Máxima Informação Modificado II, respectivamente. A terceira opção para a seleção dos itens é utilizar o método da máxima informação num determinado ponto fixo. Essa opção pode ser útil nos testes com o objetivo de classificação, onde o item selecionado maximiza a informação no ponto de corte (SPRAY; RECKASE, 1994; 1996) e será considerada na simulação do teste que tem o objetivo de classificação.

O CATSim oferece quatro métodos para a estimação da habilidade: os métodos da Máxima Verossimilhança (MV) pura ou Ponderada (MVP) e bayesianos da Esperança a Posteriori (EAP) e da Moda a Posteriori (MAP). O método MV, como visto na Seção 2.5.2, possui um problema de indeterminação no caso de respostas constantes, ou seja, enquanto as respostas aos itens forem todas as mesmas (todas corretas ou todas incorretas), o que pode ocorrer na seleção dos itens iniciais do teste, o método MV não possui um valor máximo de função. Para contornar esse problema, enquanto o padrão de resposta se mantém, o Software CATSim oferece três alternativas. Primeiro, é possível determinar o tamanho do passo, na escala de dificuldade, para a seleção do próximo item, sendo que o *default* do CATSim é tamanho do passo igual a 1, o que pode ser alterado. Por exemplo, se habilidade inicial é zero e a resposta ao primeiro item é for correta, a habilidade estimada passa a ser igual a 1. Se o próximo item ainda for respondido corretamente, a habilidade estimada passa a ser igual a 2 e assim sucessivamente até o valor 4. O mesmo raciocínio vale quando o padrão for de respostas incorretas. Esse critério pretende "quebrar" mais rapidamente o padrão de respostas, sendo que quanto maior o tamanho do passo, mais rápida será essa "quebra" (WEISS; GUYER, 2010). A segunda e a terceira alternativa oferecem a utilização dos métodos Bayesianos EAP e MAP, respectivamente. Em qualquer uma dessas três situações, uma vez "quebrado" o padrão de respostas, o método utilizado passa a ser o método MV. Nas estimações bayesianas, o *default* do CATSim considera a distribuição a priori com média igual a 0 e desvio padrão igual a 1, ou seja, priori (0,1), mas essas opções podem ser alteradas. Nesse estudo, para o método MI, serão consideradas as seguintes opções quando ocorre um padrão de respostas: com passo igual a 1, com estimação inicial Bayesiana EAP com priori (0,1) e com estimação inicial Bayesiana MAP com priori (0,1). Também serão considerados os Métodos Bayesianos puro EAP com priori (0,1) e MAP com priori (0,1) e o método da MVP.

Como nesse estudo os itens são sigilosos (Subetapa 4.3), uma importante consideração é definir a taxa de exposição dos itens, o que foi discutido na Seção 3.4.1. Considerando que os bancos de itens a serem analisados não possuem uma quantidade muito grande de itens, optou-se por utilizar uma taxa de exposição igual a 0,20 em todos os testes, para que os resultados sejam comparáveis. Isso significa que os itens devem ser expostos em no máximo 20% dos testes. Dessa forma, se o algoritmo fosse selecionar um item cuja taxa de exposição esteja acima de 0,20, esse item seria desconsiderado (como se não existisse) e outro item seria selecionado no lugar desse.

Outras restrições deveriam ser consideradas (Subetapa 4.4), tais como o balanceamento do conteúdo (vide Seção 3.4.2) e o controle de itens que não podem ser expostos para um mesmo examinando (vide Seção 3.4.3), os quais o CATSim denomina "*enemy items*". Certamente o balanceamento do conteúdo deve ser considerado, pois o teste considera vários conteúdos, conforme Olma (2008) mencionados na Seção 5.1. Porém, como não foi fornecida nenhuma informação sobre o conteúdo que cada item mede e nem a quantidade de itens de cada conteúdo que deve compor o teste, o balanceamento de conteúdo não será considerado nesse estudo. Da mesma forma, no banco de itens desse estudo podem existir itens que não deveriam ser expostos para um mesmo examinando. Porém como não foi fornecida nenhuma informação sobre isso, essa opção também não será considerada.

A Subetapa 4.5 consiste na definição dos critérios de parada. Nesse estudo, vários critérios podem ser testados. Para o teste com objetivo de estimação, optou-se por testar duas opções: nível de precisão (recomendado) e quantidade fixa de itens. Para o teste com objetivo de classificação, optou-se por utilizar o intervalo de confiança (IC), como é recomendado. O CATSim oferece seis opções para teste com tamanho variado de itens ou duas opções para teste com tamanho fixo. Entre elas, estão as opções selecionadas para esse estudo: nível de precisão onde o *default* do Software é igual a 0,2, quantidade fixa de itens e IC onde o *default* do Software é 2,0 vezes o erro padrão. Para o objetivo de classificação, o teste irá finalizar quando a estimativa da habilidade estiver afastada 2,0 vezes o erro padrão acima ou abaixo do ponto de corte (IC de 95,4%) ou quando for aplicado um máximo de 40 itens (não faz sentido aplicar mais de 40 itens, que é a quantidade fixa no teste convencional). Esse limite de itens foi determinado porque em algumas situações a habilidade estimada estará próxima do ponto de corte, o que precisaria de muitos itens para manter o IC afastado. Para o objetivo de estimação, um dos testes irá finalizar quando o erro padrão (EP) da

estimativa da habilidade for menor ou igual a 0,2 (nível de precisão) ou quando for aplicado um máximo de 40 itens, nas situações onde precisaria de muitos itens para alcançar o nível de precisão determinado. O outro teste de estimação irá finalizar quando forem aplicados 20 itens. Esse número foi definido a partir do levantamento bibliográfico sobre TAI realizado no Capítulo 3, onde vários autores mencionam que um TAI reduz cerca de 50% a quantidade de itens aplicados em um teste convencional.

O quadro 1 apresenta um resumo de todas as opções que foram selecionadas para cada elemento do algoritmo.

ELEMENTOS DO ALGORITMO	OPÇÕES SELECIONADAS
Modelo de Resposta ao item	Modelo Logístico Unidimensional de Três Parâmetros – ML3
Nível de habilidade Inicial	Mediano igual a zero. Um valor aleatório ente -1 e 1
Método de Seleção dos itens	Máxima Informação Máxima Informação Modificado I Máxima Informação Modificado II Máxima Informação no ponto de corte, se o objetivo for classificação.
Método de Estimação da habilidade	MV c/ passo = 1. MV c/ estimação inicial EAP MV c/estimação inicial MAP Bayesiano EAP Bayesiano MAP MVP
Restrições	Taxa de exposição = 0,20
Critério de parada	IC de 95,4% fora do ponto de corte (ou ponto de corte com distância de 2 EP do IC) e máximo de 40 itens, se o objetivo for classificação. Erro Padrão menor ou igual a 0,20 e máximo de 40 itens, se objetivo for estimação de θ. Quantidade fixa de 20 itens, se objetivo for estimação de θ.

Quadro 1: Opções possíveis para o algoritmo

5.5. ETAPA 5: ANÁLISE DA PRECISÃO E DA VALIDADE

Nessa etapa, foi analisado o desempenho dos algoritmos por meio de simulações realizadas pelo Software CATSim. Para simular as habilidades, duas situações distintas serão consideradas: que as habilidades provêm de uma distribuição Normal Padrão e que as habilidades provêm de uma distribuição Uniforme entre -3 e 3. Em cada caso, serão simulados 10.000 valores para as habilidades. Os resultados dessas duas situações não podem ser comparados diretamente, pois são baseados na suposição de que a habilidade provém de diferentes distribuições, o que afetará o processo de seleção dos itens, que depende, além da habilidade estimada do indivíduo, da taxa de exposição do item e, consequentemente afetará os indicadores utilizados nas análises. Por exemplo, espera-se que o EPM seja menor quando a habilidade provém de uma distribuição Normal (onde há mais respondentes na parte central) do que uma Uniforme.

Além disso, como definido na Seção 5.1, serão considerados dois objetivos distintos: teste para classificação e teste para estimação da habilidade. Nos testes de classificação (Subetapa 5.1), o objetivo é classificar o indivíduo, independente da nota que ele teve. Nesses testes, deverá ser definido o ponto de corte, o que será feito antes das simulações. Já nos testes de estimação, o objetivo é estimar a nota do indivíduo, independente da sua classificação. Os resultados dessas duas situações também não podem ser comparados diretamente, pois o objetivo do teste afetará o processo de seleção dos itens e os indicadores utilizados nas análises. Por exemplo, espera-se que teste como o objetivo de classificação utilize menos itens em alguns níveis de proficiência do que o teste com o objetivo de estimação.

Os critérios utilizados para a análise da precisão e da validade (Subetapa 5.2) dos testes serão: o erro padrão médio (EPM), a raiz quadrada do erro quadrado médio (RQEQM), o desvio empírico médio (DEM), a eficiência (EF) e a correlação linear (CL), definidos na Seção 3.5. Todos esses critérios serão calculados, porém, para cada situação, uns são mais preferíveis do que os outros, conforme foi discutido na Subetapa 5.2. Ainda, para cada situação, vários testes serão analisados e comparados (Subetapa 5.2).

5.5.1. Teste de Classificação para População Normal

O primeiro teste analisado foi o teste com objetivo de classificação considerando que as habilidades provêm de uma distribuição Normal Padrão. Essa situação é a que mais se parece com a realidade do problema.

Observando-se o Quadro 1, verifica-se que o número de algoritmos a serem testados nessa situação é de 100, considerando todas as combinações possíveis dentre as opções selecionadas. Essa quantidade de simulações é muito grande, exige muito tempo e alguns algoritmos podem não ter um desempenho satisfatório. Para diminuir esse valor, alguns algoritmos serão testados antes de outros, a fim de verificar se algum método não apresenta um desempenho satisfatório e não precise ser testado em todas as combinações possíveis. Dessa forma, decidiu-se primeiramente testar os critérios que podem ajudar no controle da exposição dos itens. Então, foram feitas simulações para cinco testes, que se diferenciam entre si segundo o nível de habilidade inicial e o método de seleção dos itens, conforme a configuração apresentada no Quadro 2.

ELEMENTOS DO ALGORITMO	OPÇÕES SELECIONADAS
Modelo de Resposta ao item	Modelo Logístico Unidimensional de Três Parâmetros – ML3
Nível de habilidade Inicial	Mediano igual a zero (Teste001, Teste002, Teste003, Teste004). Valor aleatório ente -1 e 1 (Teste005)
Método de Seleção dos itens	MI (Teste001, Teste005) MI Modificado I (Teste002) MI Modificado II (Teste003) MI no ponto de corte (Teste004)
Método de Estimação da habilidade	MV c/ passo = 1.
Restrições	Taxa de exposição = 0,20
Critério de parada	Intervalo de 95,4% de Confiança fora do ponto de corte e máximo 40 itens.

Quadro 2: Algoritmos testados

Observando-se o Quadro 2, verifica-se que os Testes 001, 002, 003 e 004 utilizam o mesmo critério para a estimativa inicial da habilidade (mediano igual a zero), mas diferem quanto ao método de seleção do item, enquanto que o Teste 005 utiliza um valor aleatório entre -1 e 1 para a habilidade inicial e o método MI puro para a seleção dos itens. Esses testes foram executados com um banco com 329 itens, resultante da terceira calibração com $a>0,7$, conforme Seção 5.3, pois inicialmente fora cogitada a inclusão desse banco de itens nas simulações. Posteriormente, optou-se trabalhar com apenas um banco de itens com 286 itens, resultante da quinta calibração com $a>1$, conforme Seção 5.3, para reduzir a quantidade de simulações. Dessa forma, a

numeração dos testes apresentada nos próximos tipos de testes não segue uma ordem sequencial direta.

Embora já esteja definido que Intervalo de Confiança será o critério de parada, falta definir qual será o ponto de corte. Conforme Segall (1997), o ponto de corte deve preservar as taxas atuais de fluxo do teste convencional, ou seja, deve-se manter a proporção de aprovados e reprovados. Dessa forma, procedeu-se da seguinte maneira:

1. Verificou-se que a proporção de examinandos reprovados em 2008, segundo os dados obtidos, foi de 0,253548.
2. Supondo que a habilidade provém de uma distribuição Normal Padrão, verificou-se que o valor do eixo x que separa os 25,3548% dos casos mais baixos era -0,663.
3. Adotou-se o valor -0,663 como ponto de corte na escala de proficiência, supondo que 25,3548% dos indivíduos na população devem ter habilidade inferior a esse valor.

A Tabela 3 apresenta o desempenho dos cinco testes analisados, segundo os critérios EPM, RQEQM, DEM, EF e CL.

Tabela 3 Desempenho dos Testes

TESTE	EPM	RQEQM	DEM	EF	CL
Teste001	0,6080	0,7447	0,2318	20,1056	0,8207
Teste002	0,6094	0,7669	0,2131	19,8947	0,8112
Teste003	0,6123	0,7781	0,2192	20,0827	0,8176
Teste004	0,6007	0,6819	0,1081	28,7146	0,8486
Teste005	0,6113	0,7712	0,2104	19,9101	0,8171

Observa-se na Tabela 3 que o Teste004 teve o pior desempenho entre todos o teste, já que a quantidade média de itens aplicado (EF) foi quase 50% superior do que nos demais. Naturalmente, essa quantidade maior de itens aplicados forneceu resultados ligeiramente melhores que os demais testes quanto ao EPM, a RQEQM, o DEM e a CL. Isso significa que o critério utilizado para a seleção dos itens que maximiza a informação no ponto de corte não teve um bom desempenho e, portanto, não será considerado nas demais simulações. Os demais testes tiveram desempenho muito parecido. A Tabela 4 apresenta o desempenho dos testes em relação ao acerto da classificação (Aprovado/Reprovado) segundo as habilidades simuladas consideradas verdadeiras.

Tabela 4 Percentual de acerto e erro na classificação

TESTE	Acerto	Ap/Ap[1]	Re/Re[2]	Erro	Ap/Re[3]	Re/Ap[4]
Teste001	92,45	94,63	85,81	7,55	5,37	14,19
Teste002	92,09	94,57	84,55	7,91	5,43	15,45
Teste003	92,26	94,62	85,08	7,74	5,38	14,92
Teste004	92,42	94,27	86,78	7,58	5,73	13,22
Teste005	91,89	94,26	84,67	8,11	5,74	15,33

[1] Percentual de aprovados que o teste classificou como aprovados.
[2] Percentual de reprovados que o teste classificou como reprovados.
[3] Percentual de aprovados que o teste classificou como reprovados[13].
[4] Percentual de reprovados que o teste classificou como aprovados[14].

A Tabela 4 mostra que os testes tiveram desempenho muito semelhante, em relação ao acerto de classificação, ficando em torno de 92%. Baseado nos resultados da Tabela 3 e da Tabela 4, conclui-se que os testes 001, 002, 003 e 005 tiveram desempenho semelhante, ou seja, a utilização de um valor aleatório entre -1 e 1 para estimar a habilidade inicial (Teste005) e a utilização dos métodos de seleção dos itens MI modificado I (Teste002) e MI modificado II (Teste003) não produziram resultados consideravelmente melhores do que o teste que utiliza o valor zero para a habilidade inicial e o método MI de seleção de itens (Teste001). Dessa forma, optou-se por não considerar os métodos MI modificados e o valor aleatório entre -1 e 1 nas próximas simulações, o que diminuirá a quantidade de combinações possíveis entre as opções selecionadas, simplificando os resultados. Entretanto, na aplicação prática, se a taxa de exposição do item não for feita após a aplicação de cada teste, esses métodos deverão ser utilizados, pois são formas alternativas de controlar a exposição dos itens sem a necessidade de atualizar a taxa de exposição a cada aplicação do teste. O Quadro 3 apresenta um resumo de todas as opções restantes para cada elemento do algoritmo.

[13] Esse resultado é conhecido como "falso negativo", ou seja, o percentual de indivíduos que foram reprovados no teste, mas na verdade não deveriam ter sido reprovados, ou seja, a habilidade verdadeira deles estava acima do ponto de corte.
[14] Esse resultado é conhecido como "falso positivo", ou seja, o percentual de indivíduos que foram aprovados no teste, mas na verdade não deveriam ter sido aprovados, ou seja, a habilidade verdadeira deles estava abaixo do ponto de corte.

ELEMENTOS DO ALGORITMO	OPÇÕES SELECIONADAS
Modelo de Resposta ao item	Modelo Logístico Unidimensional de Três Parâmetros – ML3
Nível de habilidade Inicial	Mediano igual a zero.
Método de Seleção dos itens	Máxima Informação.
Método de Estimação da habilidade	MV c/ passo = 1. MV c/ estimação inicial EAP MV c/estimação inicial MAP Bayesiano EAP Bayesiano MAP MVP
Restrições	Taxa de exposição = 0,20
Critério de parada	IC de 95,4% fora do ponto de corte (ou ponto de corte com distância de 2 EP do IC) e máximo de 40 itens.

Quadro 3: Opções possíveis para o algoritmo

Conforme disposto no Quadro 3, a quantidade de combinações possíveis nessa situação agora é 6, onde as diferenças estão focadas no apenas no método de estimação da habilidade, conforme mostra o Quadro 4.

TESTE	ESTIMAÇÂO DE θ
Teste011	MV c/ passo = 1.
Teste012	MV c/ estimação inicial EAP
Teste013	MV c/estimação inicial MAP
Teste014	Bayesiano EAP
Teste015	Bayesiano MAP
Teste016	MVP

Quadro 4: Algoritmos testados

A Tabela 5 apresenta o desempenho dos 6 testes analisados, segundo os critérios EPM, RQEQM, DEM, EF e CL, enquanto que a Tabela 6 apresenta o desempenho dos testes em relação ao acerto da classificação (Aprovado/Reprovado) segundo as habilidades simuladas consideradas verdadeiras.

Os resultados apresentados na Tabela 5 e na Tabela 6 mostram que nenhum desses testes pode ser considerado inadequado, segundo os indicadores analisados. Observa-se na Tabela 5 que o teste que utilizou o método MV com passo igual a um (Teste011) teve desempenho um inferior aos demais, em relação aos critérios EPM, RQEQM, DEM e EF

e só foi melhor que o Teste016 em relação a CL. Os testes que utilizaram o método MV com estimação inicial bayesiana (Teste012 e Teste013) tiveram desempenho ligeiramente superior aos demais, em relação a EF. Os testes com métodos bayesianos (Teste014 e Teste015) tiveram desempenho superior aos demais, em relação ao EPM, a RQEQM, ao DEM e a CL, sendo que o método MAP foi ligeiramente superior em relação a EPM, a EF.

Tabela 5 Desempenho dos Testes

TESTE	EPM	RQEQM	DEM	EF	CL
Teste011	0,6079	0,7209	0,1647	21,0787	0,8273
Teste012	0,5318	0,5490	0,0290	19,4205	0,8344
Teste013	0,5195	0,5424	0,0156	19,4090	0,8318
Teste014	0,4783	0,4794	0,0194	19,6717	0,8505
Teste015	0,4651	0,4795	-0,0093	19,5231	0,8497
Teste016	0,5563	0,6498	0,0606	19,9361	0,8242

Pela Tabela 5, observa-se que a quantidade média de itens aplicada ficou em torno de 20 nos testes simulados. Simulando novamente o Tese001 sem controlar a taxa de exposição de item e sem fixar um limite máximo de 40 itens, a fim de verificar experimentalmente qual seria a quantidade média de itens, obteve-se uma média de 54 itens, sendo que em 7,94% das situações foram aplicados todos os itens do banco. Em termos de classificação, esse teste obteve um desempenho ligeiramente superior aos demais, obtendo um percentual de acerto de 95,64% dos itens. Ou seja, um aumento médio de 34 itens aplicados proporcionou um aumento de menos de 4% na classificação correta do teste. Dessa forma, conclui-se que o mais adequado é impor um limite de itens a ser aplicado, pois a perda no percentual da classificação correta é pequena.

A Tabela 6 mostra que todos os testes tiveram desempenho muito semelhante, em relação ao acerto de classificação, ficando em torno de 92%. O maior percentual de acerto foi do Teste015 (92,81%), porém o maior percentual de aprovados que o teste classificou como aprovados foi do Teste014 (96,23%) e o maior percentual de reprovados que o teste classificou como reprovados foi do Teste016 (86,37%). Os testes com método de estimação bayesiano reprovaram menos examinandos que deveriam ter sido aprovados em relação aos demais testes, porém aprovaram mais examinandos que deveriam ter sido reprovados. O método MVP (Teste016) foi o que aprovou menos examinandos que

deveriam ter sido reprovados, porém teve um percentual de acertos levemente inferior aos demais testes. Nos dados simulados, a taxa de reprovação foi de 24,73%. O teste com a taxa de reprovação mais próxima da verdadeira foi o Teste011 (24,93).

Tabela 6 Percentual de acerto e erro na classificação

TESTE	Acerto	Ap/Ap[1]	Re/Re[2]	Erro	Ap/Re[3]	Re/Ap[4]
Teste011	91,74	94,38	83,70	8,26	5,62	16,30
Teste012	91,81	94,15	84,67	8,19	5,85	15,33
Teste013	92,05	94,53	84,51	7,95	5,47	15,49
Teste014	92,49	96,23	81,12	7,51	3,77	18,88
Teste015	92,81	96,21	82,45	7,19	3,79	17,55
Teste016	91,93	93,76	86,37	8,07	6,24	13,63

[1] Percentual de aprovados que o teste classificou como aprovados.
[2] Percentual de reprovados que o teste classificou como reprovados.
[3] Percentual de aprovados que o teste classificou como reprovados.
[4] Percentual de reprovados que o teste classificou como aprovados.

As principais características em um teste de classificação são a eficiência (terminar o teste com menor quantidade de itens aplicados) e a proporção de acertos (ter uma alta proporção de acertos na classificação), conforme discutido na Subetapa 5.2. Dentro do contexto do DETRAN-SC, a situação que parece mais grave seria aprovar um candidato que deveria ter sido reprovado, ou seja, habilitar um motorista que ainda não está preparado para dirigir. Dessa forma, em relação aos demais métodos, os procedimentos bayesianos parecem ser menos adequados por aprovar uma quantidade maior de candidatos que deveriam ter sido reprovados. Dentre os testes restantes, o que utiliza o procedimento MVP, parece o mais adequado, considerando esse critério.

Portanto, para o teste com objetivo de classificação considerando que as habilidades provêm de uma distribuição Normal Padrão, optou-se pelo Teste016 que possui a configuração descrita no Quadro 5.

ELEMENTOS DO ALGORITMO	OPÇÕES SELECIONADAS
Modelo de Resposta ao item	Modelo Logístico Unidimensional de Três Parâmetros – ML3
Nível de habilidade Inicial	Mediano igual a zero.
Método de Seleção dos itens	Máxima Informação
Método de Estimação da habilidade	MVP.
Restrições	Taxa de exposição = 0,20
Critério de parada	Intervalo de 95,4% de Confiança fora do ponto de corte e máximo 40 itens.

Quadro 5: Algoritmo Selecionado para o TAI

A Tabela 7 apresenta a quantidade de itens aplicados nos testes segundo a classificação, enquanto que a Tabela 8 apresenta o percentual associado e a habilidade média estimada (HME). Em 68,27% dos casos, o teste terminou por situar o intervalo de confiança fora do ponto de corte (foram aplicados menos de 40 itens). Observa-se na Tabela 7, por exemplo, que apenas 2 itens foram suficientes para classificar 411 testes simulados, onde a habilidade média estimada foi 0,59, sendo que em 95,38% dos casos (Tabela 8) o teste classificou corretamente. Embora o algoritmo tenha apresentado um bom desempenho, classificando corretamente os indivíduos com um alto percentual de acerto nos testes com poucos itens, deve-se verificar a validade do teste aplicado. Ou seja, na prática, uma quantidade muito baixa de itens pode não resultar em um teste válido, ainda mais quando são abordados vários conteúdos, como nesse caso. A Tabela 8 mostra que a medida que a habilidade média estimada vai se aproximando do ponto de corte, a quantidade necessária de itens para o teste vai aumentando. Observa-se na Tabela 8 que, em várias situações, onde foram aplicados mais de 21 itens, o algoritmo classificou todos os resultados corretamente. Nota-se que em 31,73% das situações, o limite máximo de 40 itens fixado no teste foi utilizado. Nesses casos, o desempenho do algoritmo foi menor, já que classificou corretamente apenas 81,09% dos casos, sendo que, em 23,63% desses casos, foram aprovados candidatos que não deveriam ter sido aprovados. A quantidade de itens mais aplicada, nos casos em que o ponto de corte esteve fora do IC, foi de quatro itens (7,35% dos casos), onde em 96,20% dos casos o teste classificou corretamente e aprovou apenas 9,43% de examinandos que deveriam ter sido reprovados.

Tabela 7 Quantidade de itens aplicados

Itens	Acerto	Ap/Ap	Re/Re	Erro	Ap/Re	Re/Ap	Total	%
2	392	310	82	19	12	7	411	4,11
3	641	433	208	34	24	10	675	6,75
4	709	565	144	28	13	15	737	7,37
5	623	499	124	19	7	12	642	6,42
6	477	375	102	17	12	5	494	4,94
7	426	348	78	9	5	4	435	4,35
8	361	281	80	8	5	3	369	3,69
9	306	252	54	9	6	3	315	3,15
10	262	217	45	9	4	5	271	2,71
11	233	188	45	5	4	1	238	2,38
12	227	183	44	6	5	1	233	2,33
13	179	141	38	2	2	0	181	1,81
14	167	134	33	3	1	2	170	1,70
15	167	135	32	3	2	1	170	1,70
16	133	102	31	5	2	3	138	1,38
17	134	109	25	2	1	1	136	1,36
18	101	83	18	2	1	1	103	1,03
19	92	75	17	2	2	0	94	0,94
20	79	63	16	1	1	0	80	0,80
21	73	56	17	0	0	0	73	0,73
22	84	64	20	1	1	0	85	0,85
23	72	58	14	1	0	1	73	0,73
24	74	55	19	3	0	3	77	0,77
25	50	43	7	1	0	1	51	0,51
26	49	38	11	3	2	1	52	0,52
27	56	43	13	0	0	0	56	0,56
28	57	44	13	1	1	0	58	0,58
29	44	36	8	0	0	0	44	0,44
30	41	29	12	0	0	0	41	0,41
31	45	34	11	2	1	1	47	0,47
32	34	22	12	2	1	1	36	0,36
33	44	30	14	2	2	0	46	0,46
34	30	19	11	0	0	0	30	0,30
35	36	22	14	0	0	0	36	0,36
36	34	26	8	2	2	0	36	0,36
37	32	23	9	1	0	1	33	0,33
38	28	21	7	1	1	0	29	0,29
39	28	25	3	4	3	1	32	0,32
40	2573	1876	697	600	347	253	3173	31,73
Total	9193	7057	2136	807	470	337	10000	100,00

Tabela 8 Percentual segundo a classificação

Itens	Acerto	Ap/Ap	Re/Re	Erro	Ap/Re	Re/Ap	HME
2	95,38	96,27	92,13	4,62	3,73	7,87	0,59
3	94,96	94,75	95,41	5,04	5,25	4,59	0,24
4	96,20	97,75	90,57	3,80	2,25	9,43	0,71
5	97,04	98,62	91,18	2,96	1,38	8,82	0,67
6	96,56	96,90	95,33	3,44	3,10	4,67	0,51
7	97,93	98,58	95,12	2,07	1,42	4,88	0,54
8	97,83	98,25	96,39	2,17	1,75	3,61	0,37
9	97,14	97,67	94,74	2,86	2,33	5,26	0,40
10	96,68	98,19	90,00	3,32	1,81	10,00	0,35
11	97,90	97,92	97,83	2,10	2,08	2,17	0,24
12	97,42	97,34	97,78	2,58	2,66	2,22	0,15
13	98,90	98,60	100,00	1,10	1,40	0,00	0,10
14	98,24	99,26	94,29	1,76	0,74	5,71	0,08
15	98,24	98,54	96,97	1,76	1,46	3,03	0,08
16	96,38	98,08	91,18	3,62	1,92	8,82	-0,02
17	98,53	99,09	96,15	1,47	0,91	3,85	0,02
18	98,06	98,81	94,74	1,94	1,19	5,26	0,02
19	97,87	97,40	100,00	2,13	2,60	0,00	-0,02
20	98,75	98,44	100,00	1,25	1,56	0,00	-0,06
21	100,00	100,00	100,00	0,00	0,00	0,00	-0,12
22	98,82	98,46	100,00	1,18	1,54	0,00	-0,15
23	98,63	100,00	93,33	1,37	0,00	6,67	-0,07
24	96,10	100,00	86,36	3,90	0,00	13,64	-0,18
25	98,04	100,00	87,50	1,96	0,00	12,50	0,02
26	94,23	95,00	91,67	5,77	5,00	8,33	-0,18
27	100,00	100,00	100,00	0,00	0,00	0,00	-0,18
28	98,28	97,78	100,00	1,72	2,22	0,00	-0,18
29	100,00	100,00	100,00	0,00	0,00	0,00	-0,12
30	100,00	100,00	100,00	0,00	0,00	0,00	-0,28
31	95,74	97,14	91,67	4,26	2,86	8,33	-0,24
32	94,44	95,65	92,31	5,56	4,35	7,69	-0,39
33	95,65	93,75	100,00	4,35	6,25	0,00	-0,36
34	100,00	100,00	100,00	0,00	0,00	0,00	-0,39
35	100,00	100,00	100,00	0,00	0,00	0,00	-0,46
36	94,44	92,86	100,00	5,56	7,14	0,00	-0,26
37	96,97	100,00	90,00	3,03	0,00	10,00	-0,29
38	96,55	95,45	100,00	3,45	4,55	0,00	-0,31
39	87,50	89,29	75,00	12,50	10,71	25,00	-0,21
40	81,09	84,39	73,37	18,91	15,61	26,63	-0,51
Total	91,93	93,76	86,37	8,07	6,24	13,63	0,05

5.5.2. Teste de Classificação para População Uniforme

O segundo teste analisado foi o teste com objetivo de classificação considerando que as habilidades provêm de uma distribuição Uniforme no intervalo [-3, 3].

Baseado nos resultados obtidos na Seção 5.5.1, algumas configurações expostas no Quadro 1 não foram testadas porque elas não tiveram um desempenho diferenciado nas análises daquela Seção. Dessa forma, o número de algoritmos testados nessa situação, considerando todas as combinações possíveis dentre as opções selecionadas, foi reduzido para 6, conforme o Quadro 3.

Nessa situação, manter a proporção de aprovados e reprovados do teste convencional, o ponto de corte teve que ser definido novamente, segundo a distribuição Uniforme no intervalo [-3, 3]. Dessa forma, procedeu-se da seguinte maneira:

1. Verificou-se que a proporção de examinandos reprovados em 2008, segundo os dados obtidos, foi de 0,253548.
2. Supondo que a habilidade provém de uma distribuição Uniforme no intervalo [-3, 3]., verificou-se que o valor do eixo x que separa os 25,3548% dos casos mais baixos era -1,479.
3. Adotou-se o valor -1,479 como ponto de corte na escala de proficiência, supondo que 25,3548% dos indivíduos na população devem ter habilidade inferior a esse valor.

Conforme disposto no Quadro 3, a quantidade de combinações possíveis nessa situação é 6, onde as diferenças estão focadas apenas no método de estimação da habilidade, conforme mostra o Quadro 6.

TESTE	ESTIMAÇÂO DE θ
Teste029	MV c/ passo = 1.
Teste030	MV c/ estimação inicial EAP
Teste031	MV c/estimação inicial MAP
Teste032	Bayesiano EAP
Teste033	Bayesiano MAP
Teste034	MVP

Quadro 6: Algoritmos testados

A Tabela 9 apresenta o desempenho dos 6 testes analisados, segundo os critérios EPM, RQEQM, DEM, EF e CL, enquanto que a Tabela 10 apresenta o desempenho dos testes em relação ao acerto da classificação (Aprovado/Reprovado) segundo as habilidades simuladas consideradas verdadeiras.

Os resultados apresentados na Tabela 9 e na Tabela 10 mostram que nenhum desses testes pode ser considerado totalmente inadequado, segundo os indicadores analisados, porém alguns possuem desempenho claramente superior aos demais. Observa-se na Tabela 9 que o teste que utilizou o método MV com passo igual a um (Teste029) teve desempenho inferior aos demais, em relação ao EPM e a EF, porém teve desempenho superior em relação a RQEQM, ao DEM e a CL Os testes que utilizaram os métodos bayesianos e o método MV com estimação inicial bayesiana tiveram desempenho ligeiramente superior aos demais, em relação ao EPM e a EF, sendo que o método MAP foi ligeiramente superior aos outros bayesianos.

Tabela 9 Desempenho dos Testes

TESTE	EPM	RQEQM	DEM	EF	CL
Teste029	0,9293	0,9384	0,1632	15,1367	0,8830
Teste030	0,6823	1,1283	-0,3734	8,8534	0,7984
Teste031	0,6529	1,1455	-0,3812	8,5140	0,7918
Teste032	0,6677	1,1433	-0,3036	8,4795	0,7923
Teste033	0,6380	1,1778	-0,2942	7,9515	0,7742
Teste034	0,7114	0,9634	-0,2994	9,6250	0,8509

Pela Tabela 9, observa-se que a quantidade média de itens aplicada ficou em torno de 10 nos testes simulados, variando entre 8 e 15 itens, que é uma variação bem significativa. O teste que utilizou método MV (Teste029) teve que aplicar consideravelmente mais itens que os demais. Como esse teste utilizou mais itens que os demais, ele classificou corretamente mais do que os demais testes acertando 96,68% dos casos, como pode-se observar na Tabela 10.

A Tabela 10 mostra que o maior percentual de acerto foi doTeste029 (95,94%), porém o maior percentual de aprovados que o teste classificou como aprovados foi do Teste033 (99,14%) e o maior percentual de reprovados que o teste classificou como reprovados foi do Teste034 (89,37%). Os testes com método de estimação bayesiano reprovaram menos examinandos que deveriam ter sido aprovados em relação aos demais testes, porém aprovaram mais examinandos que deveriam ter sido reprovados. Os testes com métodos MV (Teste029) e MVP (Teste034) foram os que aprovaram menos examinandos que deveriam ter sido reprovados. Nos dados simulados, a taxa de reprovação foi de 25,69%. Todos os testes tiveram uma taxa de

reprovação menor do que a taxa verdadeira, sendo que a mais próxima da verdadeira foi do Teste034 (24,45%).

Tabela 10 Percentual de acerto e erro na classificação

TESTE	Acerto	Ap/Ap[1]	Re/Re[2]	Erro	Ap/Re[3]	Re/Ap[4]
Teste029	95,94	98,26	89,22	4,06	1,74	10,78
Teste030	94,32	98,05	83,53	5,68	1,95	16,47
Teste031	94,29	98,10	83,26	5,71	1,90	16,74
Teste032	94,05	98,92	79,95	5,95	1,08	20,05
Teste033	93,40	99,14	76,80	6,60	0,86	23,20
Teste034	95,78	97,99	89,37	4,22	2,01	10,63

[1] Percentual de aprovados que o teste classificou como aprovados.
[2] Percentual de reprovados que o teste classificou como reprovados.
[3] Percentual de aprovados que o teste classificou como reprovados.
[4] Percentual de reprovados que o teste classificou como aprovados.

As principais características em um teste de classificação são a eficiência (terminar o teste com menor quantidade de itens aplicados) e a proporção de acertos (ter uma alta proporção de acertos na classificação), conforme discutido na Subetapa 5.2. Como mencionado na Seção 5.5.1, dentro do contexto do DETRAN-SC, a situação que parece mais grave seria aprovar um candidato que deveria ter sido reprovado, ou seja, habilitar um motorista que ainda não está preparado para dirigir. Dessa forma, dessa forma, o teste que utilizou o método MVP parece mais adequado, pois conseguiu aprovar menos examinandos que deveriam ter sido reprovados e aplicam uma baixa quantidade média de itens.

Portanto, para o teste com objetivo de classificação considerando que as habilidades provêm de uma distribuição Uniforme no intervalo [-3, 3], optou-se pelo Teste034 que possui a configuração descrita no Quadro 7.

ELEMENTOS DO ALGORITMO	OPÇÕES SELECIONADAS
Modelo de Resposta ao item	Modelo Logístico Unidimensional de Três Parâmetros – ML3
Nível de habilidade Inicial	Mediano igual a zero.
Método de Seleção dos itens	Máxima Informação
Método de Estimação da habilidade	MVP.
Restrições	Taxa de exposição = 0,20
Critério de parada	Intervalo de 95,4% de Confiança fora do ponto de corte e máximo 40 itens.

Quadro 7: Algoritmo Selecionado para o TAI

Em 88,79% dos casos, o teste terminou por situar o intervalo de confiança fora do ponto de corte. A Tabela 11 apresenta a quantidade de itens aplicados nos testes segundo a classificação, enquanto que a Tabela 12 apresenta o percentual associado e a habilidade média estimada (HME). Observa-se na Tabela 11, por exemplo, que a quantidade de itens mais aplicada, nos casos em que o ponto de corte ficou fora do IC, foi de apenas 1 item, o suficiente para classificar 2033 indivíduos simulados, onde a habilidade média estimada foi 0,74, sendo que em 94,79% dos casos (Tabela 12) o teste classificou corretamente. A Tabela 12 mostra que a medida que a habilidade média estimada vai se aproximando do ponto de corte, a quantidade necessária de itens para o teste vai aumentando, até 5 itens, conforme Tabela 11. Nos testes de 6 a 40 itens aplicados, a habilidade média estimada oscila entre -1,58 e -1,13. Em 61,69% dos casos foram aplicados até 5 itens. Observa-se na Tabela 12 que, em várias situações, onde foram aplicados mais de 8 itens, o algoritmo classificou todos os resultados corretamente. Nota-se que em 11,21% das situações, o limite máximo de 40 itens fixado no teste foi utilizado. Nesses casos, o desempenho do algoritmo foi menor, já que classificou corretamente apenas 80,11% dos casos, sendo que, em 20,87% desses casos, foram aprovados candidatos que não deveriam ter sido aprovados.

Tabela 11 Quantidade de itens aplicados

Itens	Acerto	Ap/Ap	Re/Re	Erro	Ap/Re	Re/Ap	Total	%
1	1927	1927	0	106	0	106	2033	20,33
2	790	790	0	18	0	18	808	8,08
3	1171	1127	44	22	3	19	1193	11,93
4	1231	937	294	14	3	11	1245	12,45
5	880	575	305	10	7	3	890	8,90
6	496	273	223	6	4	2	502	5,02
7	365	182	183	8	5	3	373	3,73
8	265	129	136	0	0	0	265	2,65
9	186	75	111	1	0	1	187	1,87
10	162	80	82	2	1	1	164	1,64
11	139	72	67	0	0	0	139	1,39
12	103	51	52	0	0	0	103	1,03
13	97	51	46	2	1	1	99	0,99
14	91	45	46	0	0	0	91	0,91
15	73	39	34	2	1	1	75	0,75
16	58	34	24	1	1	0	59	0,59
17	64	34	30	1	0	1	65	0,65
18	63	40	23	2	2	0	65	0,65
19	46	23	23	0	0	0	46	0,46
20	33	19	14	1	1	0	34	0,34
21	42	30	12	0	0	0	42	0,42
22	32	23	9	0	0	0	32	0,32
23	42	25	17	1	1	0	43	0,43
24	27	18	9	0	0	0	27	0,27
25	30	19	11	0	0	0	30	0,30
26	25	13	12	0	0	0	25	0,25
27	28	15	13	1	1	0	29	0,29
28	19	11	8	1	1	0	20	0,20
29	24	17	7	0	0	0	24	0,24
30	24	15	9	0	0	0	24	0,24
31	23	17	6	0	0	0	23	0,23
32	19	13	6	0	0	0	19	0,19
33	21	13	8	0	0	0	21	0,21
34	15	10	5	0	0	0	15	0,15
35	16	10	6	0	0	0	16	0,16
36	10	5	5	0	0	0	10	0,10
37	12	9	3	0	0	0	12	0,12
38	17	10	7	0	0	0	17	0,17
39	14	10	4	0	0	0	14	0,14
40	898	496	402	223	117	106	1121	11,21
Total	9578	7282	2296	422	149	273	10000	100,00

Tabela 12 Percentual segundo a classificação

Itens	Acerto	Ap/Ap	Re/Re	Erro	Ap/Re	Re/Ap	HME
1	94,79	100,00	0,00	5,21	0,00	100,00	0,74
2	97,77	100,00	0,00	2,23	0,00	100,00	0,85
3	98,16	99,73	69,84	1,84	0,27	30,16	0,70
4	98,88	99,68	96,39	1,12	0,32	3,61	-0,26
5	98,88	98,80	99,03	1,12	1,20	0,97	-0,69
6	98,80	98,56	99,11	1,20	1,44	0,89	-1,23
7	97,86	97,33	98,39	2,14	2,67	1,61	-1,41
8	100,00	100,00	100,00	0,00	0,00	0,00	-1,44
9	99,47	100,00	99,11	0,53	0,00	0,89	-1,58
10	98,78	98,77	98,80	1,22	1,23	1,20	-1,40
11	100,00	100,00	100,00	0,00	0,00	0,00	-1,38
12	100,00	100,00	100,00	0,00	0,00	0,00	-1,41
13	97,98	98,08	97,87	2,02	1,92	2,13	-1,35
14	100,00	100,00	100,00	0,00	0,00	0,00	-1,39
15	97,33	97,50	97,14	2,67	2,50	2,86	-1,34
16	98,31	97,14	100,00	1,69	2,86	0,00	-1,29
17	98,46	100,00	96,77	1,54	0,00	3,23	-1,35
18	96,92	95,24	100,00	3,08	4,76	0,00	-1,24
19	100,00	100,00	100,00	0,00	0,00	0,00	-1,41
20	97,06	95,00	100,00	2,94	5,00	0,00	-1,33
21	100,00	100,00	100,00	0,00	0,00	0,00	-1,13
22	100,00	100,00	100,00	0,00	0,00	0,00	-1,14
23	97,67	96,15	100,00	2,33	3,85	0,00	-1,32
24	100,00	100,00	100,00	0,00	0,00	0,00	-1,22
25	100,00	100,00	100,00	0,00	0,00	0,00	-1,25
26	100,00	100,00	100,00	0,00	0,00	0,00	-1,40
27	96,55	93,75	100,00	3,45	6,25	0,00	-1,40
28	95,00	91,67	100,00	5,00	8,33	0,00	-1,37
29	100,00	100,00	100,00	0,00	0,00	0,00	-1,19
30	100,00	100,00	100,00	0,00	0,00	0,00	-1,29
31	100,00	100,00	100,00	0,00	0,00	0,00	-1,17
32	100,00	100,00	100,00	0,00	0,00	0,00	-1,22
33	100,00	100,00	100,00	0,00	0,00	0,00	-1,30
34	100,00	100,00	100,00	0,00	0,00	0,00	-1,27
35	100,00	100,00	100,00	0,00	0,00	0,00	-1,30
36	100,00	100,00	100,00	0,00	0,00	0,00	-1,39
37	100,00	100,00	100,00	0,00	0,00	0,00	-1,22
38	100,00	100,00	100,00	0,00	0,00	0,00	-1,36
39	100,00	100,00	100,00	0,00	0,00	0,00	-1,22
40	80,11	80,91	79,13	19,89	19,09	20,87	-1,45
Total	95,78	97,99	89,37	4,22	2,01	10,63	-0,32

5.5.3. Teste de Estimação para População Normal com Quantidade Fixa de Itens

O terceiro teste analisado foi o teste com objetivo de estimação considerando que as habilidades provêm de uma distribuição Normal Padrão.

Baseado nos resultados obtidos na Seção 5.5.1, algumas configurações expostas no Quadro 1 não foram testadas porque elas não tiveram um desempenho diferenciado nas análises daquela Seção. Dessa forma, o número de algoritmos testados nessa situação, considerando todas as combinações possíveis dentre as opções selecionadas, foi reduzido para 6, conforme mostra o Quadro 8.

ELEMENTOS DO ALGORITMO	OPÇÕES SELECIONADAS
Modelo de Resposta ao item	Modelo Logístico Unidimensional de Três Parâmetros – ML3
Nível de habilidade Inicial	Mediano igual a zero.
Método de Seleção dos itens	Máxima Informação.
Método de Estimação da habilidade	MV c/ passo = 1. MV c/ estimação inicial EAP MV c/estimação inicial MAP Bayesiano EAP Bayesiano MAP MVP
Restrições	Taxa de exposição = 0,20
Critério de parada	Erro Padrão menor ou igual a 0,20 e máximo de 40 itens. Quantidade fixa de 20 itens.

Quadro 8: Opções possíveis para o algoritmo

Embora, nesse caso, o objetivo seja estimação, na análise também será verificado o desempenho dos algoritmos para a classificação do candidato (aprovado/reprovado). Dessa forma, adotou-se o mesmo valor utilizado na Seção 5.5.1 para o ponto de corte na escala de proficiência, ou seja, -0,663.

Conforme disposto no Quadro 8, a quantidade de combinações possíveis nessa situação é 6, onde as diferenças estão focadas apenas no método de estimação da habilidade, conforme mostra o Quadro 9.

TESTE	ESTIMAÇÂO DE θ
Teste047	MV c/ passo = 1.
Teste048	MV c/ estimação inicial EAP
Teste049	MV c/estimação inicial MAP
Teste050	Bayesiano EAP
Teste051	Bayesiano MAP
Teste052	MVP

Quadro 9: Algoritmos testados

A Tabela 13 apresenta o desempenho dos 6 testes analisados (Quadro 9), segundo os critérios EPM, RQEQM, DEM, EF e CL, enquanto que a Tabela 14 apresenta o desempenho dos testes em relação ao acerto da classificação (Aprovado/Reprovado) segundo as habilidades simuladas consideradas verdadeiras.

Os resultados apresentados na Tabela 13 e na Tabela 14 mostram que nenhum desses testes pode ser considerado totalmente inadequado, segundo os indicadores analisados, porém alguns possuem desempenho superior em relação aos demais. O critério "eficiência" (EF), foi o mesmo para todos os testes, já que foi determinada a aplicação de 20 itens fixos e, portanto, não será considerado. Primeiramente, observa-se na Tabela 13 que o teste que utilizou o método MV com passo igual a um (Teste047), teve desempenho inferior aos demais, em relação ao EPM, a RQEQM, ao DEM e a CL. Os testes que utilizaram os métodos bayesianos tiveram desempenho superior aos demais, em relação ao EPM, a RQEQM e a CL, sendo que o método EAP foi ligeiramente superior ao método EAP, em relação a RQEQM, ao DEM e a CL.

Tabela 13 Desempenho dos Testes

TESTE	EPM	RQEQM	DEM	EF	CL
Teste047	0,5369	0,6353	0,1149	20	0,8532
Teste048	0,4644	0,4634	0,0404	20	0,8892
Teste049	0,4638	0,4653	0,0367	20	0,8857
Teste050	0,4111	0,4057	0,0079	20	0,8950
Teste051	0,4032	0,4075	-0,0173	20	0,8938
Teste052	0,4577	0,4685	-0,0110	20	0,8847

A Tabela 14 mostra que o percentual de acerto foi muito parecida entre os testes, ficando em torno de 0,90. O maior percentual de acerto foi doTeste050 (90,82%), que obteve também o maior percentual de aprovados que o teste classificou como aprovados (95,26%), porém o

maior percentual de reprovados que o teste classificou como reprovados foi do Teste052 (85,36%). Todos os testes tiveram uma taxa alta na aprovação de examinandos que deveriam ter sido reprovados, variando entre 4,74% e 7,89%. Os testes com método de estimação bayesiano reprovaram menos examinandos que deveriam ter sido aprovados em relação aos demais testes, porém aprovaram mais examinandos que deveriam ter sido reprovados. O teste com métodos MVP (Testes052) foi o que aprovou menos examinandos que deveriam ter sido reprovados (14,64%). Nos dados simulados, a taxa de reprovação foi de 24,73%. Todos os testes tiveram uma taxa de reprovação semelhante, sendo que a mais próxima da verdadeira foi do Teste047 (25,60).

Tabela 14 Percentual de acerto e erro na classificação

TESTE	Acerto	Ap/Ap[1]	Re/Re[2]	Erro	Ap/Re[3]	Re/Ap[4]
Teste047	90,39	93,04	82,33	9,61	6,96	17,67
Teste048	90,56	92,88	83,50	9,44	7,12	16,50
Teste049	89,96	92,63	81,84	10,04	7,37	18,16
Teste050	90,82	95,26	77,32	9,18	4,74	22,68
Teste051	90,73	95,06	77,56	9,27	4,94	22,44
Teste052	90,44	92,11	85,36	9,56	7,89	14,64

[1] Percentual de aprovados que o teste classificou como aprovados.
[2] Percentual de reprovados que o teste classificou como reprovados.
[3] Percentual de aprovados que o teste classificou como reprovados.
[4] Percentual de reprovados que o teste classificou como aprovados.

Os principais critérios para avaliar um teste de estimação que utiliza uma quantidade fixa de itens como critério de parada são o EPM (menor-melhor), a RQEQM (menor-melhor), o DEM (mais próximo de zero - melhor) e a CL (maior-melhor), conforme discutido na Subetapa 5.2. Como mencionado na Seção 5.5.1, dentro do contexto do DETRAN-SC, a situação que parece mais grave seria aprovar um candidato que deveria ter sido reprovado, ou seja, habilitar um motorista que ainda não está preparado para dirigir. Quanto aos critérios de avaliação, os testes com método de estimação bayesiano tiveram desempenho melhor em relação ao EPM, a RQEQM e a CL, sendo que o método EAP foi ligeiramente superior ao método EAP, em relação a RQEQM, ao DEM e a CL. Os métodos bayesianos apresentaram melhor desempenho em três dos quatro indicadores e, o teste com o método EAP foi superior aos demais em relação ao desempenho do indicador DEM. Porém, como foi visto, todos os testes tiveram uma taxa alta na aprovação de examinandos que deveriam ter sido reprovados, variando

entre 4,74% e 7,89%, sendo que os testes com método de estimação bayesiano foram os que aprovaram mais e o teste com método MVP foi o que aprovou menos. Além disso, o método MVP apresenta desempenho razoavelmente bom quanto aos critérios EPM, RQEQM, DEM e CL.

Portanto, para o teste com objetivo de estimação, utilizando-se a aplicação de 20 itens como critério de parada, considerando que as habilidades provêm de uma distribuição Normal Padrão, optou-se pelo Teste052 que possui a configuração descrita no Quadro 10.

ELEMENTOS DO ALGORITMO	OPÇÕES SELECIONADAS
Modelo de Resposta ao item	Modelo Logístico Unidimensional de Três Parâmetros – ML3
Nível de habilidade Inicial	Mediano igual a zero.
Método de Seleção dos itens	Máxima Informação
Método de Estimação da habilidade	MVP.
Restrições	Taxa de exposição = 0,20
Critério de parada	Número fixo de 20 itens.

Quadro 10: Algoritmo Selecionado para o TAI

5.5.4. Teste de Estimação para População Uniforme com Quantidade Fixa de Itens

O quarto teste analisado foi o teste com objetivo de estimação considerando que as habilidades provêm de uma distribuição Uniforme no intervalo [-3, 3].

Baseado nos resultados obtidos na Seção 5.5.1, algumas configurações expostas no Quadro 1 não foram testadas porque elas não tiveram um desempenho diferenciado nas análises daquela Seção. Dessa forma, o número de algoritmos testados nessa situação, considerando todas as combinações possíveis dentre as opções selecionadas, foi reduzido para 6, conforme o Quadro 8.

Embora nesse caso o objetivo seja estimação, na análise também será verificado o desempenho dos algoritmos para a classificação do candidato (aprovado/reprovado). Dessa forma, adotou-se o mesmo valor utilizado na Seção 5.5.2 para o ponto de corte na escala de proficiência, ou seja, -1,479.

Conforme disposto no Quadro 8, a quantidade de combinações possíveis nessa situação é 6, onde as diferenças estão focadas apenas no método de estimação da habilidade, conforme mostra o Quadro 11.

TESTE	ESTIMAÇÂO DE θ
Teste065	MV c/ passo = 1.
Teste066	MV c/ estimação inicial EAP
Teste067	MV c/estimação inicial MAP
Teste068	Bayesiano EAP
Teste069	Bayesiano MAP
Teste070	MVP

Quadro 11: Algoritmos testados

A Tabela 15 apresenta o desempenho dos 6 testes analisados (Quadro 11), segundo os critérios EPM, RQEQM, DEM, EF e CL, enquanto que a Tabela 16 apresenta o desempenho dos testes em relação ao acerto da classificação (Aprovado/Reprovado), segundo as habilidades simuladas consideradas verdadeiras. Os resultados apresentados na Tabela 15 e na Tabela 16mostram que nenhum desses testes pode ser considerado totalmente inadequado, segundo os indicadores analisados, porém alguns possuem desempenho claramente superior aos demais. O critério "eficiência" (EF) foi o mesmo para todos os testes, já que foi determinada a aplicação de 20 itens fixos e, portanto, não será considerado. Observa-se na Tabela 15 que o teste que utilizou o método MV com passo igual a um teve desempenho inferior aos demais, em relação ao EPM, a RQEQM e a CL. Os testes que utilizaram os métodos bayesianos tiveram desempenho superior aos demais, em relação ao EPM e a CL, sendo que o método EAP foi ligeiramente superior ao método MAP, em relação ao EPM, ao DEM e a CL. O teste que utilizou o método MVP teve desempenho superior aos demais, em relação a RQEQM e ao DEM.

Tabela 15 Desempenho dos Testes

TESTE	EPM	RQEQM	DEM	EF	CL
Teste065	0,7726	0,7531	0,1625	20	0,9366
Teste066	0,4932	0,5199	-0,1033	20	0,9560
Teste067	0,4916	0,5327	-0,1159	20	0,9542
Teste068	0,4232	0,5613	-0,1422	20	0,9603
Teste069	0,4124	0,5883	-0,1703	20	0,9598
Teste070	0,5322	0,5143	-0,0661	20	0,9562

A Tabela 16 mostra que o percentual de acerto foi muito parecido entre os testes, ficando em torno de 95%. O maior percentual de acerto foi do Teste066 (96,07%), porém o maior percentual de aprovados que o

teste classificou como aprovados foi do Teste068 (98,88%) e o maior percentual de reprovados que o teste classificou como reprovados foi do Teste070 (92,95%). Os testes com método de estimação bayesiano reprovaram menos examinandos que deveriam ter sido aprovados em relação aos demais testes, porém aprovaram mais examinandos que deveriam ter sido reprovados. Os testes com métodos MV e MVP foram os que aprovaram menos examinandos que deveriam ter sido reprovados. Nos dados simulados, a taxa de reprovação foi de 25,69%. Todos os testes tiveram uma taxa de reprovação semelhante, sendo que a mais próxima da verdadeira foi do Teste065 (25,57%).

Tabela 16 Percentual de acerto e erro na classificação

TESTE	Acerto	Ap/Ap[1]	Re/Re[2]	Erro	Ap/Re[3]	Re/Ap[4]
Teste065	95,90	97,32	91,79	4,10	2,68	8,21
Teste066	96,07	97,48	91,98	3,93	2,52	8,02
Teste067	96,06	97,43	92,10	3,94	2,57	7,90
Teste068	95,68	98,88	86,41	4,32	1,12	13,59
Teste069	95,61	98,64	86,84	4,39	1,36	13,16
Teste070	95,81	96,80	92,95	4,19	3,20	7,05

[1] Percentual de aprovados que o teste classificou como aprovados.
[2] Percentual de reprovados que o teste classificou como reprovados.
[3] Percentual de aprovados que o teste classificou como reprovados.
[4] Percentual de reprovados que o teste classificou como aprovados.

Os principais critérios para avaliar um teste de estimação que utiliza uma quantidade fixa de itens como critério de parada são o EPM (menor-melhor), a RQEQM (menor-melhor), o DEM (mais próximo de zero - melhor) e a CL (maior-melhor), conforme discutido na Subetapa 5.2. Como mencionado na Seção 5.5.1, dentro do contexto do DETRAN-SC, a situação que parece mais grave seria aprovar um candidato que deveria ter sido reprovado, ou seja, habilitar um motorista que ainda não está preparado para dirigir. Nesse caso, como foi visto, todos os testes tiveram uma taxa baixa na aprovação de examinandos que deveriam ter sido reprovados, variando entre 7,05% e 13,59%, sendo que os testes com método de estimação bayesiano foram os que tiveram as piores taxas. Quanto aos critérios de avaliação, os testes que utilizaram os métodos bayesianos tiveram desempenho superior aos demais, em relação ao EPM e a CL, porém o teste que utilizou o método MVP teve desempenho superior aos demais, em relação a RQEQM e ao DEM. Os testes bayesianos, além de aprovarem mais examinandos que deveriam ter sido reprovados, tiveram um DEM muito alto.

Portanto, para o teste com objetivo de estimação, utilizando-se a aplicação de 40 de itens como critério de parada, considerando que as habilidades provêm de uma distribuição Uniforme no intervalo [-3, 3], optou-se pelo Teste070 que possui a configuração descrita no Quadro 12.

ELEMENTOS DO ALGORITMO	OPÇÕES SELECIONADAS
Modelo de Resposta ao item	Modelo Logístico Unidimensional de Três Parâmetros – ML3
Nível de habilidade Inicial	Mediano igual a zero.
Método de Seleção dos itens	Máxima Informação
Método de Estimação da habilidade	MVP.
Restrições	Taxa de exposição = 0,20
Critério de parada	Número fixo de 20 itens.

Quadro 12: Algoritmo Selecionado para o TAI

5.5.5. Teste de Estimação para População Normal com Erro Padrão Fixo

O quinto teste analisado será o teste com objetivo de estimação considerando que as habilidades provêm de uma distribuição Normal Padrão.

Baseado nos resultados obtidos na Seção 5.5.1, algumas configurações expostas no Quadro 1 não serão testadas porque não elas não tiveram um desempenho diferenciado nas análises daquela Seção. Dessa forma, o número de algoritmos a serem testados nessa situação, considerando todas as combinações possíveis dentre as opções selecionadas, foi reduzido para 6, conforme o Quadro 8.

Embora nesse caso o objetivo seja estimação, na análise também será verificado o desempenho dos algoritmos para a classificação do candidato (aprovado/reprovado). Dessa forma, adotou-se o mesmo valor utilizado na Seção 5.5.1 para o ponto de corte na escala de proficiência, ou seja, -0,663.

Conforme disposto no Quadro 8, a quantidade de combinações possíveis nessa situação é 6, onde as diferenças estão focadas apenas no método de estimação da habilidade, conforme mostra o Quadro 13.

A primeira simulação foi feita com o Teste077, que utiliza a mesma configuração do Teste083, porém com um banco com 329 itens, como mencionado na Seção 5.5.1 e, logo, observou-se um problema que tornaria as demais simulações inviáveis. Os resultados obtidos mostraram que apenas em 19,84% dos casos o critério de parada foi alcançado (precisão de 0,2), sendo aplicados, em média, 76,9 itens por

teste. Em 80,34% dos casos, todos os itens disponíveis no banco, respeitando a taxa de exposição, foram utilizados, sendo aplicados, em média, 79,1 itens por teste. A quantidade de itens aplicados variou entre 44 e 108. Esses resultados não foram considerados adequados (Subetapa 5.2) porque a quantidade de itens aplicados foi maior do que o teste convencional (40 itens), em todas as situações. Como nenhum item será elaborado e apenas em 19,84% dos casos o critério de parada foi alcançado (mesmo assim, a quantidade de itens aplicados foi muito alta), e a solução não será a de adicionar itens (vide Subetapa 5.3), optou-se por alterar o algoritmo. Para verificar qual a quantidade de itens necessária naquelas situações onde o teste terminou devido a utilização de todos os itens disponíveis no banco, optou-se por remover a taxa de exposição e modificar o banco de itens, substituindo o Banco 1 pelo Banco 3. Os resultados mostraram que em 81,22% dos casos o critério de parada foi alcançado, sendo aplicados, em média, 86,4 itens por teste. Em 18,78% dos casos, todos os itens disponíveis no Banco 3 (303) foram utilizados, sendo aplicados.

TESTE	ESTIMAÇÂO DE θ
Teste083	MV c/ passo = 1.
Teste084	MV c/ estimação inicial EAP
Teste085	MV c/estimação inicial MAP
Teste086	Bayesiano EAP
Teste087	Bayesiano MAP
Teste088	MVP

Quadro 13: Algoritmos testados

Na alteração do algoritmo, como a taxa de exposição dos itens é essencial nessa aplicação (itens sigilosos), optou-se por modificar o critério de parada do teste. Para identificar um novo nível de precisão, deve-se observar o comportamento do erro padrão das habilidades do banco de itens que foi apresentado no gráfico (linha tracejada) da Figura 25, na Seção 5.3. Observando-se o gráfico da Figura 25, verifica-se que o nível de precisão previamente adorado (0,2) realmente não seria alcançado em todos os níveis de proficiência nem que todos os itens dos bancos fossem aplicados. Nesse caso, não é possível ser tão rigoroso no teste. Seria mais conveniente, por exemplo, adorar um erro padrão igual a 0,5, embora esse valor seja considerado alto para uma escala (0,1). Dessa forma, optou-se por utilizar um erro padrão igual a 0,5, o qual substituirá o erro padrão igual a 0,2 que era o critério de parada anteriormente selecionado no Quadro 8.

A Tabela 17 apresenta o desempenho dos 6 testes analisados (Quadro 13), segundo os critérios EPM, RQEQM, DEM, EF e CL, enquanto que a Tabela 18 apresenta o desempenho dos testes em relação ao acerto da classificação (Aprovado/Reprovado) segundo as habilidades simuladas consideradas verdadeiras.

Os resultados apresentados na Tabela 17 e na Tabela 18 mostram que nenhum desses testes pode ser considerado totalmente inadequado, segundo os indicadores analisados, porém eles se diferenciam entre si em alguns critérios. Observa-se na Tabela 17 que os testes que utilizaram os métodos bayesianos tiveram desempenho superior aos demais, em relação ao EPM, a RQEQM e, principalmente, a EF, sendo que o método MAP foi ligeiramente superior ao método EAP em relação ao EPM e a EF e o método EAP foi ligeiramente superior ao método MAP em relação a RQEQM, ao DEM e a CL. Dessa vez, o método MVP foi o que teve o pior desempenho em relação ao DEM. O método MV puro teve desempenho inferior aos demais, em relação ao EPM, a RQEQM a CL.

Tabela 17 Desempenho dos Testes

TESTE	EPM	RQEQM	DEM	EF	CL
Teste083	0,5644	0,6339	0,0258	18,3611	0,8366
Teste084	0,5159	0,5104	-0,0110	18,4036	0,8688
Teste085	0,5162	0,5024	-0,0076	18,4452	0,8713
Teste086	0,4933	0,4725	0,0120	13,7029	0,8532
Teste087	0,4917	0,4926	-0,0136	12,2454	0,8377
Teste088	0,5154	0,5536	-0,0978	16,3552	0,8483

A Tabela 18 mostra que o percentual de acerto foi muito parecido entre os testes, ficando em torno de 86%. O maior percentual de acerto foi doTeste084 (87,10%), porém o maior percentual de aprovados que o teste classificou como aprovados foi do Teste086 (93,94%) e o maior percentual de reprovados que o teste classificou como reprovados foi do Teste088 (81,23%). Todos os testes tiveram uma taxa alta na aprovação de examinandos que deveriam ter sido reprovados, variando entre 18,77% e 33,95%. Os testes com método de estimação bayesiano reprovaram menos examinandos que deveriam ter sido aprovados, porém aprovaram mais examinandos que deveriam ter sido reprovados. Nos dados simulados, a taxa de reprovação foi de 25,57%. A taxa de reprovação entre os testes variou entre 21,56% e 30,82, sendo que a mais próxima da verdadeira foi do Teste084 (28,01%). Os testes com

métodos bayesianos tiveram taxa de reprovação abaixo de 25,57%, enquanto que os testes com métodos MV e MVP tiveram a taxa de reprovação acima desse valor.

Tabela 18 Percentual de acerto e erro na classificação

TESTE	Acerto	Ap/Ap[1]	Re/Re[2]	Erro	Ap/Re[3]	Re/Ap[4]
Teste083	86,81	89,39	79,31	13,19	10,61	20,69
Teste084	87,10	89,70	79,55	12,90	10,30	20,45
Teste085	86,67	89,10	79,59	13,33	10,90	20,41
Teste086	86,97	93,94	66,68	13,03	6,06	33,32
Teste087	86,23	93,16	66,05	13,77	6,84	33,95
Teste088	85,15	86,50	81,23	14,85	13,50	18,77

[1] Percentual de aprovados que o teste classificou como aprovados.
[2] Percentual de reprovados que o teste classificou como reprovados.
[3] Percentual de aprovados que o teste classificou como reprovados.
[4] Percentual de reprovados que o teste classificou como aprovados.

Os principais critérios para avaliar um teste de estimação que utiliza um nível de precisão como critério de parada são a EF (menor-melhor), a RQEQM (menor-melhor), o DEM (mais próximo de zero - melhor) e a CL (maior-melhor), conforme visto na Subetapa 5.2. Como mencionado na Seção 5.5.1, dentro do contexto do DETRAN-SC, a situação que parece mais grave seria aprovar um candidato que deveria ter sido reprovado, ou seja, habilitar um motorista que ainda não está preparado para dirigir. Nesse caso, como foi visto, todos os testes tiveram uma taxa alta na aprovação de examinandos que deveriam ter sido reprovados, variando entre 18,77% e 33,95%, sendo que os testes com método de estimação bayesiano foram os que aprovaram mais examinandos que deveriam ter sido reprovados. Por outro lado, quanto aos critérios de avaliação, os testes com método de estimação bayesiano tiveram desempenho melhor em relação ao EPM, a RQEQM, a EF. Nota-se que, embora tenha sido utilizado um EP igual a 0,50 como critério de parada, alguns EPM ficaram acima de 0,50 (Tabela 17), influenciados pelos casos onde foram utilizados o limite máximo de 40 itens no teste. Os métodos bayesianos apresentaram melhor desempenho em três dos cinco indicadores e, ainda assim, o desempenho do indicador DEM não foi tão ruim, porém uma taxa mais alta que os demais na aprovação de examinandos que deveriam ter sido reprovados. O teste com método MV apresentou os piores desempenhos em relação ao EPM, a RQEQM e a CL. O método MVP, embora tenha apresentado o pior desempenho quanto ao DEM, teve bom desempenho em relação

ao EPM e apresentou a menor taxa de aprovação de examinandos que deveriam ter sido reprovados.

Portanto, para o teste com objetivo de estimação, utilizando-se um erro padrão de 0,50 (com aplicação máxima de 40 itens) como critério de parada, considerando que as habilidades provêm de uma distribuição Normal Padrão, optou-se pelo Teste088 que possui a configuração descrita no Quadro 14.

ELEMENTOS DO ALGORITMO	OPÇÕES SELECIONADAS
Modelo de Resposta ao item	Modelo Logístico Unidimensional de Três Parâmetros – ML3
Nível de habilidade Inicial	Mediano igual a zero.
Método de Seleção dos itens	Máxima Informação
Método de Estimação da habilidade	MVP.
Restrições	Taxa de exposição = 0,20
Critério de parada	Erro Padrão menor ou igual a 0,50 e máximo de 40 itens.

Quadro 14: Algoritmo Selecionado para o TAI

Em 84,80% dos casos, o teste terminou por alcançar o erro padrão determinado de 0,50 (foram aplicados menos de 40 itens). A Tabela 19 apresenta a quantidade de itens aplicados nos testes segundo a classificação, enquanto que a Tabela 20 apresenta o percentual associado e a habilidade média estimada (HME). Observa-se na Tabela 19, por exemplo, que apenas quatro itens foram suficientes para classificar 80 testes simulados, onde a habilidade média estimada foi de –1,88, sendo que em 95,00% dos casos (Tabela 20) o teste classificou corretamente. Observa-se que quando houve uma quantidade baixa de itens aplicados, especialmente entre 7 e 13 itens, o teste classificou muitos examinandos como aprovados sendo que eles deveriam ter sido reprovados. Observa-se na Tabela 20 que, em todas as situações, onde foram aplicados mais de 22 itens, o algoritmo classificou todos os resultados corretamente (examinando aprovados que deveriam ter sido aprovados). Nota-se que em 15,20% das situações, o limite máximo de 40 itens fixado no teste foi utilizado. A quantidade de itens mais aplicada, nos casos em que o erro padrão foi alcançado, foi de sete itens (8,63% dos casos), onde em 66,74% dos casos o teste classificou corretamente, porém aprovou 8,7% de examinandos que deveriam ter sido reprovados. Ainda na Tabela 20, observa-se que a medida que a habilidade média estimada foi aumentando, a quantidade de itens aplicados também foi aumentando.

Tabela 19 Quantidade de itens aplicados

Itens	Acerto	Ap/Ap	Re/Re	Erro	Ap/Re	Re/Ap	Total	%
4	76	0	76	4	4	0	80	0,80
5	414	0	414	103	103	0	517	5,17
6	602	32	570	248	239	9	850	8,50
7	576	114	462	287	243	44	863	8,63
8	498	247	251	248	178	70	746	7,46
9	505	378	127	191	105	86	696	6,96
10	503	427	76	141	62	79	644	6,44
11	494	447	47	106	42	64	600	6,00
12	467	443	24	66	12	54	533	5,33
13	396	382	14	45	11	34	441	4,41
14	337	334	3	20	3	17	357	3,57
15	277	271	6	15	2	13	292	2,92
16	252	249	3	3	0	3	255	2,55
17	225	222	3	3	1	2	228	2,28
18	162	161	1	1	0	1	163	1,63
19	127	127	0	1	0	1	128	1,28
20	128	128	0	2	0	2	130	1,30
21	114	114	0	1	0	1	115	1,15
22	100	100	0	0	0	0	100	1,00
23	86	86	0	0	0	0	86	0,86
24	68	68	0	0	0	0	68	0,68
25	77	77	0	0	0	0	77	0,77
26	48	48	0	0	0	0	48	0,48
27	70	70	0	0	0	0	70	0,70
28	57	57	0	0	0	0	57	0,57
29	49	49	0	0	0	0	49	0,49
30	38	38	0	0	0	0	38	0,38
31	38	38	0	0	0	0	38	0,38
32	32	32	0	0	0	0	32	0,32
33	34	34	0	0	0	0	34	0,34
34	43	43	0	0	0	0	43	0,43
35	20	20	0	0	0	0	20	0,20
36	22	22	0	0	0	0	22	0,22
37	22	22	0	0	0	0	22	0,22
38	14	14	0	0	0	0	14	0,14
39	24	24	0	0	0	0	24	0,24
40	1520	1520	0	0	0	0	1520	15,20
Total	8515	6438	2077	1485	1005	480	10000	100,00

Tabela 20 Percentual segundo a classificação

Itens	Acerto	Ap/Ap	Re/Re	Erro	Ap/Re	Re/Ap	HME
4	95,00	0,00	100,00	5,00	100,00	0,00	-1,88
5	80,08	0,00	100,00	19,92	100,00	0,00	-1,65
6	70,82	11,81	98,45	29,18	88,19	1,55	-1,31
7	66,74	31,93	91,30	33,26	68,07	8,70	-1,03
8	66,76	58,12	78,19	33,24	41,88	21,81	-0,77
9	72,56	78,26	59,62	27,44	21,74	40,38	-0,53
10	78,11	87,32	49,03	21,89	12,68	50,97	-0,41
11	82,33	91,41	42,34	17,67	8,59	57,66	-0,26
12	87,62	97,36	30,77	12,38	2,64	69,23	-0,15
13	89,80	97,20	29,17	10,20	2,80	70,83	-0,04
14	94,40	99,11	15,00	5,60	0,89	85,00	0,09
15	94,86	99,27	31,58	5,14	0,73	68,42	0,16
16	98,82	100,00	50,00	1,18	0,00	50,00	0,28
17	98,68	99,55	60,00	1,32	0,45	40,00	0,25
18	99,39	100,00	50,00	0,61	0,00	50,00	0,37
19	99,22	100,00	0,00	0,78	0,00	100,00	0,39
20	98,46	100,00	0,00	1,54	0,00	100,00	0,46
21	99,13	100,00	0,00	0,87	0,00	100,00	0,50
22	100,00	100,00	–	0,00	0,00	–	0,56
23	100,00	100,00	–	0,00	0,00	–	0,54
24	100,00	100,00	–	0,00	0,00	–	0,57
25	100,00	100,00	–	0,00	0,00	–	0,66
26	100,00	100,00	–	0,00	0,00	–	0,66
27	100,00	100,00	–	0,00	0,00	–	0,69
28	100,00	100,00	–	0,00	0,00	–	0,68
29	100,00	100,00	–	0,00	0,00	–	0,68
30	100,00	100,00	–	0,00	0,00	–	0,68
31	100,00	100,00	–	0,00	0,00	–	0,71
32	100,00	100,00	–	0,00	0,00	–	0,70
33	100,00	100,00	–	0,00	0,00	–	0,79
34	100,00	100,00	–	0,00	0,00	–	0,77
35	100,00	100,00	–	0,00	0,00	–	0,94
36	100,00	100,00	–	0,00	0,00	–	0,83
37	100,00	100,00	–	0,00	0,00	–	0,81
38	100,00	100,00	–	0,00	0,00	–	0,97
39	100,00	100,00	–	0,00	0,00	–	0,81
40	100,00	100,00	–	0,00	0,00	–	1,51
Total	85,15	86,50	81,23	14,85	13,50	18,77	-0,12

5.5.6. Teste de Estimação para População Uniforme com Erro Padrão Fixo

O sexto teste analisado foi o teste com objetivo de estimação considerando que as habilidades provêm de uma distribuição Uniforme no intervalo [-3, 3].

Baseado nos resultados obtidos na Seção 5.5.1, algumas configurações expostas no Quadro 1 não foram testadas porque elas não tiveram um desempenho diferenciado nas análises daquela Seção. Dessa forma, o número de algoritmos a serem testados nessa situação, considerando todas as combinações possíveis dentre as opções selecionadas, foi reduzido para 6, conforme o Quadro 3. Além disso, o erro padrão foi alterado para 0,50, conforme Seção 5.5.5.

Embora nesse caso o objetivo seja estimação, na análise também será verificado o desempenho dos algoritmos para a classificação do candidato (aprovado/reprovado). Dessa forma, adotou-se o mesmo valor utilizado na Seção 5.5.2 para o ponto de corte na escala de proficiência, ou seja, -1,479.

Conforme disposto no Quadro 8, a quantidade de combinações possíveis nessa situação é 6, onde as diferenças estão focadas no banco de itens a ser utilizado e no método de estimação da habilidade, conforme mostra o Quadro 15.

TESTE	ESTIMAÇÂO DE θ
Teste101	MV c/ passo = 1.
Teste102	MV c/ estimação inicial EAP
Teste103	MV c/estimação inicial MAP
Teste104	Bayesiano EAP
Teste105	Bayesiano MAP
Teste106	MVP

Quadro 15: Algoritmos testados

A Tabela 21 apresenta o desempenho dos 6 testes analisados (Quadro 15), segundo os critérios EPM, RQEQM, DEM, EF e CL, enquanto que a Tabela 22 apresenta o desempenho dos testes em relação ao acerto da classificação (Aprovado/Reprovado), segundo as habilidades simuladas consideradas verdadeiras. Os resultados apresentados na Tabela 21 e na Tabela 22 mostram que nenhum desses testes pode ser considerado totalmente inadequado, segundo os indicadores analisados. Observa-se na Tabela 21 que os testes que utilizam os métodos bayesianos tiveram desempenho melhor do que os demais em relação ao EPM, ao DEM e a EF, porém tiveram

desempenho pior em relação a RQEQM. Em relação a EPM e a EF, o método bayesiano MAP teve desempenho levemente superior ao método bayesiano EAP, porém em relação a RQEQM, ao DEM e a CL, o método EAP mostrou-se superior. Quanto a RQEQM, o melhor desempenho foi dos testes que utilizam o método MV com estimação inicial bayesiana. O teste que utilizou MV (Teste101) teve o pior desempenho em relação a todos os critérios.

Tabela 21 Desempenho dos Testes

TESTE	EPM	RQEQM	DEM	EF	CL
Teste101	0,7977	0,7410	0,1230	21,4358	0,9331
Teste102	0,5476	0,5376	-0,0888	21,3113	0,9524
Teste103	0,5495	0,5539	-0,0821	21,2446	0,9496
Teste104	0,5036	0,6501	-0,0550	17,5590	0,9455
Teste105	0,5008	0,7132	-0,0659	16,1901	0,9403
Teste106	0,5907	0,5747	-0,0855	20,0233	0,9456

A Tabela 22 mostra que o percentual de acerto foi muito parecido entre os testes com os métodos MV e MVP e superior aos métodos bayesianos. O maior percentual de acerto foi do Teste101 (93,45%), porém o maior percentual de aprovados que o teste classificou como aprovados foi do Teste0104 (98,62) e o maior percentual de reprovados que o teste classificou como reprovados foi também do Teste101 (87,91%). Os testes com método de estimação bayesiano reprovaram menos examinandos que deveriam ter sido aprovados em relação aos demais testes, porém aprovaram mais examinandos que deveriam ter sido reprovados. Os testes com métodos MV e MVP foram os que aprovaram menos examinandos que deveriam ter sido reprovados. Nos dados simulados, a taxa de reprovação foi de 0,2597. A taxa de reprovação entre os testes variou entre 16,71% e 26,52%, sendo que a mais próxima da verdadeira foi do Teste101 (26,24%). Os testes com métodos bayesianos tiveram taxa de reprovação bem abaixo de 25,57%, enquanto que os testes com métodos MV e MVP tiveram a taxa de reprovação bem próxima desse valor.

Tabela 22 Percentual de acerto e erro na classificação

TESTE	Acerto	Ap/Ap[1]	Re/Re[2]	Erro	Ap/Re[3]	Re/Ap[4]
Teste101	93,45	95,39	87,91	6,55	4,61	12,09
Teste102	93,21	95,61	86,37	6,79	4,39	13,63
Teste103	93,27	95,91	85,75	6,73	4,09	14,25
Teste104	90,27	98,62	66,46	9,73	1,38	33,54
Teste105	88,58	98,54	60,18	11,42	1,46	39,82
Teste106	92,85	94,80	87,29	7,15	5,20	12,71

[1] Percentual de aprovados que o teste classificou como aprovados.
[2] Percentual de reprovados que o teste classificou como reprovados.
[3] Percentual de aprovados que o teste classificou como reprovados.
[4] Percentual de reprovados que o teste classificou como aprovados.

Os principais critérios para avaliar um teste de estimação que utiliza um nível de precisão como critério de parada são a EF (menor-melhor), a RQEQM (menor-melhor), o DEM (mais próximo de zero - melhor) e a CL (maior-melhor), conforme visto na Subetapa 5.2. Como mencionado na Seção 5.5.1, dentro do contexto do DETRAN-SC, a situação que parece mais grave seria aprovar um candidato que deveria ter sido reprovado, ou seja, habilitar um motorista que ainda não está preparado para dirigir. Nesse caso, como foi visto, os testes que utilizam os métodos MV e MVP tiveram uma taxa baixa na aprovação de examinandos que deveriam ter sido reprovados, variando entre 12,09% e 14,25%, e os testes que utilizam os métodos de estimação bayesiano tiveram uma taxa alta na aprovação de examinandos que deveriam ter sido reprovados, sendo 33,54% para o teste que utiliza o método EAP e 39,82% para o teste que utiliza o método MAP. Quanto aos critérios de avaliação, os testes que utilizaram os métodos bayesianos tiveram desempenho superior aos demais, em relação ao EPM, ao DEM e a EF, porém os testes que utilizaram os o método MV com estimação inicial bayesiana tiveram desempenho superior aos demais, em relação a RQEQM. Dentre os métodos MV e MVP, teste que utiliza o método MVP apresentou o melhor desempenho em relação a EF, além de apresentar o menor percentual na aprovação de examinandos que deveriam ter sido reprovados.

Portanto, para o teste com objetivo de estimação, utilizando-se um erro padrão de 0,50 (com aplicação máxima de 40 itens) como critério de parada, considerando que as habilidades provêm de uma distribuição Uniforme no intervalo [-3, 3], optou-se pelo Teste106 que possui a configuração descrita no Quadro 16.

ELEMENTOS DO ALGORITMO	OPÇÕES SELECIONADAS
Modelo de Resposta ao item	Modelo Logístico Unidimensional de Três Parâmetros – ML3
Banco de Itens	Banco 1
Nível de habilidade Inicial	Mediano igual a zero.
Método de Seleção dos itens	Máxima Informação
Método de Estimação da habilidade	MVP.
Restrições	Taxa de exposição = 0,20
Critério de parada	Erro Padrão menor ou igual a 0,50 e máximo de 40 itens.

Quadro 16: Algoritmo Selecionado para o TAI

Em 68,23% dos casos, o teste terminou por alcançar o erro padrão determinado de 0,50. A Tabela 23 apresenta a quantidade de itens aplicados nos testes segundo a classificação, enquanto que a Tabela 24 apresenta o percentual associado e a habilidade média estimada (HME). Observa-se na Tabela 23, por exemplo, que apenas quatro itens foram suficientes para classificar 174 testes simulados, onde a habilidade média estimada foi de -2,04, sendo que em 82,18% dos casos (Tabela 24) o teste classificou corretamente. A Tabela 24 mostra que, quando houve uma quantidade baixa de itens aplicados, especialmente entre 5 e 10 itens, o teste classificou muitos examinandos como aprovados sendo que eles deveriam ter sido reprovados. Observa-se na Tabela 20 que, em todas as situações, onde foram aplicados mais de 15 itens, o algoritmo classificou todos os resultados corretamente (examinando aprovados que deveriam ter sido aprovados). Nota-se que em 31,77% das situações, o limite máximo de 40 itens fixado no teste foi utilizado. A quantidade de itens mais aplicada, nos casos em que o erro padrão foi alcançado, foi de seis itens (10,68% dos casos), onde em 80,99% dos casos o teste classificou corretamente, porém aprovou 14,71% de examinandos que deveriam ter sido reprovados. Ainda na Tabela 24, observa-se que a medida que a habilidade média estimada foi aumentando, a quantidade de itens aplicados também foi aumentando.

Tabela 23 Quantidade de itens aplicados

Itens	Acerto	Ap/Ap	Re/Re	Erro	Ap/Re	Re/Ap	Total	%
4	143	1	142	31	30	1	174	1,74
5	688	76	612	190	156	34	878	8,78
6	865	291	574	203	104	99	1068	10,68
7	661	359	302	142	54	88	803	8,03
8	531	335	196	78	21	57	609	6,09
9	443	316	127	31	12	19	474	4,74
10	401	327	74	22	6	16	423	4,23
11	338	275	63	9	1	8	347	3,47
12	283	245	38	5	1	4	288	2,88
13	233	208	25	3	0	3	236	2,36
14	185	164	21	1	0	1	186	1,86
15	171	157	14	0	0	0	171	1,71
16	124	113	11	0	0	0	124	1,24
17	119	108	11	0	0	0	119	1,19
18	85	73	12	0	0	0	85	0,85
19	109	103	6	0	0	0	109	1,09
20	88	79	9	0	0	0	88	0,88
21	68	64	4	0	0	0	68	0,68
22	48	46	2	0	0	0	48	0,48
23	63	59	4	0	0	0	63	0,63
24	43	40	3	0	0	0	43	0,43
25	50	48	2	0	0	0	50	0,50
26	41	40	1	0	0	0	41	0,41
27	39	38	1	0	0	0	39	0,39
28	36	35	1	0	0	0	36	0,36
29	38	38	0	0	0	0	38	0,38
30	23	23	0	0	0	0	23	0,23
31	27	26	1	0	0	0	27	0,27
32	30	29	1	0	0	0	30	0,30
33	16	16	0	0	0	0	16	0,16
34	24	24	0	0	0	0	24	0,24
35	22	21	1	0	0	0	22	0,22
36	17	17	0	0	0	0	17	0,17
37	15	15	0	0	0	0	15	0,15
38	19	18	1	0	0	0	19	0,19
39	22	22	0	0	0	0	22	0,22
40	3177	3169	8	0	0	0	3177	31,77
Total	9285	7018	2267	715	385	330	10000	100,00

Tabela 24 Percentual segundo a classificação

Itens	Acerto	Ap/Ap	Re/Re	Erro	Ap/Re	Re/Ap	HME
4	82,18	3,23	99,30	17,82	96,77	0,70	-2,04
5	78,36	32,76	94,74	21,64	67,24	5,26	-1,94
6	80,99	73,67	85,29	19,01	26,33	14,71	-1,74
7	82,32	86,92	77,44	17,68	13,08	22,56	-1,51
8	87,19	94,10	77,47	12,81	5,90	22,53	-1,32
9	93,46	96,34	86,99	6,54	3,66	13,01	-1,11
10	94,80	98,20	82,22	5,20	1,80	17,78	-0,82
11	97,41	99,64	88,73	2,59	0,36	11,27	-0,72
12	98,26	99,59	90,48	1,74	0,41	9,52	-0,52
13	98,73	100,00	89,29	1,27	0,00	10,71	-0,35
14	99,46	100,00	95,45	0,54	0,00	4,55	-0,27
15	100,00	100,00	100,00	0,00	0,00	0,00	-0,11
16	100,00	100,00	100,00	0,00	0,00	0,00	-0,02
17	100,00	100,00	100,00	0,00	0,00	0,00	0,01
18	100,00	100,00	100,00	0,00	0,00	0,00	-0,10
19	100,00	100,00	100,00	0,00	0,00	0,00	0,26
20	100,00	100,00	100,00	0,00	0,00	0,00	0,13
21	100,00	100,00	100,00	0,00	0,00	0,00	0,35
22	100,00	100,00	100,00	0,00	0,00	0,00	0,40
23	100,00	100,00	100,00	0,00	0,00	0,00	0,35
24	100,00	100,00	100,00	0,00	0,00	0,00	0,33
25	100,00	100,00	100,00	0,00	0,00	0,00	0,54
26	100,00	100,00	100,00	0,00	0,00	0,00	0,56
27	100,00	100,00	100,00	0,00	0,00	0,00	0,56
28	100,00	100,00	100,00	0,00	0,00	0,00	0,56
29	100,00	100,00	–	0,00	0,00	–	0,74
30	100,00	100,00	–	0,00	0,00	–	0,66
31	100,00	100,00	100,00	0,00	0,00	0,00	0,52
32	100,00	100,00	100,00	0,00	0,00	0,00	0,63
33	100,00	100,00	–	0,00	0,00	–	0,88
34	100,00	100,00	–	0,00	0,00	–	0,78
35	100,00	100,00	100,00	0,00	0,00	0,00	0,62
36	100,00	100,00	–	0,00	0,00	–	0,88
37	100,00	100,00	–	0,00	0,00	–	0,85
38	100,00	100,00	100,00	0,00	0,00	0,00	0,70
39	100,00	100,00	–	0,00	0,00	–	1,04
40	100,00	100,00	100,00	0,00	0,00	0,00	1,91
Total	92,85	94,80	87,29	7,15	5,20	12,71	-0,09

5.5.7. Resumo dos Testes Selecionados

A Tabela 25 apresenta o desempenho dos 6 testes selecionados, segundo os critérios EPM, RQEQM, DEM, EF e CL, enquanto que a Tabela 26 apresenta o desempenho dos testes em relação ao acerto da classificação (Aprovado/Reprovado) segundo as habilidades simuladas consideradas verdadeiras.

Tabela 25 Desempenho dos Testes Selecionados

TESTE	EPM	RQEQM	DEM	EF	CL
Teste016	0,5563	0,6498	0,0606	19,9361	0,8242
Teste034	0,7114	0,9634	-0,2994	9,6250	0,8509
Teste052	0,4577	0,4685	-0,0110	20,0000	0,8847
Teste070	0,5322	0,5143	-0,0661	20,0000	0,9562
Teste088	0,5154	0,5536	-0,0978	16,3552	0,8483
Teste106	0,5907	0,5747	-0,0855	20,0233	0,9456

Tabela 26 Percentual de acerto e erro na classificação

TESTE	Acerto	Ap/Ap[1]	Re/Re[2]	Erro	Ap/Re[3]	Re/Ap[4]
Teste016	91,93	93,76	86,37	8,07	6,24	13,63
Teste034	95,78	97,99	89,37	4,22	2,01	10,63
Teste052	90,44	92,11	85,36	9,56	7,89	14,64
Teste070	95,81	96,80	92,95	4,19	3,20	7,05
Teste088	85,15	86,50	81,23	14,85	13,50	18,77
Teste106	92,85	94,80	87,29	7,15	5,20	12,71

[1] Percentual de aprovados que o teste classificou como aprovados.
[2] Percentual de reprovados que o teste classificou como reprovados.
[3] Percentual de aprovados que o teste classificou como reprovados.
[4] Percentual de reprovados que o teste classificou como aprovados.

Todos os testes foram selecionados, basicamente, por apresentar, em comparação aos demais, o menor percentual na aprovação de examinandos que deveriam ter sido reprovados, que é uma situação crítica para o caso do DETRAN-SC, e por apresentarem resultados aceitáveis em relação aos demais critérios. Observa-se também que todos os testes selecionados utilizaram o método de estimação MVP. Nota-se que, se esse fator crítico não fosse considerado nos testes com objetivo de estimação, os testes escolhidos seriam os que utilizam o método bayesiano EAP ou MAP, pois apresentavam geralmente bons desempenhos nos critérios EEPM, RQEQM, DEM, EF e CL.

Os testes 052 e 088 podem ser comparados entre si porque foram construídos com o mesmo objetivo e com a mesma suposição referente a

distribuição do traço latente na população, assim como os testes 070 e 106. Os demais testes não podem ser comparados entre si porque foram construídos com objetivos diferentes e com suposições diferentes referente a distribuição do traço latente na população.

O Teste052 apresenta melhor desempenho do que o Teste088 em relação a todos os critérios da Tabela 25, exceto em relação à EF. Obviamente, seu desempenho é melhor por aplicar uma quantidade média maior de itens.

O Teste070 apresenta melhor desempenho do que o Teste088 em relação a todos os critérios da Tabela 25, inclusive em relação à EF, além de apresentar os melhores percentuais de acerto e erro da Tabela 26.

Observou-se que, nos testes de classificação que utilizaram IC como critério de parada, a medida que a habilidade média estimada foi diminuindo, a quantidade de itens aplicados foi aumentando, enquanto que nos testes de estimação de habilidade que utilizaram EP fixo como critério de parada, a medida que a habilidade média estimada foi aumentando, a quantidade de itens aplicados também foi aumentando.

Em todas as situações estudadas, foi encontrado pelo menos um teste que pode ser considerado adequado (Subetapa 5.2). Assim sendo, qualquer um desses testes, considerando, contudo, o objetivo e a população, pode ser implementado.

5.6. ETAPA 6: IMPLEMENTAÇÃO

A sexta etapa compreende a implementação prática do teste, onde deve-se considerar vários recursos materiais além de uma equipe multidisciplinar. Essa Etapa não foi desenvolvida nesse estudo, mas foram feitas algumas considerações.

No caso do DETRAN-SC, todos os aspectos básicos mencionados na Subetapa 6.1 deveriam ser desenvolvidos. Primeiramente, o TAI selecionado deveria ser implementado num software, o qual, consequentemente seria implementado nos computadores utilizados para o teste. Para isso, também seriam necessários uma equipe multidisciplinar e um provedor para armazenar os dados coletados.

Como os itens são sigilosos (Subetapa 6.2), deve-se investir na segurança dos dados (informações dos itens, respostas e dados dos examinados). No caso do DETRAN-SC, seriam necessários locais onde as pessoas poderiam ter acesso para realizar a prova. Possivelmente a estrutura existente do teste tradicional poderia ser utilizada (por

exemplo, instalações, salas, fiscais, etc.) e adaptava para a versão adaptativa do teste.

5.7. ETAPA 7: APLICAÇÃO

A sétima etapa compreende a aplicação efetiva do teste, fornecendo um retorno ao examinando sobre o seu desempenho, e com a coleta de dados para constituir um banco com as respostas dos examinandos para posterior a análise e manutenção do teste. Essa Etapa não foi desenvolvida nesse estudo, mas foram feitas algumas considerações.

No DETRAN-SC, a aplicação do teste poderia se dar conforme a Subetapa 7.1; Primeiramente, com a identificação do examinando, que poderá possibilitar um controle no sentido de verificar quantas vezes ele fez o teste e quais os itens que foram aplicados para que eles não sejam reaplicados novamente, dentro do possível. Em seguida, o teste seria aplicado e, ao final, forneceria um *feedback* para o examinando, com a sua nota ou classificação num relatório resumido do seu desempenho, identificando os conteúdos que ele domina e os que ele não domina. Por fim, os dados obtidos (identificação e respostas) seriam armazenado no servidor.

A questão da atualização da taxa de exposição do item poderia ser definida se seria feita após a finalização de cada teste ou na manutenção (Etapa 8).

A finalização da aplicação dos testes poderia ser, por exemplo, ao final de cada dia ou uma vez por semana ou por mês, conforme a demanda de examinandos e a quantidade de dados coletada.

Após a finalização dos testes (Subetapa 7.2), poderia se verificar se foram identificados problemas nas aplicações práticas ou na análise prévia dos dados coletados. Se houver necessidade de modificar ou atualizar algo, procede-se para a manutenção, caso contrário, os teste podem continuar a serem aplicados.

5.8. ETAPA 8: MANUTENÇÃO

A oitava etapa compreende a manutenção do teste, que deverá ser periódica e, se necessário, poderá resultar em alterações no teste, evidenciadas por meio da análise do desempenho do TAI com os dados reais coletados. Essa Etapa não foi desenvolvida nesse estudo, mas foram feitas algumas considerações.

No caso do DETRAN-SC, várias manutenções poderiam ser realizadas periodicamente: atualização da taxa de exposição dos itens (Subetapa 8.1), calibração de itens novos (Subetapa 8.2), análise dos

dados coletados para verificar se algum item existente no banco precisa ser excluído ou modificado ou se itens novos precisam ser adicionados (Subetapa 8.3).

Eventualmente, se necessário, poderia ser feita alguma modificação no algoritmo do teste (Subetapa 8.4). Após a manutenção, o teste pode continuar a ser aplicado.

5.9. CONSIDERAÇÕES FINAIS

Nesse estudo de caso a SITAI foi aplicada a fim de ser validada. Infelizmente não foi possível aplicar a SITAI em sua plenitude. Entretanto, nas etapas que ela foi aplicada, pode-se constatar que ela proporciona uma boa referência para a implantação de TAIs.

Quanto ao estudo de caso, constatou-se que o banco de itens possui deficiência em relação a baixa quantidade de itens difíceis, o que prejudica a estimação da habilidade de respondentes com proficiência alta. Por outro lado, essa pequena quantidade de itens difíceis fez com que o ponto de corte, nas situações estudadas, ficasse localizado entre os itens fáceis na escala (0, 1). Certamente, se forem elaborado mais itens difíceis, esse ponto de corte deveria ser recalculado, já que a taxa de aprovação seria menor, caso fossem mantidos os pontos de corte especificados.

6. CONCLUSÕES E TRABALHOS FUTUROS

Nesse capítulo são apresentadas as conclusões em relação aos objetivos determinados e aos resultados obtidos, e as propostas sugeridas para os trabalhos futuros, tanto em relação à sistemática elaborada quanto em relação ao estudo de caso.

6.1. CONCLUSÕES

O objetivo geral desse trabalho era desenvolver uma sistemática para a implantação de Testes Adaptativos Informatizados baseados na Teoria da Resposta ao Item, com a finalidade de servir como um suporte para a orientação nas etapas do desenvolvimento e elaboração de um Teste Adaptativo Informatizado. Na literatura, existem poucos guias para a elaboração de testes adaptativos, sendo que esses são específicos a um determinado tipo de avaliação, ou restritos a algumas características, ou incompletos, ou voltados para a parte mais técnica do teste ou abordam métodos mais antigos. Havia uma necessidade da existência de um material que fornecesse um método para a implantação de um TAI e informações, principalmente, sobre a implementação, a aplicação e a manutenção de um TAI. Esse objetivo foi alcançado e a sistemática elaborada, denominada SITAI, foi apresentada com detalhes no Capítulo 4.3. Entretanto, houve dificuldades para atingir esse objetivo. Uma das dificuldades encontradas, como mencionado, foi a escassez de referências acerca da implantação efetiva de um TAI, especialmente com relação à implementação e a aplicação do TAI. Outra dificuldade encontrada foi a falta de experiência prática do autor na implantação ou na participação de um processo de implantação de um TAI, o que poderia ter contribuído para avaliar e melhorar as etapas de implementação, aplicação e manutenção da SITAI.

Um dos objetivos específicos era realizar um levantamento bibliográfico exaustivo sobre Testes Adaptativos Informatizados, o qual seria utilizado para contribuir com o desenvolvimento da sistemática proposta nessa tese. Esse objetivo foi alcançado e o levantamento bibliográfico encontra-se no Capítulo 3. Inicialmente, pretendia-se analisar todas as referências encontradas, porém a excessiva quantidade de referências tornou essa pretensão inviável na prática. Na revisão bibliográfica, estão incluídos testes, relatórios técnicos, artigos de periódico e congressos de várias épocas. Porém, devido ao excesso de literatura, foram priorizadas as publicações de periódicos científicos, principalmente as mais recentes. Por outro lado, essa imensa quantidade de literatura está basicamente focalizada no que diz respeito ao

algoritmo do TAI (critérios de seleção de itens, métodos de estimação, regras de parada, controle da exposição dos itens, etc.). Na revisão bibliográfica sobre TAIs, optou-se por descrever vários desses métodos, embora poucos foram utilizados no estudo de caso.

Outro objetivo específico era realizar um levantamento bibliográfico sucinto sobre a Teoria da Resposta ao Item (TRI) voltado para a aplicação de TAIs. Esse objetivo foi alcançado e esse levantamento foi apresentado no Capítulo 2. A TRI foi descrita sucintamente, onde foram apresentadas suas características e o Modelo Logístico Unidimensional de Três Parâmetros (ML3) que foi utilizado no estudo de caso dessa tese.

Outro objetivo específico foi definir as etapas necessárias para a implantação de um TAI, segundo as características específicas do teste. Esse objetivo foi alcançado e as etapas necessárias foram incorporadas na sistemática elaborada, apresentada no Capítulo 4.3. Existem várias características do teste que foram consideradas, por exemplo, a possibilidade de aproveitar um teste tradicional existente, o objetivo do teste, o tipo de teste (com itens sigilosos ou não), quantidade de itens a serem elaborados, o tamanho da amostra, a existência de informação prévia dos examinandos, etc. Também foram apresentadas alternativas para reduzir a quantidade de parâmetros, alterar o algoritmo do teste ou elaborar mais itens, nos casos em que o teste podia estar comprometido devido ao tamanho pequeno de amostra ou à pouca quantidade de itens no banco.

Outro objetivo específico consistia em identificar os mais diferentes métodos e critérios utilizados para a elaboração de um TAI, segundo as suas características e especificações. Esse objetivo foi alcançado em parte, uma vez que não foram explorados todos os métodos e critérios encontrados na literatura pesquisada, apresentada no Capítulo 3. Nesse trabalho, devido a grande variedade de métodos e critérios existentes, procurou-se explorar os principais métodos e critérios, e indicar as referências que utilizam outros métodos e critérios.

O último objetivo específico consistia em aplicar a sistemática desenvolvida em um estudo de caso, a fim de validar a mesma. Essa aplicação foi feita na avaliação teórica para a obtenção da carteira de habilitação de motorista realizada pelo Departamento de Trânsito do Estado de Santa Catarina – DETRAN-SC. Esse objetivo foi alcançado em parte, uma vez que não foi possível utilizar todas as etapas da sistemática, principalmente em relação às etapas de implementação, aplicação e manutenção. Embora a SITAI não tenha sido aplicada em sua plenitude, nas etapas em que ela foi aplicada, pode-se constatar que

ela proporciona uma boa referência para a implantação de TAIs. Não houve grandes dificuldades para seguir a SITAI em relação à definição do teste, à calibração do banco de itens, à elaboração do algoritmo e à análise da precisão e da validade.

Em relação ao estudo de caso, constatou-se que o banco de itens possui deficiência em relação à baixa quantidade de itens difíceis na escala elaborada, o que prejudica a estimação da habilidade de respondentes com proficiência alta. Por outro lado, essa pequena quantidade de itens difíceis fez com que o ponto de corte, nas situações estudadas, ficasse localizado entre os itens fáceis na escala (0, 1). Certamente, se forem elaborados mais itens difíceis, esse ponto de corte deverá ser recalculado, já que a taxa de aprovação seria menor, caso fossem mantidos os pontos de corte especificados.

Embora tenham sido identificadas deficiências no banco de itens, em todas as situações estudadas, foi encontrado pelo menos um teste que pode ser considerado adequado. Dessa forma, qualquer um desses testes, considerando, contudo, o objetivo e a população, pode ser implementado.

Todos os testes escolhidos foram selecionados, basicamente, por apresentar, em comparação aos demais, o menor percentual na aprovação de examinandos que deveriam ter sido reprovados, que é uma situação crítica para o caso do DETRAN-SC, e por apresentarem resultados aceitáveis em relação aos demais critérios. Observa-se também que todos os testes selecionados utilizaram o método de estimação MVP., ou seja, esse método proporcionou um menor percentual na aprovação de examinandos que deveriam ter sido reprovados. Nota-se que, se esse fator crítico não fosse considerado nos testes com objetivo de estimação, os testes escolhidos seriam os que utilizam o método bayesiano EAP ou MAP, pois apresentavam geralmente bons desempenhos em relação aos critérios EPM, RQEQM, DEM, EF e CL.

Observou-se que, nos testes de classificação que utilizaram IC como critério de parada, à medida que a habilidade média estimada foi diminuindo, a quantidade de itens aplicados foi aumentando, enquanto que nos testes de estimação de habilidade que utilizaram erro padrão fixo como critério de parada, a medida que a habilidade média estimada foi aumentando, a quantidade de itens aplicados também foi aumentando.

6.2. TRABALHOS FUTUROS

Quanto aos trabalhos futuros, várias sugestões foram levantadas durante a elaboração dessa tese, tanto em relação à sistemática elaborada quanto ao estudo de caso.

Em relação à sistemática, pretende-se expandi-la para outros tipos de TAIs, como, por exemplo, o os Testes Autoadaptativos Informatizados, onde o examinando tem o controle sobre o teste escolhendo o nível de dificuldade do teste ou dos itens, antes do início do mesmo. O mesmo pode ser feito com os Testes Sequenciais Adaptativos Informatizados, onde o algoritmo seleciona um conjunto de itens ao invés de um item de cada vez, e com outros tipos de TAIs.

Quanto ao estudo de caso apresentado, sugere-se levar em conta outras situações nas simulações, tais como, avaliar outras taxas de exposição do item, incorporar o balanceamento de conteúdo e verificar o desempenho do MRN na calibração do banco de itens.

Como, nesse teste, o indivíduo que é reprovado, pode repeti-lo, depois de um certo tempo, quantas vezes desejar ou até ser aprovado, sugere-se a utilização da informação dos testes anteriores realizados pelo indivíduo. Por meio da identificação do indivíduo, é possível verificar quantos testes ele fez anteriormente, quais os itens respondidos, quais foram suas respostas e qual foi a sua avaliação (nota estimada ou classificação). Esses registros podem (ou não) ser utilizados para evitar a administração de itens que já foram respondidos e para estimar a habilidade inicial do indivíduo no próximo teste que ele será submetido. Sugere-se fazer simulações que considerem esse tipo de situação, comparando os resultados nas situações em que essas informações históricas são ou não utilizadas.

Também sugere-se a utilização de outro software de simulação de TAI, a fim de verificar outros métodos que podem ser utilizados no algoritmo, já que o CATSim possui algumas limitações. Por exemplo, o software CATSim não oferece outro método de seleção de itens que não seja baseado no método MI (por exemplo, os critérios MIG e MIE). Segundo Costa (2009), o critério MIG é mais eficiente que o método MI no início do teste, quando a quantidade de respostas do examinando ainda é pequena. O software CATSim também não permite que a opção do método MI no ponto de corte seja utilizada apenas no início do teste, conforme recomendado por Bergstrom e Lunz, (1999) e Renom e Doval (1999), no caso dos testes de classificação.

A quinta calibração resultou na permanência de itens com alto poder de discriminação, o que é bom para um TAI. Por outro lado, apenas 16 estão posicionados na parte superior da escala, o que não é

bom para um TAI. Essa pouca quantidade de itens pode prejudicar a estimação da habilidade daqueles que possuem habilidade acima da média. Dessa forma, outra sugestão é considerar outras calibrações menos rigorosas do banco de itens, a fim de acrescentar mais itens para a parte superior da escala e reduzir um provável impacto negativo na estimação da habilidade daqueles que estão acima da média. Por exemplo, pode-se calibrar um banco de itens constituído por itens com média e alta discriminação (parâmetro $a > 0,7$) ou um banco misto, com itens mais discriminativos na parte inferior da escala (onde há abundância de itens) e itens menos discriminativos na parte superior da escala (onde há escassez de itens), para serem testados nas simulações do TAI.

Como mencionado por Olma (2008), o teste teórico do DETRAN-SC aborda vários conteúdos, o que poderia indicar uma possível avaliação multidimensional. Sugere-se ajustar uma modelo multidimensional, simular TAIs multidimensionais e comparar o desempenho com o teste unidimensional.

Nesse estudo, não foi verificada a validade do teste tradicional existente, porém foi suposto que o teste era válido. Numa situação prática para a efetiva implantação de um TAI, sugere-se que seja verificada a validade do teste existente, caso ele seja aproveitado na elaboração do TAI.

Ainda em relação ao estudo de caso, sugere-se também: a elaboração de mais itens com boa qualidade, principalmente com grau de dificuldade maior (posicionados na parte superior da escala), um estudo mais aprofundado sobre o ponto de corte ideal, e a implementação efetiva do teste, baseada na sistemática desenvolvida.

A elaboração de itens com boa qualidade fará com que o teste seja mais eficiente, ou seja, será possível alcançar um resultado preciso com poucos itens administrados, o que é adequado no caso dos testes onde o critério de parada é determinado pelo erro padrão. No estudo de caso abordado, uma consequência disso seria a adoção de um erro padrão menor do que 0,5, que é considerado alto.

A elaboração de itens com grau de dificuldade maior e com boa qualidade irá contribuir para a redução do erro padrão da habilidade estimada de indivíduos que possuem habilidade acima da média.

A definição de um ponto de corte adequado irá contribuir para garantir a aprovação de candidatos que realmente estejam acima do nível de conhecimento necessário para adquirir a CNH. Diferente de como foi feito nesse estudo, onde o ponto de corte foi feito baseado na taxa de aprovação do teste existente, uma alternativa para a definição do

ponto de corte seria definir em qual nível da escala é atingida a quantidade mínima de conhecimento para a aquisição da CNH. A definição desse nível depende da interpretação da escala, ou seja, isso deve ser feito conjuntamente com os especialistas nos assuntos envolvidos na avaliação teórica do DETRAN-SC.

6.3. CONSIDERAÇÕES FINAIS

A sistemática desenvolvida nessa tese pretende servir como um método, que pode ser utilizado como um guia, principalmente ao usuário leigo que deseja desenvolver e implantar um TAI, orientando-o quanto à criação das etapas do teste, salientando os cuidados necessários na elaboração do mesmo, e indicando os métodos e critérios adequados para o teste a ser construído, de acordo com as características e especificações do teste. Essa tese pretende se tornar um referencial teórico para o usuário que deseja desenvolver um TAI, o qual poderá consultá-la tanto em relação à sistemática quanto ao levantamento bibliográfico realizado. Entretanto, os TAIs evoluem rapidamente, novos estudos são publicados a cada dia, apresentando novas técnicas e procedimentos. O usuário deve tomar cuidado para manter-se atualizado em relação a esses aspectos.

Esse trabalho apresentou muitos aspectos que devem ser observados na elaboração de um TAI. No estudo de caso abordado, nem todos esses aspectos foram abordados da forma que deveria, caso o TAI fosse efetivamente implantado. Para fins didáticos, muitas suposições foram feitas, por exemplo, não foi verificada a validade do teste existente aproveitado, não foi verificada a validade do TAI nos casos em que foram aplicados poucos itens nas simulações, e não foi considerado o balanceamento de conteúdo, que é necessário em testes que abordam vários conteúdos, como o teste teórico do DETRAN-SC. Numa situação prática, esses aspectos devem ser verificados com maior rigor, para que o TAI elaborado seja válido.

Como foi visto, diversos países do mundo já aderiram aos TAIs em várias avaliações de larga escala, incluindo avaliações educacionais, de proficiência de língua estrangeira, testes de aptidão e de admissão para emprego. Os benefícios obtidos com a implantação de um TAI são muitos e superam as desvantagens e as dificuldades em muitos casos. Entre os benefícios, podem ser observados: redução de custos (deslocamento, espaço físico e aquisição de materiais), redução de material físico (papéis, impressoras, tinta), conformidade com os padrões atuais de sustentabilidade (energia limpa, resíduo zero), redução no tempo de execução da prova (maior precisão nos resultados),

divulgação imediata do resultado, redução de erros (de correção e de elaboração de questões), redução de "cola" ou vazamento de provas. Vários testes também permitem ao candidato uma flexibilidade de data e o horário para realizar o teste e podem ser realizado até mesmo em casa, via internet.

Os TAIs possuem um caminho livre para seguir no Brasil. Obviamente, a implantação de TAI depende de tempo e de recursos financeiros consideráveis, embora os benefícios resultantes superem os custos envolvidos, como mencionado por Fetzer et al. (2008) e Hornke (1999). Nesse trabalho, foram apresentadas as vantagens e desvantagens da utilização dos TAIs. O leitor pode observar que o ganho relacionado às vantagens supera as desvantagens, na maioria dos casos. Muitos programas de avaliações internacionais em larga escala utilizam com sucesso o TAI, já faz algumas décadas. No Brasil, assim como a TRI, que começou a ser utilizada tardiamente em avaliações de larga escala (comparando com as avaliações internacionais), porém com sucesso, os TAIs estão começando a ser utilizados. Algumas avaliações existentes, como o ENEM e o vestibular da UAB (Universidade Aberta do Brasil), podem ser consideradas como potenciais candidatas para uma futura aplicação de um TAI. A tendência é que no futuro, essas e outras avaliações nacionais sejam feitas baseadas num TAI.

REFERÊNCIAS

ABAD, F. J.; OLEA, J.; AGUADO, D.; PONSODA, V. Deterioro de parámetros de los ítems en tests adaptativos informatizados: estudio con eCAT. **Psicothema,** v. 22, n. 2, p. 340-347, 2010.

ABAD, F. J.; OLEA, J.; PONSODA, V. Analysis of the optimum number alternatives from the Item Response Theory. **Psicothema,** v. 13, n. 1, p. 152-158, 2001.

ABAD, F. J.; OLEA, J.; PONSODA, V.; XIMÉNEZ, M. C.; MAZUELA, P. Efectos de las Omisiones en la Calibración de un Test Adaptativo Informatizado. **Metodología de las Ciencias del Comportamiento**, Suplemento 2004, p. 1-6, 2004.

ABAD, F. J.; OLEA, J.; REAL, E.; PONSODA, V. Estimación de habilidad y precisión en tests adaptativos informatizados y tests óptimos. Un caso práctico. **Revista Electrónica de Metodología Aplicada**, v. 7, n. 1, p. 1-20, 2002.

ABERNATHY, L. J. **Computerized placement tests: A revolution in testing instruments**. New York: College Board, 1986.

ACKERMAN, T. Developments in Multidimensional Item Response Theory. **Applied Psychological Measurement,** 20: 309-310, 1996a.

ACKERMAN, T. Graphical Representation of Multidimensional Item Response Theory Analyses **Applied Psychological Measurement,** 20: 309-310, 1996b.

ADEMA, J. J. The construction of customized two-stage tests. **Journal of Educational Measurement**, 27, 241–253, 1990.

ADEMA, J. J.; BOEKKOOI-TIMMINGA, E.; VAN DER LINDEN, W. Achievement test construction using 0-1 linear programming. **European Journal of Operational Research**, n. 55, p. 103-111, 1991.

AGUADO, D.; RUBIO, V. J.; HONTANGAS, P. M.; HERNÁNDEZ, J. M. Propiedades psicométricas de un test adaptativo informatizado para la medición del ajuste emocional. **Psicothema,** V. 17, n. 3, p. 484-491, 2005.

AGUADO, D.; SANTA CRUZ, C.; DORRONSORO, J. R.; RUBIO, V. Algoritmo mixto mínima entropia-máxima información para la selección de ítems en un test adaptativo. **Psicothema**, 12, 12-14, 2000.

AL-A'ALI, M. Implementation of an Improved Adaptive Testing Theory. **Educational Technology & Society**, 10 (4), 80-94, 2007.

AL-AMRI, S. Computer-based testing vs. paper-based testing: a comprehensive approach to examining the comparability of testing modes. **Essex Graduate Student Papers in Language & Linguistics**, v. 10, p. 22-44, 2008.

ALMOND, R. G.; MISLEVY, R. J. Graphical Models and Computerized Adaptive Testing. **Applied Psychological Measurement**, Vol. 23 No. 3, p. 223–237, 1999.

ALEXANDRE, J. W. C.; ANDRADE, D. F.; VASCONCELOS, A. P.; ARAUJO, M. A. S. Aplicação da Teoria da Resposta ao Item na Gestão da Qualidade: Proposta de um Modelo Probabilístico. In: XX Encontro Nacional de Engenharia de Produção – ENEGEP, 2001, Salvador. **Anais...** Salvador: ABEPRO, 2001.

ALEXANDRE, J. W. C.; ANDRADE, D. F.; VASCONCELOS, A. P.; ARAUJO, A. M. S. Uma proposta de análise de um construto para a medição dos fatores críticos da gestão pela qualidade através da teoria da resposta ao item. **Gestão & Produção**, v.9, n.2, p.129-141, 2002a.

ALEXANDRE, J. W. C.; ANDRADE, D. F.; VASCONCELOS, A. P.; ARAUJO, A. M. S.; BATISTA, M. J. Teoria da Resposta ao Item: Aplicação do Modelo de Escala Gradual na Gestão pela Qualidade. In: XXII Encontro Nacional de Engenharia de Produção – ENEGEP, 2002b, Curitiba. **Anais...** Curitiba: ABEPRO, 2002b.

ALEXANDRE, J. W. C.; ANDRADE, D. F.; VASCONCELOS, A. P.; ARAUJO, A. M. S.; BATISTA, M. J. Uma Proposta de Análise da Maturidade Organizacional na GQT via Teoria da Resposta ao Item. In: XXXV SBPO, 2003, Natal. **Anais...** Rio de Janeiro: SOBRAPO, 2003a. v. 1. p. 1-20.

ALEXANDRE, J. W. C.; ANDRADE, D. F.; VASCONCELOS, A. P.; ARAUJO, A. M. S.; BATISTA, M. J. Análise do Número de Categorias da Escala de Likert Aplicada à Gestão pela Qualidade Total Através da Teoria da Resposta ao Item. In: XXIII Encontro Nacional de Engenharia de Produção – ENEGEP, 2003, Ouro Preto. **Anais...** Santa Bárbara d´Oeste - SP: ABEPRO, 2003b. v. 1. p. 1-20.

ALMEIDA, V. L. **Avaliação do desempenho ambiental de estabelecimentos de saúde, por meio da Teoria da Resposta ao Item, como incremento da criação do conhecimento organizacional.** 2009.

189 f. Tese (Doutorado em Engenharia e Gestão do Conhecimento) - Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2009.

ANASTASI, A. **Testes psicológicos.** 2ª edição, São Paulo: EPU, 1977.

ANATCHKOVA, M. D. et al. Development and Preliminary Testing of a Computerized Adaptive Assessment of Chronic Pain. **The Journal of Pain**, v. 10, n. 9, p. 932-943. 2009.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. **Teoria da resposta ao item: conceitos e aplicações.** São Paulo: ABE - Associação Brasileira de Estatística, 2000.

ANDRADE, D. F.; VALLE, R. C. Introdução à Teoria da resposta ao item: conceitos e aplicações. **Estudos em Avaliação Educacional**, p. 13-32, 1998.

ANDRICH, D. A rating formulation for ordered response categories. **Psychometrika**, n. 43, p. 561-573, 1978.

ANDRICH, D.; LUO, G. A Hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. **Applied Psychological Measurement**, n. 17, p. 253-276, 1993.

ANDRIOLA, W. B. Funcionamento Diferencial dos Itens (DIF): Estudo com Analogias para Medir o Raciocínio Verbal. **Psicologia: Reflexão e Crítica**, v. 13, n. 3, p.475-483, 2000.

ARIAS, M. R. M.; RIVAS, M. T. Análisis de Escalas Acumulativas: Modelo Probabilistico de Mokken para Itens Dicotomicos. **Psicothema**, v. 3,. n. 1, p. 119-218, 1991.

ARIEL, A.; VAN DER LINDEN, W. J.; VELDKAMP, B. P. A Strategy for Optimizing Item-Pool Management. **Journal of Educational Measurement**. Vol. 43, No. 2, pp. 85–96, 2006.

ARIEL, A.; VELDKAMP, B. P. Optimal testlet pool assembly for multistage testing designs. **Applied Psychological Measurement**, 30, 204-215, 2006.

ARIEL, A.; VELDKAMP, B. P.; VAN DER LINDEN, W. J. Constructing rotating item pools for constrained adaptive testing. **Journal of Educational Measurement**, 41, 345-360, 2004.

ARMSTRONG, R. D.; JONES, D. H.; KOPPEL, N. B.; PASHLEY, P. J. Computerized Adaptive Testing With Multiple-Form Structures. **Applied Psychological Measurement**, V. 28 N. 3, p. 147–164, 2004.

ASSESSMENT SYSTEMS CORPORATION. **MicroCAT Testing System.** St. Paul. MN: Author, 1994.

AZEVEDO, C. L. N. **Métodos de Estimação na Teoria da Resposta ao Item.** 2003. 133 f. Dissertação (Mestrado em Estatística) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2003.

AZEVEDO, C. L. N. **Modelos Longitudinais de Grupos Múltiplos Multiníveis na Teoria da Resposta ao Item: Métodos de Estimação e Seleção Estrutural sob uma Perspectiva Bayesiana.** 2008. 265 f. Tese (Doutorado em Ciências) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2008.

BACCHIERI, G.; BARROS, A. J. D. Acidentes de trânsito no Brasil de 1998 a 2010: muitas mudanças e poucos resultados. Rev. **Saúde Pública,** São Paulo, Epub 16-Set-2011. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89102011005000069&lng=pt&nrm=iso>. Acesso em: 29/09/2011.

BAEK, S.-G. Computerized adaptive attitude testing using the partial credit model. **Dissertation Abstracts International Section A**: Humanities and Social Sciences, 1995.

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

BAKER, F. B.; KIM, S. **Item Responde Theory: parameter estimation techniques.** 2. ed. revised and expanded. New York: Marcel Dekker, 2004.

BAN, J.; HANSON, B. A.; WANG, T.; YI, Q.; HARRIS, D. J. A comparative study of online pretest item calibration/scaling methods in CAT. **Annual meeting of the AERA**, New Orleans, USA, 2000.

BARBERO, M. I. M. Gestión informatizada de bancos de ítems. Em J.Olea, V. Ponsoda y G. Prieto (Eds). **Tests informatizados. Fundamentos y aplicaciones**.(pp. 63-83). Madri, Espanha, 1999: Pirámide.

BARRADA, J. R.; ABAD, F. J.; OLEA, J. Varying the Valuating Function and the Presentable Bank in Computerized Adaptive Testing. **The Spanish Journal of Psychology**, Vol. 14 No. 1, p. 500-508, 2011.

BARRADA, J. R.; ABAD, F. J.; VELDKAMP, B. P. Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. **Psicothema**, v. 21, n. 2, p. 313-320, 2009.

BARRADA, J. R.; MAZUELA, P.; OLEA, J. Maximum information stratification method for controlling item exposure in computerized adaptive testing. **Psicothema,** vol. 18, nº 1, pp. 156-159, 2006.

BARRADA, J. R.; OLEA, J.; ABAD, S. 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. Methods for restricting maximum exposure rate in computerized adaptative testing. **Methodology,** 3, 14-23, 2007.

BARRADA, J. R.; OLEA, J.; PONSODA, V. Reglas de selección de ítems en Tests Adaptativos Informatizados. **Metodología de las Ciencias del Comportamiento**, Suplemento 2004, p. 55-61, 2004.

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**, 34: p. 438-452, 2010.

BARRADA, J. R.; OLEA, J.; PONSODA, V.; ABAD, F. J. Estrategias de selección de ítems en un test adaptativo informatizado para la evaluación del inglés escrito. **Psicothema**, 18, 828-834, 2006.

BARRADA, J. R.; OLEA, J.; PONSODA, V.; ABAD, F. J. Incorporating randomness to the Fisher information for improving item exposure control in CATs. **British Journal of Mathematical and Statistical Psychology**, 61, p. 493-513, 2008.

BARRADA, J. R.; OLEA, J.; PONSODA, V.; ABAD, F. J. **Item bank disclosure in computerized adaptive testing: What makes an item selection rule safer?** Manuscrito submitido para publicação, 2011. Disponível em: <http://www.uam.es/becarios/jbarrada/papers/disclosure.pdf>. Acesso em: 01/07/2011.

BARRADA, J. R.; OLEA, J.; PONSODA, V.; ABAD, F. J. Item selection rules in Computerized Adaptive Testing: Accuracy and security. **Methodology**, 5, 7-17, 2009.

BARRADA, J. R.; VELDKAMP, B. P.; OLEA, J. Multiple Maximum Exposure Rates in Computerized Adaptive Testing. **Applied Psychological Measurement,** v. 33, n. 1, p. 58–73, 2009.

BARTRAM, D.; HAMBLETON, R. K. **Computer-based testing and the internet: Issues and advances.** Chichester, UK: Wiley. 2006.

BARTROFF, J. Modern Sequential Analysis and its Applications to Computerized Adaptive Testing. **Psychometrika**, V. 73, N. 3, p. 473-486, 2008.

BATISTA, M. J.; ALEXANDRE, J. W. C. Teoria da resposta ao item: proposta de análise na gestão pela qualidade total. In: 16° SINAPE - Simpósio Nacional de Probabilidade e Estatística, 2004, Caxambu - MG. **Resumos...** São Paulo: ABE - Associação Brasileira de Estatística, 2004.

BATISTA, M. J.; VASCONCELOS, A. P.; ALEXANDRE, J. W. C.; ANDRADE, D. F.; ARAUJO, A. M. S. Teoria Clássica de Medida e Teoria da Resposta ao Item: uma abordagem comparativa na gestão da qualidade. In: 34ª Reunião Regional da Associação Brasileira de Estatística, 2002, Fortaleza - CE. **Resumos.** São Paulo: ABE - Associação Brasileira de Estatística, 2002, p. 14.

BAZÁN, J. L. **Uma Família de Modelos de Resposta ao Item Normal Assimétrica.** 2005. 133 f. Tese (Doutorado em Estatística) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2005.

BEATON, A. E.; ALLEN, N. L. Interpreting Scales through Scale Anchoring. **Journal of Educational Statistics**, n. 17, p. 191-204, 1992.

BEJAR, I. I. A generative approach to psychological and educational measurement. In: FREDERIKSEN, N; MISLEVY, R. J.; BEJAR, I. I. (Eds.), **Test theory for a new generation of tests** (pp. 323-357). Hillsdale, NJ: Lawrence Erlbaum, 1993.

BEJAR, I. I.; LAWLESS, R. R.; MORLEY, M. E.; WAGNER, M. E.; BENNET, R. E.; REVUELTA, J. A Feasibility Study of On-the-Fly Item Generation in Adaptive Testing. **Journal of Technology, Learning, and Assessment**, v. 2, n. 3, 2003.

BELOV, D. I.; ARMSTRONG, R. A Monte Carlo approach to the design, assembly, and evaluation of multistage adaptive tests. **Applied Psychological Measurement**, 32, 119-137, 2008.

BELOV, D. I.; ARMSTRONG, R. Automatic Detection of Answer Copying via Kullback-Leibler Divergence and K-Index. **Applied Psychological Measurement**, v. 34, n. 6, p. 379-392, 2010.

BELOV, D. I.; ARMSTRONG, R. Direct and Inverse Problems of Item Pool Design for Computerized Adaptive Testing. **Educational and Psychological Measurement,** V. 69, N. 4, 533-547, 2009.

BELOV, D. I.; ARMSTRONG, R Monte Carlo test assembly for item pool analysis and extension. **Applied Psychological Measurement**, 29, 239–261, 2005.

BELOV, D. I.; ARMSTRONG, R.; WEISSMAN, A. A Monte Carlo approach for Adaptive Testing With Content Constraints. **Applied Psychological Measurement**, 32, n. 6, p. 431-446, 2008.

BERGSTROM, B. CAT Fit Analysis, **Rasch Measurement Transactions,** v. 4:3, p. 112, 1990.

BERGSTROM, B. A.; LUNZ, M. E. CAT for certification and licensure. In: DRASGOW, F.; OLSON-BUCHANAN, J. B. (Eds.). **Innovations in computerized assessment.** Mahwah, NJ: LEA. p. 67-92, 1999.

BERGSTROM, B. A.; LUNZ, M. E. Confidence in Pass/Fail Decisions for Computer Adaptive and Paper and Pencil Examinations. **Evaluation and the Health Professions**, 15, 4, p. 453-464, 1992.

BERGSTROM, B. A.; LUNZ, M. E.; GERSHON, R. C. Altering the Level of Difficulty in Computer Adaptive Testing. **Applied Measurement in Education,** v. 5, n. 2, 1992.

BINET, A.; SIMON, TH. A. M´ethodes nouvelles pour le diagnostic du niveau intellectual des anormaux. **l'Anne´e Psychologie**, 11, 191–336, 1905.

BIRNBAUM, A. Some Latent Trait Models and Their Use in Infering an Examinee's Ability. In: LORD, F. M.; NOVICK, M. R. **Statistical Theories of Mental Test Scores.** Reading, MA: Addison-Wesley, 1968.

BJORNER, J. B.; CHANG, C.-H.; THISSEN, D.; REEVE, B. B. Developing tailored instruments: Item banking and computerized adaptive assessment. **Quality of Life Research,** 16, 95-108, 2007.

BJORNER, J. B.; KOSINSKI, M.; WARE JUNIOR, J. E. Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT™). **Quality of Life Research**, 12: p. 913–933, 2003.

BLAIS, J-G.; RAÎCHE, G. Features of the sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules. **11th International Objective Measurement Worksho**p, New Orleans, USA, 2002.

BLAIS, J-G.; RAÎCHE, G. Features of the sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules. **Journal of Applied Measurement,** v. 11, n. 4, 2010.

BLOXOM, B.; VALE, C. D. Multidimensional adaptive testing: an approximate procedure for updating. **Meeting of the Psychometric Society**, Montreal, USA, 1987.

BMDP Statistical Software. **BMDP Statistical Software Manual.** Release 7, vol. 1 e 2. Los Angeles: BMDP Statistical Software, 1992.

BOCK, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. **Psychometrika**, v. 37, p. 29-51, 1972.

BOCK, R. D.; AITKIN, M. Marginal maximum likelihood estimation of item parameters: An application of a EM algorithm. **Psychometrika**, v. 46, p. 433-459, 1981.

BOCK, R.; GIBBONS, R.; SCHILLING, S.; MURAKI, E. W.; WOOD, R., **TESTFACT 4** (Computer software). Lincolnwood, IL: Scientific Software International, 2003.

BOCK, R. D.; LIEBERMAN, M. Fitting a response model for *n* dichotomously scored items. **Psychometrika**, v. 35, p. 179-197, 1970.

BOCK, R. D.; MISLEVY, R. J. Adaptive EAP estimation of ability in a microcomputer environment. **Applied Psychological Measurement**, 6, 4, p. 431-444, 1982.

BOCK, R. D.; ZIMOWSKI, M. F. Multiple Group IRT. In: VAN DER LINDER, W. J.; HAMBLETON, R. K. **Handbook of Modern Item Response Theory.** New York: Spring-Verlag, 1997.

BODMANN, A. M; ROBINSON, D. H. Speed and Performance Differences Among Computer-Based and Paper-Pencil Tests. **J. Educational Computing Research**, V. 31(1), p. 51-60, 2004.

BOEKKOOI-TIMMINGA, E. A method for designing Rasch model-based item banks. **Annual meeting of the Psychometric Society,** Princeton, NJ, USA, 1991.

BOOMSMA, A.; VAN DUIJN, M. A. J.; SNIJDERS, T. A. B. **Essays On Item Response Theory. Lecture Notes In Statistics.** Springer Verlag Pod, 2001.

BORNIA, A. C.; ANDRADE, D. F.; POSSAMAI, O.; MAFRA, P. M. R.; ALMEIDA, V. L. Satisfação do congressista em relação ao Congresso Brasileiro de Custos por meio da teoria da resposta ao item. In: XVI Congresso Brasileiro de Custos, 2009, Fortaleza. **Anais...** Fortaleza, 2009.

BORTOLOTTI, S. L. V. **Aplicação de um modelo de desdobramento graduado generalizado da teoria da resposta ao item – TRI.** 2003. 107 f. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2003.

BORTOLOTTI, S. L. V. **Resistência à Mudança Organizacional: Medida de Avaliação por meio da Teoria da Resposta ao Item.** 2010. 291 f. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2010.

BORTOLOTTI, S. L. V.; ANDRADE, D. F. Aplicação de um modelo de desdobramento graduado generalizado – GGUM da teoria da resposta ao item. **Estudos em Avaliação Educacional**, São Paulo, v. 18,. n. 37, p. 157-188, 2007.

BORTOLOTTI, S. L. V.; MOREIRA JUNIOR, F. J.; SOUZA JUNIOR, A. F.; ANDRADE, D. F. Proposta de Avaliação da Satisfação por meio do Modelo Logístico de Dois Parâmetros da Teoria da Resposta ao Item. In: VI CNEG - Congresso Nacional de Excelência em Gestão, 2010, Rio de Janeiro. **Anais do VI CNEG.** Niterói, RJ: UFF - Universidade Federal Fluminense, 2010a.

BORTOLOTTI, S. L. V.; MOREIRA JUNIOR, F. J.; SOUZA JUNIOR, A. F.; ANDRADE, D. F. Teoria Da Resposta ao Item - Medida de Satisfação por meio do Modelo Logístico de Dois Parâmetros. In: 19 SINAPE - Simpósio Nacional de Probabilidade e Estatística, 2010, São Pedro, SP. **Resumos**, 2010b.

BORTOLOTTI, S. L. V.; SOUZA JUNIOR, A. F.; ANDRADE, D. F. Uma Metodologia para Avaliação da Satisfação através da Teoria da Resposta Ao Item - TRI. In: VI SEGeT - Simpósio de Excelência em Gestão e Tecnologia, 2009, Rezende - RJ. **Anais...** Resende - RJ: AEDB - Associação Educacional Dom Bosco, 2009.

BOWLES, R.; POMMERICH, M. An examination of item review on a CAT using the specific information item selection algorithm. **Annual meeting of the National Council of Measurement in Education**, Seattle, WA, 2001.

BOYD, A.; DODD, B. G.; FITZPATRICK, F. A comparison of exposure control procedures in CAT systems based on different measurement models for testlets using the verbal reasoning section of the MCAT. **Annual meeting of the National Council on Measurement in Education,** Chicago, IL, USA, 2003.

BRADLOW, E. T.; WAINER, H.; WANG, X. A bayesian random effects model for testlets. **Psychometrika**, 64, 153-168, 1999.

BRADLOW, E. T.; WEISS, R. E. **Outlier measures and norming methods for computerized adaptive tests** (Research report). Iowa City, IA, USA: ACT, 1999.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Edital No 4, de 24 de Setembro de 2010. Exame Nacional Do Ensino Médio - ENEM 2010. **Diário Oficial da União**, Brasília, DF, 27 de setembro de 2010, Seção 3, p. 59-60.

BRASIL, Ministério das Cidades. Portaria n° 400 de 2 de setembro de 2005. Altera a Portaria nº 227, de 4 de julho de 2003 e dá outras providências. **Diário Oficial da União**, Brasília, DF, 5 de setembro de 2005, Seção 1, p. 77-78.

BREITHAUPT, K.; ARIEL, A. A.; HARE, D. R. Assembling an Inventory of Multistage Adaptive Testing Systems. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing**. Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p. 247-266.

BREITHAUPT, K.; HARE, D. R. Automated Simultaneous Assembly of Multistage Testlets for a High-Stakes Licensing Examination. **Educational and Psychological Measurement,** V. 67, N. 1, p. 5-20, 2007.

BRIDGEMAN, B.; CLINE, F. **Variations in Mean Response Times for Questions on the Computer- Adaptive GRE General Test: Implications for Fair Assessment.** GRE Board Professional Report No. 96-20P. Princeton, USA: ETS - Educational Testing Service, 2000.

BRIDGEMAN, B.; CLINE, F.; HESSINGER, J. **Effect of Extra Time on GRE® Quantitative and Verbal Scores.** GRE Board Report No. 00-03P. Princeton, USA: ETS - Educational Testing Service, 2003.

BUGBEE JUNIOR, A. C. The equivalence of paperand-pencil and computer-based testing. **Journal of Research on Computing in Education**, 28, 282-290, 1996.

BURGHOF, K. L. Assembling an item-bank for computerised linear and adaptive testing in Geography. **International Education Journal.** Vol 2, No 4, 2001.

BURT, W.; DAVIS, L. L.; DODD, B. G. A Comparison of Item Exposure Control Procedures Using a CAT System Based on the Generalized Partial Credit Model. **Annual meeting of the American Educational Research Association,** Chicago, 2003.

CASSETTARI, N. Pagamento por performance na educação básica. In: **31ª. Reunião Anual da ANPED**, Caxambu-MG. Rio de Janeiro: ANPED - Associação Nacional de Pós-Graduação e Pesquisa em Educação, 2008.

CASTRO JUNIOR, J. S. **Método de Avaliação de Maturidade para a Implantação de Sistemas de Informação Estratégica em Empresas de tecnologia da Informação e Comunicação.** 2007, 103 f. Dissertação (Mestrado em Engenharia Elétrica). Programa de Pós-Graduação em Engenharia Elétrica. Departamento de Engenharia Elétrica, Faculdade de Tecnologia, Universidade de Brasília. Brasília, 2007.

CECCATO, M. G. B. et al. Compreensão da terapia anti-retroviral: uma aplicação de modelo de traço latente. **Cad. Saúde Pública**. Rio de Janeiro, vol. 24, n. 7, p. 1689-1698, 2008.

CELLA, D.; GERSHON, R. **Assessment Center[SM] User Manual.** 2010. Disponível em: <http://www.assessmentcenter.net/ac1/AssessmentCenter_Manual.pdf>. Acesso em: 09/09/2011.

CELLA, D.; GERSHON, R.; LAI, J-S.; CHOI, S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. **Qual Life Res,** 16:133–141, 2007.

CHAKRAVARTY, E. F.; BJORNER, J. B.; FRIES, J. F. Improving Patient Reported Outcomes Using Item Response Theory and Computerized Adaptive Testing. **The Journal of Rheumatology**, 34:6, p. 1426-1431 2007.

CHALHOUB-DEVILLE, M. **Issues in computer-adaptive testing of reading proficiency.** Studies in Language Testing 10. University of Cambridge, NY: 2000.

CHALHOUB-DEVILLE, M.; DEVILLE, C. Computer adaptive testing in second language contexts. **Annual Review of Applied Linguistics**, 19, 273-299, 1999.

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

CHANG, S. W.; ANSLEY, T. N.; LIN, S. H. Performance of item exposure control methods in computerized adaptive testing: Further explorations. **Annual meeting of the American Educational Research Association,** New Orleans, LA, USA, 2000.

CHANG, S. W.; HARRIS, D. J. Redeveloping the Exposure Control Parameters of CAT Items when a Pool is Modified. **Annual Meeting of the American Educational Research Association**, New Orleans, LA, USA, 2002.

CHANG, S-R.; PLAKE; B. S.; KRAMER, G. A.; LIEN, S.-M. Development and Application of Detection Indices for Measuring Guessing Behaviors and Test-Taking Effort in Computerized Adaptive Testing. **Educational and Psychological Measurement**, 71(3) p. 437–459, 2011.

CHANG, H. H.; QIAN, J.; YING, α. A-Stratified Multistage Computerized Adaptive Testing with *b*-Blocking. **Applied Psychological Measurement**, n. 25, p. 333-341, 2001.

CHANG, S.-W.; TWU, B.-Y. **A comparative study of item exposure control methods in computerized adaptive testing** (Research report 1998-3). Iowa City, IA, USA: ACT, 1998.

CHANG, S-W.; TWU, B-Y. Effects of changes in the examinees' ability distribution on the exposure control methods in CAT. **Annual Meeting of the American Educational Research Association,** Seattle, USA, 2001.

CHANG, H. H.; VAN DER LINDEN, W. J. Optimal stratification of item pools in α-stratified computerized adaptive testing. **Applied Psychological Measurement**, 27, 262-274, 2003.

CHANG, H. H.; YING, Z. A Global Information Approach to Computerized Adaptive Testing. **Applied Psychological Measurement**, n. 20, p. 213-229, 1996.

CHANG, H. H.; YING, Z. α-stratified Multistage Computerized Adaptive Testing. **Applied Psychological Measurement**, n. 23, p. 211-222, 1999.

CHANG, H.-H.; YING, Z. Nonlinear sequential designs for logistic item response models with applications to computerized adaptive tests. **The Annals of Statistics**, 37, 1466–1488, 2009.

CHANG, H. H.; YING, Z. To weight or not to weight? Balancing influence of initial items in adaptive testing. **Psychometrika**, 73, 441–450, 2008.

CHANG, H. H.; ZHANG, J. Hypergeometric family and item overlap rates in computerized adaptive testing. **Psychometrika**, 67, 387–398, 2002.

CHANG, Y-C. I. Application of sequential interval estimation to adaptive mastery testing. **Psychometrika**, v. 70, n. 4, p. 685-713, 2005.

CHANG, Y-C. I.; LU, H-Y. Online Calibration Via Variable Length Computerized Adaptive Testing. **Psychometrika**, v. 79, n. 1, p. 140-157, 2010.

CHEN, S.-Y. A Procedure for Controlling General Test Overlap in Computerized Adaptive Testing. **Applied Psychological Measurement**, n. 34 (6), p. 393-409, 2010.

CHEN, S.-Y.; ANKENMANN, R. D. Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. **Journal of Educational Measurement**, 41, 149-174, 2004.

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**, n. 24, p. 241-255, 2000.

CHEN, S.; ANKENMANN, R. D.; SPRAY, J. A. The relationship between item exposure and test overlap in computerized adaptive testing. **Journal of Educational Measurement,** 40(2), 129-145, 2003.

CHEN, S. Y.; DOONG, S. H. Predicting item exposure parameters in computerized adaptive testing. **British Journal of Mathematical and Statistical Psychology**, 61, 75-91, 2008.

CHEN, S., HOU, L.; DODD, B. G. A comparison of maximum likelihood estimation and expected a posteriori estimation on CAT using the partial credit model. **Educational and Psychological Measurement**, 53, 61-77, 1998.

CHEN, S. K., HOU, L., FITZPATRICK, S. J., DODD, B. G. The effect of population distribution and method of theta estimation on computerized adaptive testing (CAT) using the rating scale model. **Educational and psychological measurement**, 57, 422-439, 1997.

CHEN, S-Y.; LEI, P. W. Controlling Item Exposure and Test Overlap in Computerized Adaptive Testing. **Applied Psychological Measurement**, n. 29, p. 204-217, 2005.

CHEN, S. Y.; LEI, P. W.; LIAO. W. H. Controlling item exposure and test overlap on the fly in computerized adaptive testing. **British Journal of Mathematical and Statistical Psychology,** 61, p. 471-492, 2008.

CHENG, Y. Improving Cognitive Diagnostic Computerized Adaptive Testing by Balancing Attribute Coverage: The Modified Maximum Global Discrimination Index Method. **Educational and Psychological Measurement**, vol. 70, n. 6: p. 902-913, 2010.

CHENG, Y. When cognitive diagnosis meets computerized adaptive testing: CD-CAT. **Psychometrika**, v. 74, n. 4, p. 619-632, 2009.

CHENG, Y.; CHANG, H.-H. The maximum priority index method for severely constrained item selection in computerized adaptive testing. **British Journal of Mathematical and Statistical Psychology**, 62, 369–383, 2009.

CHENG, Y.; CHANG, H.-H.; YI, Q. Two-Phase Item Selection Procedure for Flexible Content Balancing in CAT. **Applied Psychological Measurement,** Vol. 31 No. 6, p. 467–482, 2007.

CHENG, Y.; CHANG, H.-H.; DOUGLAS, J.; GUO, F. Constraint-Weighted a-Stratification for Computerized Adaptive Testing With Nonstatistical Constraints: Balancing Measurement Efficiency and Exposure Control. **Educational and Psychological Measurement**, v. 69, n. 1, p. 35-49, 2009.

CHENG, P. E.; LIOU, M. Computerized adaptive testing using the nearest-neighbors criterion. **Applied Psychological Measurement,** 27, 204-216, 2003.

CHENG, P. E.; LIOU, M. Estimation of trait level in computerized adaptive testing. **Applied Psychological Measurement,** 24, 257-265, 2000.

CHOI, S. W. Firestar: Computerized Adaptive Testing Simulation Program for Polytomous Item Response Theory Models. **Applied Psychological Measurement.** V. 33 N. 8, p. 644-645, 2009.

CHOI, I.; KIM, K.; BOO, J. Comparability of a paper-based language test and a computer-based language test. **Language Testing**, vol. 20(3), 295-320, 2003.

CHOI, S. W.; GRADY, M. W.; DODD, B. G. A New Stopping Rule for Computerized Adaptive Testing. **Educational and Psychological Measurement,** 71(1), p. 37–53, 2011.

CHOI, S. W.; SWARTZ, R. J. Comparison of CAT Item Selection Criteria for Polytomous Items. **Applied Psychological Measurement**, V. 33, N. 6, p. 419-440, 2009.

CHOI, S.W.; TINKLER, T. Evaluating comparability of paper-and-pencil and computer-based assessment in a K–12 setting. **Annual meeting of the National Council on Measurement in Education**, New Orleans, LA, USA, 2002.

CISAR, S. M. et al. Computer Adaptive Testing of Student Knowledge. **Acta Polytechnica Hungarica**, v. 7, n. 4, 2010.

CLARÉS, J. Propuesta de Desarrollo de Test Informatizado Adaptándolo a las Respuestas del Usuario. **Pixel-Bit. Revista de Medios y Educación**, n. 31, p. 19-30, 2008.

COHEN, J. A coefficient of agreement for nominal scales. **Educacional and Psychological Measurement**, v. 20, p. 37-46, 1960.

COLLINS, J.; GREER, J.; HUANG, S. Adaptive assessment using granularity hierarchies and bayesian nets. In: FRASSON, C.; GAUTHIER, G.; LESGOLD, A. (Eds) **Proceedings of the Third International Conference on Intelligent Tutoring Systems**, Montreal (Canadá): Springer, p. 569-577, 1996.

COLTON, G. D. Exam security and high-tech cheating. **The Bar Examiner**, 67(3), p. 13-35, 1998.

COMPUTER ADAPTIVE TECHNOLOGIES. **CAT software System.** Chicago, IL: Author., 1994.

CONDÉ, F. N.; LAROS, J. A. Unidimensionalidade e a Propriedade de Invariância das Estimativas da Habilidade pela TRI. **Avaliação Psicológica**, Porto Alegre, v. 6, n. 2, p. 205-215, 2007.

CONEJO, R.; GUZMAN, E.; MILLAN, E.; TRELLA, M.; PEREZ-DE-LA-CRUZ, J. L; RIOS, A. "SIETTE: A Web–Based Tool for Adaptive Testing". **International Journal of Artificial Intelligence in Education**, vol.14, pp.29-61, 2004.

CONEJO, R.; MILLÁN, E.; CRUZ J. L. P.; TRELLA, M. Modelado del Alumno: um enfoque bayesiano. Inteligencia Artificial, **Revista Iberoamerica de Inteligencia Artificial,** n. 12, p. 50-58, 2001.

COOK, K. F. et al. Development and Psychometric Evaluation of the Flexilevel Scale of Shoulder Function. **Medical Care,** v. 41, n. 7, p. 823–835, 2003

COOK, K. F. et al. Development of a Flexilevel Scale for use with computeradaptive testing for assessing shoulder function **J Shoulder Elbow Surg,** v. 14, n. 1S, 2005.

COSTA, D. R. **Métodos Estatísticos em Testes Adaptativos Informatizados.** Dissertação. 2009. 120 f. (Mestrado em Estatística) – Departamento de Métodos Estatísticos, Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2009.

COSTA, D. R.; FERNANDES, P. G. M. Protótipo CAT para provas de proficiência em línguas estrangeiras da Universidade de Brasília. **I CONBRATRI - I Congresso Brasileiro de Teoria de Resposta ao Item**, apresentação de poster, Florianópolis, 2009.

COSTA, D. R.; KARINO, C. A. MOURA, F. A. S.; ANDRADE, D. F. A Comparison of Three Methods of Item Selection for Computerized Adaptive Testing. In D. J. Weiss (Ed.), **Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing,** 2009.

COSTA, M. B. F.; CHAVES NETO, A. Aplicação da Teoria da Resposta ao Item (TRI) na avaliação do grau de satisfação do consumidor em um serviço específico. In: 34ª Reunião Regional da Associação Brasileira de Estatística, 2002, Fortaleza - CE. **Resumos.** São Paulo: ABE - Associação Brasileira de Estatística, 2002, p. 31-31.

COSTER, W. J. et al. Assessing Self-Care and Social Function Using a Computer Adaptive Testing Version of the Pediatric Evaluation of Disability Inventory. **Archives of Physical Medicine and Rehabilitation**, v. 89, n. 4, p. 622-629, 2008.

CRESWELL, J. W. **Projeto de Pesquisa.** 2ª edição. Porto Alegre: Artmed, 2007.

CROCKER, L.; ALGINA, J. **Introduction to Classical and Modern Test Theory.** New York: Holt, Rinehart and Winston.,1986.

CRONBACH, L. J. **Fundamentos da Testagem Psicológica.** Porto Alegre: Artes Médicas, 1996.

CRONBACH, L. J.; GLESER, G. C. **Psychological test and personnel decisions** (2nd ed.). Urbana: University of Illinois Press, 1957.

CUESTA, M. Unidimensionalidad. En J. MUÑIZ (Coord.) **Psicometría**.(pp. 239-292). Madri, Espanha: Universitas, 1996.

CUNHA, R. C. L. V.; SENA JUNIOR, M. R.; MATOS, G. S. Medindo Satisfação do Consumidor através do Modelo Rasch. In: 34ª Reunião Regional da Associação Brasileira de Estatística, 2002, Fortaleza - CE. **Resumos.** São Paulo: ABE - Associação Brasileira de Estatística, 2002.

CURA JUNIOR, C. et al. Uma Ferramenta Adaptativa de Avaliação da Aprendizagem Baseada no Perfil Cognitivo e Metacognitivo do Estudante. **Revista de Informática Aplicada,** v. 3, n. 2, p. 41-48, 2007.

CURI, M.; PITON-GONÇALVES, J.; RICARTE, T. A. M.; ALUÍSIO, S. M. Métodos de Seleção de Itens em Teste Adaptativo Multidimensional. **I CONBRATRI - I Congresso Brasileiro de Teoria de Resposta ao Item**, apresentação de poster, Florianópolis, 2009.

DALPIAZ, M. G. G. **SIA on-line: Sistema Integrado de Avaliação On-line baseado em Testes Adaptativos Informatizados (TAIs).** 2007, 75 f. Trabalho de Conclusão de Curso (Tecnólogo em Análise e Desenvolvimento de Sistemas). Faculdade de Tecnologia SENAC/RS, Porto Alegre, 2007.

DAMANDO, F. S. **Ferramenta Computacional de Apoio Pedagógico Baseada em Testes Adaptativos Informatizados e Teoria de Resposta ao Item.** 2003, 103 f. Dissertação (Mestrado em Engenharia da Computação). Programa de Pós-Graduação em Engenharia Elétrica e de Computação, Escola de Engenharia Elétrica, Universidade Federal de Goiás, Goiânia, 2003.

DAMANDO, F. S.; GUEDES, L. G. R. Ferramenta Computacional de Apoio Pedagógico Baseada em Testes Adaptativos Informatizados e Teoria de Resposta ao Item. 2° Seminário Nacional ABED de Educação a Distância, 2004, Campo Grande. **Anais...** São Paulo: ABED, 2004, v. 1, p. 85-92.

DAMANDO, F. S.; GUEDES, L. G. R. Testes Adaptativos Informatizados baseados em Teoria de Resposta ao Item utilizados em ambientes virtuais de aprendizagem. **Novas Tecnologias na Educação**, Porto Alegre, v. 3, n. 2, p. 1-8, 2005.

DAMANDO, F. S.; RIBEIRO, L.; MARTINS, W.; GUEDES, L. G. R. Ferramenta Avaliativa Dinâmica a partir da Teoria de Resposta ao Item. In: I Encontro Regional em Modelagem e Análise Computacional de Sistemas, Goiânia - ERMACS, 2004. **Anais...** Goiânia: UCG – Universidade Católica de Goiás, 2004a.

DAMANDO, F. S.; RIBEIRO, L.; MARTINS, W.; GUEDES, L. G. R. Ferramenta Avaliativa Pedagógica para Cursos a Distância Baseada em Testes Adaptativos Informatizados e Teoria de Resposta ao Item. In: XII SEMINCO – Seminário de Computação, 2004, Blumenau, **Anais do XIII SEMINCO.** Blumenau: FURB, 2004b.

DAVEY, T.; FAN, M. Specific information item selection for adatpive testing. **Annual meeting of the National Council on Measurement in Education**, New Orleans, 2000.

DAVEY, T.; NERING, M. Controlling item exposure and maintaining item security. In: MILLS, C.; POTENZA, M. T.; FREMER, J. J.; WARD, W. C. (Eds.). **Computer-Based Testing: Building the**

Foundation for Future Assessments, Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2002.

DAVEY, T.; OSHIMA, T. C.; LEE, K. Linking Multidimensional Item Calibrations. **Applied Psychological Measurement**, 20: 405-416, 1996.

DAVEY T.; PARSHALL, C. G. New algorithms for item selection and exposure control with computerized adaptive testing. **Annual meeting of the American Educational Research Association**, San Francisco, CA, USA, 1995.

DAVIS, L. L. Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. **Applied Psychological Measurement,** 28(3), 165-185, 2004.

DAVIS, L. L.; DODD B. G. **An examination of testlet scoring and item exposure constraints in the verbal reasoning section of the MCAT.** MCAT Monograph Series: Association of American Medical Colleges, 2001.

DAVIS, L. L.; DODD, B. G. Item Exposure Constraints for Testlets in the Verbal Reasoning Section of the MCAT. **Applied Psychological Measurement**, 27(5), 335-356, 2003.

DAVIS, L. L.; DODD, B. G. Strategies for Controlling Item Exposure in Computerized Adaptive Testing with the Partial Credit Model. **Journal of Applied Measurement,** v. 9, n. 1, p. 1-17, 2008.

DAVIS, L. L.; PASTOR, D. A.; DODD, B. G.; CHIANG, C.; FITZPATRICK, S. An examination of exposure control and content balancing restrictions on item selection in CATs using the partial credit model. **Journal of Applied Measurement**, v. 4, n. 1, p. 24-42, 2003.

DE AYALA, R. J. A Comparison of the Nominal Response Model and the Three-Parameter Logistic Model en Computerized Adaptive Testing. **Educational and Psychological Measutement,** v. 49, p. 789-805, 1989.

DE AYALA, R. J. The influence of dimensionality on CAT ability estimation. **Educational and psychological measurement**, 52, 513-528, 1992.

DE AYALA, R. J. The nominal response model in computerized adaptive testing. **Applied Psychological Measurement,** 16, 327-342, 1992.

DE AYALA, R. J. **The Theory and Practice of Item Response Theory.** New York, USA: The Guilford Press, 2008.

DE AYALA, R. J.; DODD, B. G.; KOCH, W. R. A comparison of the partial credit and graded response models in computerized adaptive testing. **Applied Measurement in Education**, 5, 17–34, 1992.

DE AYALA, R. J.; SAVA-BOLESTA, M. Item parameter recovery for the nominal response model. **Applied Psychological Measurement**, 23, 3-19, 1999.

DE AYALA, R. J.; SCHAFER, W.; SAVA-BOLESTA, M. An investigation of the standard errors for expected a posteriori ability estimates. **British Journal of Mathematical and Statistical Psychology,** 47, 385-405, 1995.

DE BEER, M. A comparison of learning potential results at various educational levels. Paper presented at the 6th Annual Society for Industrial and Organisational Psychology of South Africa (SIOPSA) conference, **Anals...,** 2003.

DE BEER, M. Development of the Learning Potential Computerised Adaptive Test (LPCAT). **South African Journal of Psychology**, v. 35; n. 4, p. 717-747, 2005.

DE BEER, M. Utility of learning potential computerised adaptive test (LPCAT) scores in predicting academic performance of bridging students: A comparison with other predictors. Paper presented at the 5th Annual Industrial Psychology Conference, Pretoria, CSIR, **Anals...,** 2002.

DE BOECK, P.; WILSON, M. **Explanatory item response models: A generalized linear and nonlinear approach.** New York, USA: Springer, 2004.

DE LA TORRE, R.; VISPOEL, W. P. **The development and evaluation of a computerized adaptive testing system.** ERIC Accession No: ED338711, 1991.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm (with discussion). **Journal of the Royal Statistical Society**, Series B, n. 39, p; 1-38, 1977.

DENG, N. **References of Non-Commercial Software for IRT Analyses.** 2011. Disponível em

<http://www.umass.edu/remp/software/CEA-699_6-2-11.doc>. Acesso em: 31/08/2011.

DENG, N.; HAMBLETON, R. **20 Software Packages for Assessing Test Dimensionality.** 2007. Disponível em <http://www.umass.edu/remp/software/IRTSoftware_TestDimensionality.pdf>. Acesso em: 31/08/2011.

DENG, H.; ANSLEY, T. To Stratify or Not: An Investigation of CAT Item Selection Procedures under Practical Constraints. **2003 NCME annual meeting**, Chicago, IL, USA, 2003.

DENG, H.; CHANG, H. H. A-stratified computerized adaptive testing with unequal item exposure across strata. **Annual Meeting of the American Educational Research Association**, Seattle, WA, 2001.

DESCOVI, L. M. G. Experimento IDA com sistema informático SCOMAX. X Encontro Gaúcho de Educação Matemática. **Anais 10. Encontro Gaúcho de Educação Matemática.** Sociedade Brasileira de Educação Matemática – Regional do Rio Grande do Sul; Departamento de Física, Estatística e Matemática –Universidade Regional do Noroeste do Estado do Rio Grande do Sul. Ijuí: Ed. UNIJUÍ, 2009.

DIAO, Q.; VAN DER LINDEN, W. J. Automated Test Assembly Using lp_Solve Version 5.5 in R. **Applied Psychological Measurement,** v. 35, n. 5, p. 398-409, 2011.

DIAO, Q.; VAN DER LINDEN, W. J.; YEN, S. J. Exposure Control Using Item-ineligibility Probabilities in Multidimensional Computerized Adaptive Testing with Shadow Test.**Annual meeting of the National Council on Measurement in Education**, 2001.

DODD, B. G. The Effect of Item Selection Procedure and Stepsize on Computerized Adaptive Attitude Measurement Using the Ratting Scale Model. **Applied Psychological Measurement**, v. 14, n. 4, p. 355-366, 1990.

DODD, B. G.; DE AYALA, R. J.; KOCH, W. R. Computerized adaptive testing with polytomous items. **Applied Psychological Measurement**, n. 19, p. 5-22, 1995.

DODD, B. G.; KOCH, W. R.; DE AYALA, R. J. Computerized adaptive testing using the partial credit model effects of item pool characteristics and different stopping rules. **Educational and psychological measurement**, 53, 61-77, 1993.

DODD, B. G.; KOCH, W. R.; DE AYALA, R. J. Operational characteristics of adaptive testing procedures using the graded response model. **Applied Psychological Measurement**, 13, 129-143, 1989.

DOONG, S. H. A Knowledge-Based Approach for Item Exposure Control in Computerized Adaptive Testing. **Journal of Educational and Behavioral Statistics.** Vol. 34, No. 4, p. 530–558, 2009.

DRASGOW, F.; OLSON-BUCHANAN, J. B. **Innovations in computerized assessment**. Mahwah, NJ: Erlbaum, 1999.

EDMONDS, J.; ARMSTRONG, R. A mixed integer programming model for multiple stage adaptive testing. **European Journal of Operational Research**, v. 193, n. 2, 1, p. 342-350, 2009.

EDWARDS, M. C.; THISSEN, D. Exploring potential designs for multi-form structure computerized adaptive tests with uniform item exposure. In D. J. Weiss (Ed.), **Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing,** 2007.

EGEBERINK, I. J. L. et al. Detection of aberrant item score patterns in computerized adaptive testing: An empirical example using the CUSUM. **Personality and Individual Differences**, v. 48, n. 8, p. 921-925, 2010.

EGGEN, T. J. H. M. **Contributions to the Theory and Practice of Computerized Adaptive Testing**. Arnhem, Holanda: Citogroep, 2004.

EGGEN, T. J. H. M. Item selection in adaptive testing with the sequential probability ratio test. **Applied Psychological Measurement**, 23, 249-261, 1999.

EGGEN, T. J. H. M. **Overexposure and underexposure of items in computerized adaptive testing** (Measurement and Research Department Reports 2001-1). Arnhem, The Netherlands: Citogroep, 2001.

EGGEN, T. J. H. M. Three-Category Adaptive Classification Testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing.** Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p 373-388.

EGGEN, T. J. H. M.; STRAETMANS, G. J. J. M. Computerized adaptive testing for classifying examinees into three categories. **Educational and Psychological measurement**, 60, 713-734, 2000.

EGGEN, T. J. H. M.; VERSCHOOR, A. J. Optimal Testing With Easy or Difficult Items in Computerized Adaptive Testing. **Applied Psychological Measurement**, V. 30, N. 5, p. 379–393, 2006.

EGGEN, T. J. H. M.; VERHELST, N. D. Item calibration in incomplete testing designs. **Psicológica**, 32, p. 107-132, 2011.

EIGNOR, D. R. **Deriving comparable scores for computer adaptive and conventional tests: An example using the SAT** (Research Report 93-55). Princeton, NJ, USA: Educational Testing Service, 1993.

EIGNOR, D. R.; STOCKING, M. L.; WAY, W. D.; STEFFEN, M. **Case Studies in Computer Adaptive Test Design through Simulation.** RR 93-56. Princeton, NJ, USA: Educational Testing Services, 1993.

EIGNOR, D. R.; WAY, W. D.; AMOSS, K. E. Establishing the comparability of the NCLEX using CAT with traditional NCLEX examinations. **Annual meeting of the National Council on Measurement in Education**, New Orleans, USA, 1994.

EIGNOR, D.; TAYLOR, C.; KIRSCH, I.; JAMIESON, I. **Development of a scale for assessing the level of computer familiarity of TOEFL examinees.** TOEFL Research Reports, Report 60. Princeton, NJ, USA: Educational Testing Services, 1998.

EMBRETSON, S. E. Generating Items During Testing: Psychometric Issues And Models. **Psychometrika**, v. 64, n. 4, p. 407-433, 1999.

EMBRETSON, S. E.; REISE, S.P. **Item Response Theory for Psychologists.** New Jersey, USA: Lawrence Erlbaum Associates, 2000.

FAYERS, P. M. Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. **Qual Life Res.** 16, p. 187–194, 2007.

FAYERS, P. M.; MACHIN, D. **Quality of Life: The Assessment, Analysis and Interpretation of Patient-reported Outcomes**, Second Edition. Wiley: 2007.

FERNANDES, P. G. M. **Sistema Computadorizado de Avaliação Adaptativa em Larga Escala (SCAALE).** 2009, 152 f. Monografia (Graduação em Ciência da Computação). Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade de Brasília, Brasília, 2009.

FERRANDO, P. J. Evaluación de la unidimensionalidad de los ítems mediante análisis factorial. **Psicothema**, v. 8, n. 2, p. 397-410, 1996.

FETZER, M.; DAINIS, A.; LAMBERT, S.; MEADE, A. **Computer Adaptive Testing (CAT) in an Employment Context.** White paper. Roswell, USA: PreVisor, 2008.

FINGER, M. S. A Review of MicroFACT 2.0: A Microcomputer Factor Analysis Program for Ordered Polytomous Data and Mainframe Size Problems. **International Journal of Testing,** V. 4, N. 1, p. 83 – 89, 2004.

FINKELMAN, M. D. On Using Stochastic Curtailment to Shorten the SPRT in Sequential Mastery Testing. **Journal of Educational and Behavioral Statistics**, V. 33, N. 4, p. 442–463, 2008.

FINKELMAN, M. D.; WEISS, D. J.; KIM-KANG, G. Item Selection and Hypothesis Testing for the Adaptive Measurement of Change. **Applied Psychological Measurement** 34: p. 238-254, 2010.

FLAUGHER, R. Item Pools. In: WAINER, H. **Computerized Adaptive Testing: A Primer.** New Jersey, USA: Lawrence Erlbaum Associates, 2000.

FLIEGE, H.; BECKER, J.; WALTER, O. B.; BJORNER, J. B., KLAPP, B. F., ROSE, M. Development of a computer-adaptive test for depression (d-cat). **Quality of Life Research**, v. 14, p. 2277–2291, 2005.

FOLK, V.G.; GREEN, B.F. Adaptive estimation when the unidimensionality assumption of IRT is violated. **Applied Psychological Measurement,** 13, 373-389, 1989.

FONTANIVE, N. S.; ELLIOT, L. G.; KLEIN, R. Os desafios da apresentação dos resultados da avaliação de sistemas escolares a diferentes públicos. **REICE - Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación,** v. 5, n. 2e, 2007.

FORBEY, J. D.; BEN-PORATH, Y. S. Computerized Adaptive Personality Testing: A Review and Illustration With the MMPI–2 Computerized Adaptive Version. **Psychological Assessment**, v. 19, n. 1, p. 14-24, 2007.

FORBEY, J. D.; BEN-PORATH, Y. S.; GARTLAND, D. Validation of the MMPI-2 Computerized Adaptive Version (MMPI-2-CA) in a

Correctional Intake Facility. **Psychological Services**, v. 6, n 4, p. 279-292, 2009.

FRANCISCO, R. **Aplicação da Teoria da Resposta ao Item (TRI) no Exame Nacional de Cursos (ENC) da Unicentro.** 2005. 144 f. Dissertação (Mestrado em Ciências) - Pós-Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná, Curitiba, 2005.

FRENCH, B. F.; THOMPSON, T. D. The Evaluation of Exposure Control Procedures for an Operational CAT. **Annual meeting of the American Educational Research Association.** Chicago, IL, USA, 2003.

FREY, A.; SEITZ, N.-N. Hypothetical Use of Multidimensional Adaptive Testing for the Assessment of Student Achievement in the Programme for International Student Assessment. **Educational and Psychological Measurement,** 71(3) 503–522, 2011.

FREY, A.; SEITZ, N.-N. Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges **Studies In Educational Evaluation,** v 35, n. 2-3, p. 89-94, 2009.

FRIES, J. B.; BRUCE; B.; CELLA, D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. **Clin Exp Rheumatol,** 23(5 Suppl 39):S53-7, 2005.

GARCÍA, D. A.; CRUZ, C. S.; DORRONSORO, J. R.; FRANCO, V. J. R. Algoritmo mixto mínima entropía-máxima información para la selección de ítems en un test adaptativo informatizado. **Psicothema**, v. 12, Suplemento. n. 02, p. 12-13, 2000.

GARCÍA, J. R.; REVUELTA, J. Métodos para Controlar la Sobrexposición e Infraexposición de Ítems en Tests Adaptativos Informatizados. **VII Congreso de Metodología de las Ciencias Sociales y de la Salud.** Valencia, España, 2003.

GARCÍA JIMÉNEZ, E.; GIL, J.; RODRÍGUEZ GÓMEZ, G. La evaluación de tests adaptativos informatizados. **RELIEVE**, v. 4, n. 2, 1998. Disponível em <http://www.uv.es/RELIEVE/v4n2/RELIEVEv4n2_6.htm>. Acesso em 11/09/2009.

GARCÍA-PÉREZ, M. A.; ALCALÁ-QUINTANA, R.; GARCÍA-CUETO, E. A Comparison of Anchor-Item Designs for the Concurrent Calibration of Large Banks of Likert-Type Items. **Applied Psychological Measurement** 34(8), 580-599, 2010.

GARDNER, W.; KELLEHER, K. J.; PAJER, K. A. Multidimensional Adaptive Testing for Mental Health Problems in Primary Care. **Medical Care,** v. 40, n. 9, p. 812–823, 2002.

GARDNER, W.; SHEAR. K.; KELLEHER, K. J.; PAJER, K. A.; MAMMEN, O.; BUYSSE, D.; FRANK, E. Computerized adaptive measurement of depression: A simulation study. **BMC Psychiatry**, 4:13, 2004.

GARRET, H. E. **Estadística en Psicología y Educación.** Buenos Aires, Editorial Paidos, 1979.

GEORGIADOU, E.; TRIANTAFILLOU, E.; ECONOMIDES, A. A review of item exposure control strategies for Computerized Adaptive Testing developed from 1983 to 2005. **Journal of Technology, Learning, and Assessment**, n. 5, 2007.

GEORGIADOU, E.; TRIANTAFILLOU, E.; ECONOMIDES, A. Evaluation parameters for computer-adaptive testing. **British Journal of Educational Technology**, v. 37, n. 2, p. 261-278, 2006.

GERSHON, R. Computer Adaptive Testing. **Journal of Applied Measurement**, v. 6, n. 1, 2005.

GERSHON, R.; BERGSTROM, B. A. CAT and test-wiseness. **Rasch Measurement Transactions**, v. 9:1 p.415, 1995.

GIBBONS, R. D. et al. Using Computerized Adaptive Testing to Reduce the Burden of Mental Health Assessment. **Psychiatric Services,** V. 59 N. 4, 2008.

GIERL, M. J.; ZHOU, J. Computer adaptive-attribute testing: A new approach to cognitive diagnostic assessment. **Zeitschrift für Psychologie,** v. 216, n. 1, p. 29-39, 2008.

GIL, A. C. **Como Elaborar Projetos de Pesquisa.** 5. ed. São Paulo: Atlas, 2010.

GIOUROGLOU, H.; ECONOMIDES, A. The Development of the Adaptive Item Language Assessment (AILA) for Mixed-Ability Students. In: RICHARDS, G. (Ed.). **Proceedings E-Learn 2005 World**

Conference on E-Learning in Corporate, Government, Helthcare, and Higher Education, p. 643-650, Vancouver, Canada, 2005.

GLAS, C. A. W. Item calibration and parameter drift. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic, 2000, p 183-199.

GLAS, C. A. W. Item Parameter Estimation and Item Fit Analysis. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing**. Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p. 269-288.

GLAS, C. A. W.; DAGOHOY, A. V. T. A person fit test for IRT models for polytomous items. **Psychometrika**, v. 72, n. 2, p. 159–180, jun 2007.

GLAS, C. A. W.; VAN DER LINDEN, W. J. Computerized Adaptive Testing With Item Cloning. **Applied Psychological Measurement**, Vol. 27 No. 4, 247–261, 2003.

GLAS, C. A. W.; VAN DER LINDEN, W. J.; GEERLINGS, H. Estimation of the Parameters in an Item-Cloning Model for Adaptive Testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing**. Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p. 289-314.

GLAS, C. A. W.; VOS, H. J. Adaptive Mastery Testing Using a Multidimensional IRT Model. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing.** Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p 409-431.

GLAS, C. A. W.; VOS, H. J. **Testlet-Based Adaptive Mastery Testing** Computerized Testing Report 99-11. Newtown, PA: Law School Admission Council, 2006.

GLAS, C. A. W.; WAINER, H.; BRADLOW, E. T. MML and EAP estimation in testlet based adaptive testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic, 2000, p 271-287.

GLOWACKI, M. L.; MCFADDEN, A. C.; PRICE, B. J. Developing computerized tests for classroom teachers: A pilot study. **Annual Meeting of the Mid-South Educational Research Asssociation**, Biloxi, MS, USA, 1995.

GOLDBERG, A. L.; PEDULLA, J. J. Performance Differences According to Test Mode and Computer Familiarity on a Practice GRE. **Journal of Educational and Psychological Measurement**, 2002.

GORIN, J. S.; DODD, B. G.; FITZPATRICK, S. J.; SHIEH, Y. Y. Computerized Adaptive Testing With the Partial Credit Model: Estimation Procedures, Population Distributions, and Item Pool Characteristics. **Applied Psychological Measurement,** 29; 433-456, 2005.

GREEN, B. F. Alternate methods for scoring computer-based adaptive tests. **Annual meeting of the NCME**, Chicago, USA, 1997.

GREEN, B. F. System design and operation. In WAINER, H. (Ed.) **Computerized Adaptive Testing: A Primer**. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

GREEN, B. F. et al. **Applications of item response theory.** Educational Research Institute of British Columbia: Vancouver, Canada, 1983.

GREEN, B. F.; BOCK, R. D.; HUMPHREYS, L. G.; LINN, R. L.; RECKASE, M. D. Technical guidelines for assessing computerized adaptive tests. **Journal of Educational Measurement**, 4:13, 1984.

GROENWALD, C. L. O.; BECHER, E. L. Características do Pensamento Algébrico de Estudantes do 1º Ano do Ensino Médio. IV SIPEM – Seminário Internacional de Pesquisa em Educação Matemática. **Livro de Resumos.** Brasília, DF: SBEM - Sociedade Brasileira de Educação Matemática, 2009.

GROENWALD, C. L. O.; RUIZ, L. M. Formação de professores de Matemática: uma proposta de ensino com novas tecnologias. **Acta Scientiae**, Canoas, v.8, n. 2, p.19-28 jul./dez. 2006.

GU, L.; RECKASE, M. D. Designing Optimal Item Pools for Computerized Adaptive Tests with Sympson-Hetter Exposure Control. In D. J. Weiss (Ed.), **Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing,** 2007.

GUEWEHR, K. **Teoria da Resposta ao Item na Avaliação de Qualidade de Vida de Idosos.** 2007. 179 f. Dissertação (Mestrado em

Epidemiologia) –Programa de Pós- Graduação em Epidemiologia, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul. Porto Alegre, 2007.

GUILFORD, J. P. **Psychometric Methods.** McGraw-Hill Education, 1954.

GULLIKSEN, H. **Theory of mental tests.** New York: Wiley; 1950.

GUZMÁN, E.; CONEJO, R. A Brief Introduction to the New Architecture of SIETTE. In: NEJDL, W.; DE BRA, P. (Eds.). **Adaptive Hypermedia and Adaptive Web-Based Systems.** Lecture Notes in Computer Science, v. 3137, 2004, p. 405–408.

GUZMÁN, E.; CONEJO, R. A Model for Student Knowledge Diagnosis Through Adaptive Testing. 7th International Conference Intelligent Tutoring Systems, ITS2004, **Proceedings...** Brasil, 2004.

GUZMÁN, E.; CONEJO, R.; GARCÍA-HERVÁS, E. An Authoring Environment for Adaptive Testing. Educational. **Technology & Society**, 8 (3), p. 66-76, 2005.

GUZMÁN, E.; CONEJO, R.; PÉREZ-DE-LA-CRUZ, J-L. Adaptive testing for hierarchical student models. **User Model User-Adap Inter** 17, p.119–157, 2007.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise Multivariada de Dados.** 6ª Edição. Porto Alegre: Bookman, 2009.

HALEY, S. M.; NI, P.; DUMAS, H. M.; FRAGALA-PINKHAM, M. A.; HAMBLETON, R. K.; MONTPETIT, K.; BILODEAU, N.; GORTON, G. E.; WATSON, K.; TUCKER, C. A. Measuring global physical health in children with cerebral palsy: illustration of a multidimensional bi-factor model and computerized adaptive testing. **Quality of Life Research**, 18, p. 359–370, 2009a.

HALEY, S. M.; NI, P.; JETTE, A. M.; TAO, W.; MOED, R.; MEYERS, D.; LUDLOW, L. H. Replenishing a computerized adaptive test of patient-reported daily activity functioning. **Quality of Life Research,** 18, p. 461–471, 2009b.

HALEY, S. M. et al. Computerized Adaptive Testing for Follow-Up After Discharge From Inpatient Rehabilitation: I. Activity Outcomes. **Archives of Physical Medicine and Rehabilitation**, v. 87, n. 8, p. 1033-1042, 2006.

HALEY, S. M. et al. Computerized Adaptive Testing for Follow-Up After Discharge From Inpatient Rehabilitation: II. Participation Outcomes. **Archives of Physical Medicine and Rehabilitation**, v. 89, n. 2, p. 275-283, 2008.

HALKITIS, P. N. CAT algorithm. **Rasch Measurement Transactions,** v. 6:4, p.254-5, 1993.

HALKITIS, P. N. CAT with a Limited Item Bank. **Rasch Measurement Transactions**, v. 9:4, p. 471, 1996.

HALKITIS, P. N. The effect of item pool restriction on the precision of ability measurement for a Rasch-based CAT: comparisons to traditional fixed length examinations. **Journal of Outcome Measurement**, 2(2), 97-122, 1998.

HAMBLETON, R. K.; SLATER, S.; NARAYANAN, P.; SEDIATI, H. Construcción automatizada de los tests: conceptos básicos, avances técnicos y aplicaciones. In: MUÑIZ (Ed.). **Psicometría**. Madri: Universitas, 1996.

HAMBLETON, R. K.; SWAMINATHAN, H. **Item Response Theory: Principles and Applications.** Boston: Kluwer Nijoff; 1985.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H. J. **Fundamentals of item response theory.** Newbury Park, CA: Sage, 1991.

HAMBLETON, R.; ZAAL, J. N.; PIETERS, J. P. M. Computerized adaptive testing: Theory, applications, and standards. In: HAMBLETON, R. K; ZAAL, J. N. (Eds.). **Advances in Educational and psychological testing,** Boston: Kluwer, 1991, p. 341-366.

HAMILTON, L. S.; KLEIN, S. P.; LORIE, W. Using web-based testing for large-scale assessment. **Science Foundation, Division of Elementary, Secondary, and Informal Education.** p. 1–40, 2000.

HAN, K. T. A gradual maximum information ratio approach to item selection in computerized adaptive testing. In D. J. Weiss (Ed.), **Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing**, 2009.

HAN, K. T. **SimulCAT: Windows application that simulates computerized adaptive test administration**. 2010. Disponível em: <http://www.hantest.net/simulcat>. Acesso em: 31/08/2011.

HAN, N. **Using moving averages to assess test and item security in computer based testing** (Center for Educational Assessment Research Report No. 468). Amherst, MA: University of Massachusetts, School of Education, 2003.

HAN, N.; HAMBLETON, R. Detecting Exposed Test Items in Computer-Based Testing.Meeting of the NCME, San Diego, USA, 2004.

HARMES, J. C.; KROMREY, J. D.; PARSHALL, C. G. **Online item parameter recalibration: application of missing data treatments to overcome the effects of sparse data conditions in a computerized adaptive version of the MCAT.** Report submitted to the Association of American Medical Colleges, Section for the MCAT, University of South Florida, 2001.

HART, D. L.; DEUTSCHER, D.; CRANE, P. K.; WANG, Y.-C. Differential item functioning was negligible in an adaptive test of functional status for patients with knee impairments who spoke English or Hebrew. **Qual Life Res.** 18(8):1067-83, 2009.

HART, D. L.; DEUTSCHER, D.; WERNECK, M. W.; HOLDER, J.; WANG, Y.-C. Implementing Computerized Adaptive Tests in Routine Clinical Practice: Experience Implementing CATs. **Journal of Applied Measurement,** v. 11, n. 3, 2010.

HART, D. L.; MIODUSKI, J. E.; STRATFORD, P. W. Simulated computerized adaptive tests for measuring functional status were efficient with good discriminant validity in patients with hip, knee, or foot/ankle impairments. **Journal of Clinical Epidemiology**, 58(6), 629–638, 2005.

HART, D. L.; WANG, Y.-C.; STRATFORD, P. W.; MIODUSKI, J. E. A computerized adaptive test for patients with hip impairments produced valid and responsive measures of function. **Archives of Physical Medicine and Rehabilitation**, 89, 2129-2139, 2008a.

HART, D. L.; WANG, Y.-C.; STRATFORD, P. W.; MIODUSKI, J. E. Computerized adaptive test for patients with knee impairments produced valid and responsive measures of function. **Journal of Clinical Epidemiology**, 61, 1113-1124, 2008b.

HART, D. L.; WANG, Y.-C.; STRATFORD, P. W.; MIODUSKI, J. E. Computerized adaptive test for patients with foot or ankle impairments

produced valid and responsive measures of function. **Qual Life Res.** 17:p. 1081–1091, 2008c.

HARVEY, R. J.; HAMMER, A. L. Item Response Theory. **The Counseling Psychologist**, V. 27, N. 3, p. 353-383, 1999.

HAU, K.; CHANG, H. H. Item selection in computerized adaptive testing: Should more discriminating items be used first? **Journal of Educational Measurement**, p. 38, 249-266, 2001.

HÄUSLER, J. Adaptive success control in computerized adaptive testing. **Psychology Science**, v. 48(4), p. 436-450, 2006.

HAYDEN, J. J. Shocking Our Students to the Next Level: Language Loss and Some Implications for Teaching Chinese as a Foreign Language. **Journal of the Chinese Language Teachers Association**, Volume 38:3, p. 1-20, 2003.

HERRANDO, S. Tests adaptativos computerizados: una sencilla solución al problema de la estimación con puntuaciones perfectas y cero. **II Conferencia Española de Biometría. Biometric Society**, Segovia, España, 1989.

HETTER, R. D.; SEGALL, D. O.; BLOXOM, B. M. Evaluating item calibration medium in computerized adaptive testing. In W.A. Sands, B.K. Waters, and J.R. McBride (Eds.), **Computerized Adaptive Testing: From Inquiry to Operation** (Chapter 16, pp. 161–167). Washington, DC, USA: American Psychological Association, 1997.

HETTER, R. D.; SYMPSON, B. Item exposure control in CAT-ASBAV. In: SANDS, W. A.; WATERS, B. K.; MCBRIDE, J. R. (Eds.). **Computerized Adaptive Testing: from inquiry to operation**, Washington, USA: American Psychological Association, 1997.

HICKS, M. **The TOEFL computerized placement test: adaptive conventional measurement.** (ETS Reports No. 89-12). Princeton, NJ: Educational Testing Services, 1989.

HO, R.-G.; YEN, Y.-C. Design and evaluation of an xml-based platform independent computerized adaptive testing system. **IEEE Transaction on Education**, 2005.

HOCKEMEYER, C.; ALBERT, D. The adaptive tutoring system RATH: A prototype. In: AUERY, M. E.; RESSLER, U. (eds.), **ICL99 Workshop Interative Computer Aided Learning: Tools and Aplicactions**. Willach, Austria, 1994.

HOE, L. S.; KIONG, L. N.; SAM, H. K.; USOP, H. B. Improving educational assessment: A computer-adaptive multiple choice assessment using NRET as the scoring method. **US-China Education Review,** v. 6, n. 5, p. 51-60,2009.

HOIJTINK, H. A Latent trait model for dichotomous choice data. **Psychometrika**, n. 55, p. 641-656, 1990.

HOL, A. M.; VORST, H. C. M.; MELLENBERGH, G. J. A Randomized Experiment to Compare Conventional, Computerized, and Computerized Adaptive Administration of Ordinal Polytomous Attitude Items. **Applied Psychological Measurement**, Vol. 29 No. 3,, p. 159–183, 2005.

HOL, A. M.; VORST, H. C. M.; MELLENBERGH, G. J. Computerized Adaptive Testing for Polytomous Motivation Items: Administration Mode Effects and a Comparison With Short Forms. **Applied Psychological Measurement**, Vol. 31 No. 5, September 2007, 412–429.

HOL, A. M.; VORST, H. C. M.; MELLENBERGH, G. J. Computerized Adaptive Testing of Personality Traits. **Zeitschrift für Psychologie,** v. 216, n. 1, p. 12-21, 2008.

HONTANGAS, P.; OLEA, J.; PONSODA, V. Elección de la dificultad de los tests autoadapatados informatizados: un estudio piloto. **RELIEVE**, vol. 4, n. 2, 1998. Disponível em <http://www.uv.es/RELIEVE/v4n2/RELIEVEv4n2_3.htm>. Acesso em 11/09/2009.

HONTANGAS, P.; OLEA, J.; PONSODA, V.; REVUELTA, J.; WISE, S. L. Assisted Self-Adapted Testing: A Comparative Study. **European Journal of Psychological Assessment**, v. 20, n. 1, 2-9, 2004.

HONTANGAS, P.; PONSODA, V.; OLEA, J. Procedimientos de integración numérica y estimación bayesiana en tests adaptativos informatizados. **VI Congreso de Metodología de las Ciencias Sociales y de la Salud**, Oviedo, Espanha, 1999.

HONTANGAS, P.; PONSODA, V.; OLEA, J.; ABAD, F. Los tests adaptativos informatizados en la frontera del siglo XXI: Una revisión. **Metodología de las Ciencias del Comportamiento,** v. 2, n. 2, p. 183-216, 2000a.

HONTANGAS, P.; PONSODA, V.; OLEA, J.; WISE, S. L. The choice of item difficulty in Self-adapted testing. **European Journal of Psychological Assessment**, 16, 3-12, 2000b.

HORNKE, L. F. Benefits from computerized adaptive testing as seen in simulation studies. **European Journal of Psychological Assessment**, 15(2), 91-98, 1999.

HORNKE, L. F. Item response times in computerized adaptive testing. **Psicológica**, 21 (1-2), 175-189 2000.

HORNKE, L. F.; HABON, M. W. Rule based item bank construction and evaluation within the linear logistic framework. **Applied Psychological Measurement**, 10 (4) 369-380, 1986.

HOU, L.; CHEN, S.; DODD, B. G.; FITZPATRICK, S. J. The effects of methods of theta estimation, prior distribution, and number of quadrature points on CAT using the graded response model. **Annual meeting of the AERA**, New York, 1996.

HUANG, S. X. A content-balanced adaptive testing algorithm for computer-based training systems. In: FRASSON, C.; GAUTHIER, G.; LESGOLD, A. ITS – Intelligent Tutorial Systems, Third International Conference, ITS '96, Montréal, Canada, June 1996, Proceedings. **Lecture Notes in Computer Science,** v. 1086, Springer 1996a, p. 306-315.

HUANG, S. X. On content-balanced adaptive testing. In: SÁNCHEZ, A. D. I; CASTRO, I. F. Computer aided learning and instruction in science and engineering: Third International Conference, CALISCE '96, San Sebastian, Spain, July 29-31, 1996, Proceedings. **Lecture Notes in Computer Science,** v. 1108, Springer 1996b, p. 60-68.

HWANG, G.-J. et al. On the Development of a Computer-Assisted Testing System With Genetic Test Sheet-Generating Approach. **IEEE Transactions on systems, man, and cybernetics - Part c: Applications and Reviews,** v. 35, n. 4, 2005.

ILOG, Inc. **CPLEX 9.0 [Computer Program and Manual].** Incline Village, NV: Author, 2003.

IP, H. et al. Development of a computerized adaptive test for assessing balance function in patients with stroke. **Phys Ther.**, 90(9), p. 1336-1344, 2010.

IRAMANEERAT, C.; STAHL, J. Optimizing Item Pool Characteristics to Control Item Exposure in a Computerized Adaptive Test. **Annual American Educational Research Association Meeting,** Chicago, Illinois, USA, 2007.

IRVINE, S.; KYLLONEN, P. **Item generation for tests development**. Hillsdale, NJ: LEA, 2002.

JACOBUSSE, G.; VAN BUUREN, S. Computerized adaptive testing for measuring development of young children. **Statist. Med.** 26, p. 2629-2638, 2007.

JANSKY, L. J.; HUANG, J. C. A Multi-Method Approach to Assess Usability and Acceptability: A Case Study of the Patient-Reported Outcomes Measurement System (PROMIS) Workshop. **Social Science Computer Review.** V. 27, N. 2, p. 262—270, 2009.

JETTE, A. M.; HALEY, S. M. Contemporary Measurement Techniques for Rehabilitation Outcomes Assessment. **J Rehabil Med,** 37, p. 339–345, 2005.

JETTE, A. M.; HALEY, S. M.; NI, P.; OLARSCH, S.; MOED, R. Creating a Computer Adaptive Test Version of the Late-Life Function and Disability Instrument. **Journal of Gerontology: Medical Sciences**, V. 63A, N. 11, p. 1246–1256, 2008.

JOARISTI. L.; LIZASOAIN, L. Estudio de la dimensionalidad empleando análisis factorial clásico y análisis factorial de información total: análisis de pruebas de matemáticas de primaria (5º y 6º cursos) y secundaria obligatoria. **Revista ELectrónica de Investigación y EValuación Educativa – RELIEV**, v. 14, n. 2, p. 1-18, 2008.

JONES, N. BULATS: A case study comparing computer based and paper-and-pencil tests. **Research Notes,** v. 3, p. 10-13, 2000.

JÖRESKOG, K. G.; SÖRBOM, D. **LISREL 8: User's Reference Guide.** Scientific Software International, 1996a.

JÖRESKOG, K. G.; SÖRBOM, D. **PRELIS 2: User's Reference Guide**. Scientific Software International, 1996b.

JULIAN, E. CAT: What feedback? **Rasch Measurement Transactions,** v. 6:4, p. 246, 1993.

KALOHN, J. C.; SPRAY, J. A. The effective of model misspecification on classification decisions made using a computerized test. **Journal of Education Measurement,** 36 (1), p. 47-59, 1999.

KARINO, C. A.; COSTA, D. R.; LAROS, J. A. Adequacy of an Item Pool Measuring English Language Proficiency for Implementing CAT. In D. J. Weiss (Ed.), **Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing,** 2009.

KENG, L.; HO, T.-H.; CHEN, T.-A. A.; DODD, B. G. A Comparison of Item and Testlet Selection Procedures in Computerized Adaptive Testing. **AERA & NCME Conference**, 2008.

KIM, J. K.; NICEWANDER, W. A. Ability estimation for conventional tests. **Psychometrika**, 58, 4, 587-599, 1993.

KIM-KANG, G.; WEISS, D. J. Adaptive measurement of individual change. **Zeitschrift fur Psychologie**, 216, 49-58, 2008.

KIM-KANG, G. K.; WEISS, D. J. Comparison of computerized adaptive testing and classical methods for measuring individual change. **GMAC conference on computerized adaptive testing,** Minneapolis, MN, USA, 2007.

KINGSBURY, G. G. Adapting adaptive testing: Using the MicroCAT Testing System in a local School District. **Educational Measurement: Issues and Practice**, 9, 2, 3-6, 1990.

KINGSBURY, G. G. An Empirical Comparison of Achievement Level Estimates from Adaptive Tests and Paper-and-Pencil Tests. **American Educational Research Association annual meeting**. New Orleans, LA, USA, 2002.

KINGSBURY, G. G. Item review and adaptive testing. **an the annual meeting of the NCME,** New York, USA, 1996.

KINGSBURY, G. G.; HOUSER, R. L. A comparison of achievement level estimates from computerized adaptive testing and paper-and-pencil testing. **Annual Meeting of the American Educational Research Association**, New Orleans, LA, USA, 1988.

KINGSBURY, G. G.; HOUSER, R. L. Developing computerized adaptive tests for school children. In: DRASGOW, F.; OLSON-BUCHANAN, J. B. (Eds.). **Innovations in computerized assessment.** Mahwah, NJ: LEA, 1999. p. 93-116.

KINGSBURY, G. G.; HOUSER, R. L. ICAT: An adaptive testing procedure for the identification of idiosyncratic knowledge patterns. **Zeitschrift für Psychologie,** v. 216, n. 1, p. 40-48, 2008.

KINGSBURY, G. G.; WEISS, D.J. A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. WEISS (Ed.), **New horizons in testing: Latent trait theory and computerized adaptive testing** (pp. 237-254). New York: Academic Press, 1983.

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

KIRSCH, I.; JAMIESON, J.; TAYLOR, C.; EIGNOR, D. **Computer familiarity among TOEFL examinees. TOEFL Research Reports, .Report 59.** Princeton, NJ, USA: Educational Testing Services, 1998.

KLEIN, R.; FONTANIVE, N. S. Avaliação em larga escala: Uma proposta inovadora. **Em Aberto,** Brasília, ano 15, n.66, abr./jun. 1995.

KLEIN, R.; FONTANIVE, N. S.; ELLIOT, L. G. O Exame Nacional do Ensino Médio – Tecnologia e Principais Resultados Em 2005. **REICE - Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación,** v. 5, n. 2e, p. 116-131, 2007.

KOCALEVENT, R.-D. et al. An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. **Journal of Clinical Epidemiology**, v. 62, n. 3, p. 278-287.e3, 2009.

KOCH, W. R.; DODD, B. G. An investigation of procedures for computerized adaptive testing using partial credit scoring. **Applied Measurement in Education**, 2(4), 335-337, 1989.

KOCH, W. R.; DODD, B. G. An investigation of procedures for computerized adaptive testing using the successive intervals Rasch model. **Educational and Psychological Measurement**, 55, 976-990, 1995.

KOLEN, M. J.; BRENNAN, R. L. **Test Equating - Methods and Pratices**. New York, USA: Springer, 1995.

KOSINSKI, M.; BAYLISS, M. S.; BJORNER, J. B.; WARE JUNIOR, J. E.; GARBER, W. H.; BATENHORST, A.; CADY, R. A six-item

short-form survey for measuring headache impact: The HIT-6™. **Quality of Life Research**, n. 12, p. 963–974, 2003.

KOSINSKI, M.; BJORNER, J. B.; WARE JUNIOR, J. E.; SULLIVAN, E.; STRAUS, W. L. An evaluation of a patient-reported outcomes found computerized adaptive testing was efficient in assessing osteoarthritis impact. **Journal of Clinical Epidemiology,** v. 59, n. 7, p. 715-723, 2006.

KREITER, C. D.; FERGUSON, K.; GRUPPEN, L. D. Evaluating the usefulness of computerized adaptive testing for medical in-course assessment. **Academic Medicine**, 74(10), 1125-1128, 1999.

KUO, T.-H.; LIN, H.-T.; YUAN, S.-M. Web-Based Adaptive Testing System. **Wuhan University Journal of Natural Sciences**, v. 11, n. 1, p. 313-322, 2006.

LAI, J. S.; CELLA, D.; CHANG, C. H.; BODE, R. K.; HEINEMANN, A. W. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. **Quality of Life Research**, 12(5), 485-501, 2003.

LAI, J. S.; CELLA, D.; DINEEN, K.; BODE, R.; VON ROENN, J.; GERSHON, R. C., et al. An item bank was created to improve the measurement of cancer-related fatigue. **Journal of Clinical Epidemiology**, 58(2), 190-197, 2005.

LANGE, R. Binary Items and Beyond: A Simulation of Computer Adaptive Testing Using the Rasch Partial Credit Model. **Journal of Applied Measurement,** v. 9, n. 1, 2008.

LARKIN, K. C.; WEISS, D. J. **An empirical comparison of two-stage and pyramidal adaptive ability testing** (Research Report, 75-1). Minneapolis: Psychometrics Methods Program, Department of Psychology, University of Minnesota, 1975.

LAU, C. A.; WANG, T. A New Item Selection Procedure for Mixed Item Type in Computerized Classification Testing. **AERA Annual Meeting** , New Orleans, Louisiana, USA, 2000.

LAU, C. A.; WANG, T. Comparing and combining dichotomous and polytomous items with SPRT procedure in computerized classification testing. **Annual meeting of the American Educational Research Association**, San Diego, USA, 1998.

LAU, C. A.; WANG, T. Computerized classification testing under practical constraints with a polytomous model. **annual meeting of the American Educational Research Association**, Montreal, Canada, 1999.

LEE, Y.-H.; IP, E. H.; FUH, C.-D. A Strategy for Controlling Item Exposure in Multidimensional Computerized Adaptive Testing. **Educational and Psychological Measurement**, v. 68, n. 2, p. 215-232, 2008.

LEE, Y.-H.; PARK, J.-H.; PARK, I-.Y. Estimation of an Examinee's Ability in the Web-Based Computerized Adaptive Testing Program IRT-CAT. **Journal of Educational Evaluation for Health Professions**, 3:4, p. 1-9, 2006.

LEUNG, C.-K.; CHANG, H.-H.; HAU, K.-T. An enhanced stratified computerized adaptive testing design. **Annual meeting of the AERA**, Montreal, USA, 1999.

LEUNG, C.-K.; CHANG, H.-H.; HAU, K.-T. An Examination of Item Selection Rules by Stratified CAT Designs Integrated with Content Balancing Methods. **AERA Annual Meeting**, Seattle, USA, 2001.

LEUNG, C.-K.; CHANG, H.-H.; HAU, K.-T. Comparing Three Item Selection Approaches for Computerized Adaptive Testing with Content Balancing Requirement. **NCME Annual Meeting**, 2002a.

LEUNG, C.-K.; CHANG, H.-H.; HAU, K.-T. Computerized adaptive testing: a mixture item selection approach for constrained situations. **British Journal of Mathematical and Statistical Psychology**, 58, p. 239-257, 2005.

LEUNG, C.-K.; CHANG, H.-H.; HAU, K.-T. Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. **Educational and Psychological Measurement,** 63, 257-270, 2003.

LEUNG, C.-K.; CHANG, H.-H.; HAU, K.-T. Item selection in computerized adaptive testing: improving the a-stratified design with the Sympson- Hetter algorithm. **Applied Psychological Measurement**, 26, 376-392, 2002b.

LEUNG, C.-K.; CHANG, H.-H.; HAU, K.-T.; WEN, Z. Computerized Adaptive Testing: A Comparison of Three Content Balancing Methods **NCME Annual Meeting,** 2003.

LEWIS, C. Some thoughts on controlling item exposure in adaptive testing. In D. J. Weiss (Ed.), **Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing**, 2007.

LEWIS, G.; PELOSI, A. J.; GLOVER, E.; WILINSON, G.; STANFIELD, S. A., P., W. The development of a computerized assessment for minor psychiatric disorder. **Psychological Medicine**, 18, 737-745, 1988.

LI, Y.; LI, S.; WANG, L. **Application of a General Polytomous Testlet Model to the Reading Section of a Large-Scale English Language Assessment.** RR 10-21. Princeton, NJ, USA: Educational Testing Services, 2010.

LI, Y. H.; SCHAFER, W. D. Increasing the homogeneity of CAT's item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. **Journal of Educational Measurement**, 42, 245-269, 2005a.

LI, Y. H.; SCHAFER, W. D. Multidimensional Computerized Adaptive Testing in Recovering Reading and Mathematics Abilities. **Annual meeting of the American Educational Research Association**, Chicago, IL, USA: April, 21-25, 2003a.

LI, Y. H.; SCHAFER, W. D. The Context Effects of Multidimensional CAT on the Accuracy of Multidimensional Abilities and the Item Exposure Rates. **Presented at the 2004 annual meeting of the American Educational Research Association**, San Diego CA, USA: 2004.

LI, Y. H.; SCHAFER, W. D. The Effect of Item Selection Methods on the Accuracy of CAT's Ability Estimates When Item Parameters Are Contaminated with Measurement Errors. **Annual meeting of the National Council on Measurement in Education**, Chicago, IL, USA, 2003b.

LI, Y. H.; SCHAFER, W. D. Trait Parameter Recovery Using Multidimensional Computerized Adaptive Testing in Reading and Mathematics. **Applied Psychological Measurement**, v. 29, n. 1, p. 3-25, January 2005b.

LIGHTSTONE, K.; SMITH, S. M. Student Choice between Computer and Traditional Paper-and-Pencil University Tests: What Predicts Preference and Performance? **Revue internationale des technologies en pédagogie universitaire,** V. 6, n. 1, p. 30-45, 2009.

LILLEY, M.; BARKER, T. Students' Perceived Usefulness of Formative Feedback for a Computer-adaptive Test" **The Electronic Journal of e-Learning**, V. 5 n. 1, p. 31 - 38, 2007.

LILLEY, M.; BARKER, T.; BRITTON, C. The development and evaluation of a software prototype for computer-adaptive testing. **Computers & Education**, V. 43, N. 1-2, p. 109-123, 2004.

LIN, C.-J. Item Selection Criteria With Practical Constraints for Computerized Classification Testing. **Educational and Psychological Measurement,** 71(1) 20–36, 2011.

LIN, C.-J.; SPRAY, J. **Effects of item-selection criteria on classification testing with the sequential probability ratio test**. (Research report 2000-8). Iowa City, IA, USA: ACT, 2000.

LINACRE, J. M. CAT: Maximum possible ability. **Rasch Measurement Transactions,** v. 12:3 p. 657-8, 1998.

LINACRE J. M. Computer-adaptive testing - How many questions are enough. **American Educational Research Association (AERA) Annual meeting,** Washington DC, 1987a.

LINACRE, J. M. **Computer-Adaptive Testing (CAT) by Microcomputer (with computer program - UCAT).** MESA Psychometric Laboratory University of Chicago, MESA Memorandum N. 40, 1987b. Disponível em: <http://www.rasch.org/memo40.htm>. Acesso em: 01/07/2011.

LINACRE, J. M. **Computer-Adaptive Testing CAT: A Methodology Whose Time Has Come.** MESA Psychometric Laboratory University of Chicago, MESA Memorandum N. 69, 2000. Disponível em: <http://www.rasch.org/memo69.pdf>. Acesso em: 06/08/2010.

LINACRE, J. M. Computer-Adaptive Tests (CAT), Standard Errors and Stopping Rules. **Rasch Measurement Transactions**, v. 20:2 p. 1062, 2006.

LINACRE, J. M. A Bayesian Maximum-Falsification approach. **Rasch Measurement Transactions,** v. 9:1 p.412, 1995.

LÓPEZ-CUADRADO, J.; ARMENDARIZ, A. J.; LATAPY, M.; LOPISTÉGUY, P. A Genre-Based Perspective for the Development of Communicative Computerized Adaptive Tests. **Educational Technology & Society**, v. 11 n. 1, p. 87-101, 2008.

LÓPEZ-CUADRADO, J.; FERNÁNDEZ, J. M. S. GenTAI: generador de tests adaptativos informatizados. **Revista Iberoamericana de Informática Educativa**, n. 2, p. 9-24, 2005.

LÓPEZ-CUADRADO, J.; PÉREZ, T. A.; ARMENDARIZ, A. J. Evaluación mediante Tests: ¿Por qué no usar el ordenador? **Revista Iberoamericana de Educación,** n. 36/11, 2005. Disponível em: <http://www.rieoei.org/deloslectores/1040Lopez.PDF>. Acesso em: 18/05/2010.

LORD, F. M. **Applications of Item Response Theory to Practical Testing Problems.** Hillsdale, USA: Lawrence Erlbaum Associates, Inc, 1980.

LORD, F. M. **A theory of test scores** (No. 7). Psychometric Monograph, 1952.

LORD, F. M. Maximun likelihood and Bayesian parameter estimation in item response theory. **Journal of Educational Measurement**, 23, 157-162, 1986.

LORD, F. M. Tailored testing, an application of stochastic approximation. **Journal of American Statistical Association,** n. 66, p. 707-711, 1971c.

LORD, F. M. The self-scoring flexilevel test. **Journal of Educational Measurement,** n. 8, p. 147-151, 1971b.

LORD, F. M. The theoretical study of the measurement effectiveness of flexilevel tests. **Educational and Psychological Measurement**, n. 31, p. 805-813, 1971a.

LORD, F., NOVICK, M. R. **Statistical theories of mental test scores.** Reading, MA:Adisson-Wesley, 1968.

LOZZIA, G. S. et al. Tests informatizados. Nuevos desfíos practicos y éticos para la Evaluación Psicológica. **SUMMA Psicológica UTS**, v. 6, n. 1, p. 135-148, 2009.

LU, Y.; HAMBLETON, R. K. Statistics for detecting disclosed items in CAT environment. **Metodologia de las Ciencias del Comportamiento,** 5, 225-242, 2004.

LU, Y.; RIZAVI, S. Methods for Item Set Selection in Adaptive Testing. **Annual meeting of the National Council on Measurement in Education**, Chicago, USA, 2003.

LUECHT, R. M. Exposure control using adaptive multi-stage item bundles. **Annual Meeting of the National Council on Measurement in Education**, Chicago, IL, USA, 2003.

LUECHT, R. M. Multidimensional Computerized Adaptive Testing in a Certification or Licensure Context. **Applied Psychological Measurement**, 20: 389-404, 1996.

LUECHT, R. M.; BRUMFIELD, T.; BREITHAUPT, K. A Testlet Assembly Design for the Uniform CPA Examination. **Annual Meeting of the National Council on Measurement in Education**, New Orleans, USA, 2002.

LUECHT, R. M.; DECHAMPLAIN, A.; NUNGESTER, R. J. Maintaining content validity in computerized adaptive testing. In: SCHERPBIER, A.; VAN DER VLEUTEN, C.; RETHANS, J.; VAN DER STEEG, A. (Eds.). **Advances in Medical Education**, pp. 416-419. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1997.

LUECHT, R. M.; HADADI, A.; NUNGESTER, R. J. Heuristic-Based CAT: Balancing Item Information, Content, and Exposure. **Annual Meeting of the National Council on Measurement in Education.** New York, NY, USA, 1996.

LUECHT, R. M.; NUNGESTER, R. J. Computer-Adaptive Sequential Testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic, 2000, p 117-128.

LUECHT, R. M.; NUNGESTER, R. J. Some practical applications of computer- adaptive sequential testing. **Journal of Educational Measurement**, 35, 229-240, 1998.

LUMSDEN, J. Test Theory. **Annual Review of Psychology**, 27, 251-280, 1976.

LUNZ, M. E.; BERGSTROM, B. A.; WRIGHT, B. D. The effect of review on student ability and test efficiency for computerized adaptive tests. **Applied psychological measurement**, 16 (1), 33-40, 1992.

LUNZ, M. E.; DEVILLE, C. W. Validity of item selection: A comparison of automated computerized adaptive and manual paper and pencil examinations. **Teaching and Learning in Medicine**, 8(3), 152-157, 1996.

LUNZ M. E.; O'NEILL T. R. CAT: Taking items twice? **Rasch Measurement Transactions**, v. 12:3 p. 656-7, 1998.

LUNZ, M. E.; STAHL, J. A. Patterns of item exposure using a randomized CAT algorithm. **Annual meeting of the National Council on Measurement in Education,** San Diego, CA, USA, 1998.

MADSEN, H. Evaluating A Computer-Adaptive ESL Placement Test. **CALICO Journal,** v. 4, n. 2, 1986.

MAGIS, D.; RAICHE, G. **catR: an R package to generate IRT adaptive tests.** R package version 1.4, 2010.

MAGIS, D.; RAICHE, G. catR: an R package for Computerized Adaptive Testing. **Applied Psychological Measurement,** 35(7) 576–577, 2011.

MARCONI, M. A.; LAKATOS, E. V. **Metodologia científica.** 5ª edição. São Paulo: Atlas, 2007.

MARTÍNEZ-CARDEÑOSO, J. M.; CUESTA, M.; MUÑIZ, J. Dimensionalidad y Función de Información de los Tests. **Psicothema**, v. 8, n. 1, p. 215-220, 1996.

MASTERS, G. N. A Rasch model for partial credit scoring. **Psychometrika**, n. 47, p. 149-174, 1982.

MAZZA, A.; PUNZO, A.; MCGUIRE, B. **KernSmoothIRT: Nonparametric Item Reponse Theory.** R package version 1.0, 2011.

MCBRIDE, J. R. A computerized adaptive version of the Psychological Corporation´s Differential Aptitude Battery. **annual meeting of APA,** Atlanta, GA, USA, 1988.

MCBRIDE, J. R. Some properties of a Bayesian adaptive ability testing strategy. **Applied Psychological Measurement**, 1, 121-140, 1977.

MCBRIDE, J. R.; MARTIN, J. T. Reliability and validity of adaptive ability tests in a military setting. In: WEISS, D. J. (Ed.). **New Horizons in testing**. New York, USA: Academic Press, 1983, p. 223-236.

MCBRIDE, J. R.; PADDOCK, A. F.; WISE, L. L.; STRICKLAND, W. J.; WATERS, B. K. **Testing via the internet: A literature review and analysis of issues for Department of Defense internet testing of the Armed Services Vocational Aptitude Battery (ASVAB)** (FR-01-12). Alexandria, VA: Human Resources Research Organization, 2001.

MCHORNEY C. A. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. **Ann Intern Med.** 127, p. 743-50, 1997.

MCHORNEY C. A. Ten Recommendations for Advancing Patient-Centered Outcomes Measurement for Older Persons. **Ann Intern Med.** 139, p. 403-409-50, 2003.

MCKINLEY, R. L.; WAY, W. D. **The Feasibility of Modeling Secondary TOEFL® Ability Dimensions Using Multidimensional IRT Models.** TOEFL – Technical Report TR-5. New Jersey: ETS, 1992.

MCLEOD, L.; LEWIS, C.; THISSEN, D. A Bayesian method for the detection of item preknowledge in computerized adaptive testing. **Applied Psychological Measurement**, 27, 121-137, 2003.

MEAD, A. D.; DRASGOW, F. Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. **Psychological Bulletin**, 114, 449-458, 1993.

MEIJER, R. R.; NERING, M. L. Computerized Adaptive Testing: Overview and Introduction. **Applied Psychological Measurement**, Vol. 23 No. 3, p. 187–194, 1999.

MEIJER, R. R.; VAN KRIMPEN-STOOP, E. M. L. A. Detecting Person Misfit in Adaptive Testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing**. Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p. 315-330.

MELICAN, G. J.; BREITHAUPT, K.; ZHANG, Y. Designing and Implementing a Multistage Adaptive Test: The Uniform CPA Exam. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing.** Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p. 167-190.

MENDES, E. L. **Uma metodologia para avaliação da satisfação do consumidor com os serviços prestados pelas distribuidoras de energia elétrica.** 2006. 148f. Tese (Doutorado em Engenharia Elétrica) – Programa de Pós-Graduação em Engenharia Elétrica, Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2006.

MEUNIER, L. E. Computer Adaptive Language Tests (Calt) Offer a Great Potential for Functional Testing. Yet, Why Don't They? **CALICO Journal,** v. 11, n. 6, 1994.

MIGUEL, G. B. **Testes Psicométricos e Projetivos: Medidas Psico-Educacionais.** São Paulo, Edições Loyola, 1974.

MILLMAN, J.; ARTER, J. A. Issues in item banking. **Journal of Educational Measurement,** 21, 315-330, 1984.

MILLS, C. N.; POTENZA, M. T.; FREMER, J. J.; WARD, W. C. **Computer-based testing: Building the foundation for future assessments.** Mahwah, NJ: Erlbaum, 2002.

MILLS, C. N.; STEFFEN, M. The GRE computer adaptive test: Operational issues. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic, 2000, p 75-99.

MILLS, C. N.; STOCKING, M. L. Practical issues in large-scale computerized adaptive testing. **Applied measurement in education**, 9(4), 1996.

MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada.** Belo Horizonte: Editora UFMG, 2007.

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

MISLEVY, R. J.; CHANG, H. Does adaptive testing violate local independence? **Psychometrika**, 65, 149-156, 2000.

MISLEVY, R. J.; WU, P.-K. **Inferring examinee ability when some items response are missing** (Research Report 88-48-ONR). Princeton, NJ: Educational Testing Service, 1988.

MOKKEN, R. J. **A Theory and Procedure of Scale Analysis.** Den Haag: Mouton, 1971.

MOLINA, J. G.; PAREJA, I.; SANMARTIN, J. Modeling item banking: Analysis and design of a computerized system. **Revista Electrónica de Metodología Aplicada**, V. 13 n. 2, p. 1-14, 2008.

MORAIS, J. H. M. Estudo de validação: padrões e dimensões da organização. In: IV Congresso Brasileiro de Avaliação Psicológica,

2009, Campinas. **Resumos - Painéis.** Campinas: IBAP - Instituto Brasileiro de Avaliação Psicológica, 2009.

MOREIRA JUNIOR, F. J. Aplicações da Teoria da Resposta ao Item (TRI) no Brasil. **Revista Brasileira de Biometria**, São Paulo, v.28, n.4 , p. 137-170, out.-dez. 2010.

MOREIRA JUNIOR, F. J.; BORTOLOTTI, S. L. V.; ANDRADE, D. F.; SOUZA JUNIOR, A. F. Avaliação da Satisfação através da utilização da Teoria da Resposta ao Item (TRI). In: X SEPROSUL - Semana de la Engeniería de la Producción Sudamericana, 2010, **Anais...** Santiago, Chile, 2010.

MOREIRA JUNIOR, F. J.; VARGAS, V. C. C.; ANDRADE, D. F. Avaliação dos Intangíveis em Indústrias Gaúchas e Catarinenses. In: X SEPROSUL - Semana de la Engeniería de la Producción Sudamericana, 2010, **Anais...** Santiago, Chile, 2010.

MOREIRA JUNIOR, F. J.; TEZZA, R.; ANDRADE, D. F. Teste Adaptativo Informatizado para Avaliar a Usabilidade de Sites. In: X SEPROSUL - Semana de la Engeniería de la Producción Sudamericana, 2010, **Anais...** Santiago, Chile, 2010.

MORENO, K. E. CAT-ASVAB Operational Test and Evaluation. In: SANDS, W. A.; WATERS, B. K.; MCBRIDE, J. R. (Eds.). **Computerized Adaptive Testing: From Inquiry to Operation** (p. 199–206). Washington, DC, USA: American Psychological Association, 1997.

MORENO, K. E.; SEGALL, D. O. Reliability and Construct Validity of CAT-ASVAB. In: SANDS, W. A.; WATERS, B. K.; MCBRIDE, J. R. (Eds.). **Computerized Adaptive Testing: From Inquiry to Operation.** Washington, DC, USA: American Psychological Association, 1997. p. 169–174.

MORLEY. J. Methods of Assessing Learning in Distance Education Courses. **Education at a Distance**, v. 13, n. 1, Janeiro, 2000.

MORRISON, C.; SUBHIYAH, R.; NUNGESTER, R. Item exposure rates for unconstrained and content-balanced computerized adaptive tests. **Annual Meeting of the American Educational Research Association**. San Francisco, CA, USA, 1995.

MORRISON, C. A.; NUNGESTER, R. J. Computerized Adaptive Testing in a Medical Licensure Setting: A Comparison of Outcomes

Under the One- and Two-Parameter Logistic Models. **Meeting of the National Council on Measurement in Education**, San Francisco, CA, USA, 1995.

MUCKLE, T.; BERGSTROM, B. A.; BECKER, K.; STAHL, J. A. Impact of altering randomization intervals on precision of measurement and item exposure. **Annual meeting of the American Educational Research Association**, Montreal, Canada, 2005.

MULDER, J.; VAN DER LINDEN, W. J. Multidimensional adaptive Testing with Kullback–Leibler Information Item Selection. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing.** Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p 77-102.

MULDER, J.; VAN DER LINDEN, W. J. Multidimensional adaptive testing with optimal design criteria for item selection. **Psychometrika**, v. 74, n. 2, p. 273-296, 2009.

MUÑIZ, J. La validez desde una óptica psicométrica. **Acta Comportamentalia**, 13, 9-20, 2005.

MUÑIZ, J. **Introducción a la teoría de respuesta a los ítems.** Madrid: Pirámide, 1997.

MUÑIZ, J.; HAMBLETON, R. Evaluación psicométrica de los tests informatizados. In: OLEA, J.; PONSODA, V.; PRIETO, G. (Eds.). **Tests informatizados: Fundamentos y aplicaciones.** Madrid: Pirámide, 1999, p. 23-52.

MURAKI, E. A generalized partial credit model: Application of an EM algorithm. **Applied Psychological Measurement,** n. 16, p. 159-176, 1992.

MURPHY, D. L.; DODD, B. G.; VAUGHN, B. K. A Comparison of Item Selection Techniques for Testlets. **Applied Psychological Measurement**, 34: p. 424-437, 2010.

NANDAKUMAR, R.; ROUSSOS, L. **CATSIB: a modified SIBTEST prodecure to detect differential item functioning in computerized adaptive tests** (Report research). Newtonwn: Law School Admission Council, 1997a.

NANDAKUMAR, R.; ROUSSOS, L. Evaluation of the CATSIB DIF Procedure in a Pretest Setting. **Journal of Educational and Behavioral Statistics**, Vol. 29, No. 2, p. 177–199, 2004.

NANDAKUMAR, R.; ROUSSOS, L. Validation of CATSIB to investigate DIF of CAT data. **Annual meeting of the AERA,** Chicago, 1997b.

NAVAS, M. J. Equiparación de puntuaciones. In: MUÑIZ, J. (Coord.). **Psicometría**. Madrid: Universitas, 1996, p. 293-369.

NERING, M. L. The distribution of indexes of person fit within the computerized adaptive testing environment. **Applied Psychological Measurement,** 21, p. 115-127, 1997.

NICEWANDER, W. A; THOMASSON, G. L. Some Reliability Estimates for Computerized Adaptive Tests. **Applied Psychological Measurement**, Vol. 23 No. 3, p. 239–247, 1999.

NICK, E. **Estatística e Psicometria.** Rio de Janeiro: J. Ozon + Editor, 1963.

NITKO, A. J.; HSU, T. A comprehensive Microcomputer System for Classroom Testing. **Journal of Educational Measurement**. Vol 21. Nº 4, p. 337-390, 1984.

NOGAMI, Y.; HAYASHY, N. A Japanese Adaptive Test of English as a Foreign Language: Developmental and Operational Aspects. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing**. Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p. 191-211.

NOJOSA, R. T. **Modelos Multidimensionais para a Teoria da Resposta ao Item.** Dissertação (Mestrado em Estatística) - Programa de Pós-Graduação em Estatística, Departamento de Estatística, Universidade Federal de Pernambuco, Recife, 2001.

NOJOSA, R. T. Teoria da resposta ao item (TRI) - Modelos Multidimensionais. **Estudos em Avaliação Educacional**, n. 25, p. 123-166, 2002.

NUNNALLY, J. C. **Introducción a la Medición Psicológica.** Buenos Aires: Editorial Paidós, 1970.

NUNES, C. H. S. S.; PRIMI, R. Impacto do Tamanho da Amostra na Calibração de Itens e Estimativa de Escores por Teoria de Resposta do Item. **Avaliação Psicológica,** v. 4, n. 2, p. 141-153, 2005.

NWEA. **Computerized adaptive testing user's manual.** Portland, USA: NWEA, 1997.

OECD. **PISA 2003 Data Analysis Manual.** OECD - Organization for Economic Co-operation and Development, 2005.

OLEA, J.; ABAD, F. J.; BARRADA, J. R. Tests Informatizados y Otros Nuevos Tipos de Tests. **Papeles del Psicólogo**, v. 31, n. 1, p. 94-107, 2010.

OLEA, J.; ABAD, F. J.; PONSODA, V.; XIMÉNEZ, M. C. Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: diseño y comprobaciones psicométricas. **Psicothema**, v. 16, n. 03, p. 519-525, 2004.

OLEA, J.; HONTANGAS, P. Tests informatizados de primera generación. In: OLEA, J.; PONSODA, V.; PRIETO, G. (Eds.). **Tests informatizados: Fundamentos y aplicaciones.** Madrid: Pirámide, 1999, p. 111-125.

OLEA, J.; PONSODA, V. Tests adaptativos informatizados. In: MUÑIZ, J. (Coord.). **Psicometría**. Madrid: Universitas, 1996, p. 730-783.

OLEA, J.; PONSODA, V. Tests informatizados y adaptativos informatizados: investigación en España. **RELIEVE**, v. 4, n. 2, 1998. Disponível em <http://www.uv.es/RELIEVE/v4n2/RELIEVEv4n2_0.htm>. Acesso em 11/09/2009.

OLEA, J.; PONSODA, V.; PRIETO, G. **Tests informatizados: Fundamentos y aplicaciones**. Madri, Espanha: Pirámide, 1999.

OLEA, J.; PONSODA, V.; REVUELTA, J. ; BELCHI, J. Propiedades psicométricas de un test adaptativo informatizado de vocabulario inglés. **Estudios de Psicología**, n. 55, p. 61-73, 1996.

OLEA, J.; REVUELTA, J.; XIMÉNEZ, C.; ABAD, F. J. Psychometric and psychological effects of review on computerized fixed and adaptive tests. **Psicológica,** n. 21, p. 157-173, 2000.

OLIVA, A. P. V. **Banco de itens para a Avaliação do Raciocínio Diagnóstico (BIARD).** 2008, 84 f. Tese (Doutorado em Enfermagem) - Escola de Enfermagem, Universidade de São Paulo. São Paulo, 2008.

OLIVEIRA, L. H. M. **Testes Adaptativos Sensíveis ao Conteúdo do Banco de Itens: Uma Aplicação em Exames de proficiência em Inglês para Programas de Pós-Graduação.** 2002. 220 f. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional) -

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2002.

OLIVEIRA, L. H. G. **Livro Didático e aprendizado de leitura no início do ensino fundamental. Rio de Janeiro,** 2007, 132 p. Dissertação (Mestrado em Educação) – Programa de Pós-Graduação em Educação, Departamento de Educação, Centro de Teologia e Ciências Humanas, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

OLMA, M. **Consciência Sobre Rodas: Primeira Habilitação.** 12a. ed. Porto Alegre: Águia, 2008.

OLSEN, J. B.; MAYNES, D. M.; SLAWSON, D. A. Comparison and equating of paper-administered, computer-administered and computerized adaptive tests of achievement. **Meetings of the American Educational Research Association,** San Francisco, CA, 1986.

O'NEILL T. R.; LUNZ M. E.; THIEDE, K. The Impact of Receiving the Same Items on Consecutive Computer Adaptive Test Administrations. **Journal of Applied Measurement**, v. 1, n. 2, 2000.

ORTNER, T. M.; CASPERS, J. Consequences of Test Anxiety on Adaptive Versus Fixed Item Testing. **European Journal of Psychological Assessment,** v. 27, n. 3, p. 157-163, 2011.

OSTERLIND, S. J. **Constructing Test Items: Multiple-Choice, Constructed- Response, Performance, and Other Formats.** 2a. edição. Kluwer Academic Publishers, 1997.

OSTINI, R.; NERING, M. L. **Polytomous Item Response Theory Models. Series: Quantitative Applications in the Social Sciences.** 1. ed. SAGE Publications, USA: 2006.

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

PACHECO JÚNIOR, W.; PEREIRA, V. L. D. V.; PEREIRA FILHO, H. V. **Pesquisa científica sem tropeços.** São Paulo: Atlas, 2007.

PAEK, P. **Recent Trends in Comparability Studies**. Pearson, 2005.

PAPANASTASIOU, E. C. A 'Rearrangement Procedure' for Scoring Adaptive Tests with Review Options. **National Council of Measurement in Education,** New Orleans, LA, 2002.

PAPANASTASIOU, E. C. Item review and the rearrangement procedure. Its process and its results. **Educational Research and Evaluation**, 11 (4), 303-321, 2005.

PAPANASTASIOU, E. C.; RECKASE, M. D. A 'rearrangement procedure' for scoring adaptive tests with review options. **International Journal of Testing**, 7(4), 1-21, 2007.

PARSHALL, C. G.; DAVEY, T.; NERING, M. L. Test Development Exposure Control for Adaptive Testing. **Annual meeting of the National Council on Measurement in Education,** San Diego, CA, USA, 1998.

PARSHALL, C. G.; DAVEY, T.; PASHLEY, P. J. Innovative item types for computerized testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic Publishers, 2000, p. 129-148.

PARSHALL, C. G.; HARMES, J. C.; DAVEY, T.; PASHLEY, P. J. Innovative item types for computerized testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing**. Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p. 215-230.

PARSHALL, C.; HARMES, J. C.; KROMREY, J. D. Item exposure control in computer-adaptive testing: the use of freezing to augment stratification. **Florida Journal of Educational Research,** 40, 28-52, 2000.

PARSHALL, C. G.; HOGARTY, K. Y.; KROMREY, J. D. Item exposure in adaptive tests: an empirical investigation of control strategies. **Annual meeting of the Psychometrics Society**, Lawrence, KS, 1999.

PARSHALL, C. G.; KROMREY, J. D. Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect. **Annual Meeting of the American Educational Research Association**, Atlanta, GA USA, 1993.

PARSHALL, C. G.; KROMREY, J. D.; HARMES, J. C.; SENTOVICH, C. Nearest Neighbors, Simple Strata, and Probabilistic Parameters: An Empirical Comparison of Methods for Item Exposure Control in CATs. **Annual Meeting of the National Council on Measurement in Education,** Seattle, WA, 2001.

PARSHALL, C. G.; SPRAY, J. A.; KALOHN, J. C; DAVEY, T. **Practical Considerations in Computer-Based Testing.** Springer-Verlag: New York, 2002.

PARTCHEV, I. **Irtoys: Simple interface to the estimation and plotting of IRT models.** R package version 0.1.4, 2011.

PASQUALI, L. **Teoria e Métodos de Medida em Ciências do Comportamento.** Brasília: Laboratório de Pesquisa em Avaliação e Medida / Instituto de Psicologia / UnB: INEP, 1996.

PASQUALI, L. Princípios de elaboração de escalas psicológicas. **Revista de Psiquiatria Clínica,** n. 25, v. 5, Edição Especial, p. 206-213, 1998.

PASSOS, V. L.; BERGER, M. P. F.; TAN, F. E. Test design optimization in CAT early stage with the nominal response model. **Applied Psychological Measurement**, 31, p. 213-232, 2007.

PASSOS, V. L.; BERGER, M. P. F.; TAN, F. E. The D-Optimality Item Selection Criterion in the Early Stage of CAT: A Study With the Graded Response Model. **Journal of Educational and Behavioral Statistics**, V. 33, N. 1, p. 88–110, 2008.

PASTOR, D. A.; CHIANG, C.; DODD, B. G.; YOCKEY, R. Performance of the Sympson-Hetter exposure control algorithm with a polytomous item bank. **Annual meeting of American Educational Research Association,** Montreal, Canadá, 1999.

PASTOR, D. A.; DODD, B. G.; CHANG, H. H. A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. **Applied Psychological Measurement**, 26, 147-163, 2002.

PENFIELD, R. D. Applying Bayesian item selection approaches to adaptive tests using polytomous items. **Applied Measurement in Education**, 19, p. 1-20, 2006.

PENFIELD, R. D. Estimating the Standard Error of the Maximum Likelihood Ability Estimator in Adaptive Testing Using the Posterior-Weighted Test Information Function. **Educational and Psychological Measurement,** V. 67, N. 6, p. 958-975, 2007.

PEREIRA, L. M. **Um método para medir o grau da maturidade das empresas para implantar o comércio eletrônico B2B através da Teoria da Resposta ao Item.** 2007. Dissertação (Mestrado em

Engenharia Elétrica e de Computação). Programa de Pós-Graduação em Engenharia Elétrica e de Computação, Escola de Engenharia Elétrica, Universidade Federal de Goiás, Goiânia, 2007.

PEREIRA, V. R. **Métodos Alternativos no Critério Brasil para Construção de Indicadores Sócio-econômico: Teoria da Resposta ao Item.** 2004. 103 f. Dissertação (Mestrado em Engenharia Elétrica) – Programa de Pós-Graduação em Engenharia Elétrica, Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2004.

PETERSEN, M. A. et al. Development of computerised adaptive testing (CAT) for the EORTC QLQ-C30 dimensions – General approach and initial results for physical functioning. **European Journal of Cancer,** v. 46, n. 8, p. 1352-1358, 2010.

PETERSEN, M. A.; GROENVOLD, M.; AARONSON, N.; FAYERS, P.; SPRANGERS, M.; BJORNER, J. B. Multidimensional computerized adaptive testing of the EORTC QLQ-C30: Basic developments and evaluations. **Quality of Life Research,** n. 15, p. 315–329, 2006.

PHANKOKKRUAD, M.; WORARATPANYA, K. Web Service Architecture for Computer-Adaptive Testing on e-Learning. **International Journal of Behavioral, Cognitive, Educational and Psychological Sciences,** v. 1, p. 43-47, 2009.

PINA, J. A. L.; MONTESINOS, M. D. H. Fitting Rasch Model using Appropriateness Measure Statistics. **The Spanish Journal of Psychology,** v. 8, n. 1, p. 100-110, 2005.

PITKIN, A. K.; VISPOEL, W. P. Differences Between Self-Adapted and Computerized Adaptive Tests: A Meta-Analysis. **Journal of Educational Measurement**, v. 38, n. 3, p. 235-247, 2001.

PITON-GONÇALVES, J. **A Integração de Testes Adaptativos Informatizados e Ambientes Computacionais de Tarefas para o Aprendizado do Inglês Instrumental.** 2004. 142 f. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional), Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2004.

PITON-GONÇALVES, J.; MONZÓN, A. J. B.; ALUÍSIO, S. M. Métodos de avaliação informatizada que tratam o conhecimento parcial do aluno e geram provas individualizadas. In: XX Simpósio Brasileiro

de Informática na Educação. **Anais do XX Simpósio Brasileiro de Informática na Educação,** Florianópolis: UFSC, 2009.

PITONIAK, M. **Automatic item generation methodology in theory and practice** (Center for Educational Assessment Research Report No. 444). Amherst, MA: University of Massachusetts, School of Education, 2002.

POMMERICH, M. Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. **Journal of Technology, Learning, and Assessment**, v. 2, n. 6, 2004.

POMMERICH, M. The effect of using item parameters calibrated from paper administrations in computer adaptive test administrations. **Journal of Technology, Learning, and Assessment**, v. 5, n. 7, 2007.

POMPLUN, M.; FREY, S.; BECKER, D. F. The score equivalence of paper and computerized versions of a speeded test of reading comprehension. **Educational and Psychological Measurement**, 62(2), 337-354, 2002.

PONSODA, V. Overview of the computerized adaptive testing special section. **Psicológica,** 21, p. 115-120, 2000.

PONSODA, V.; HONTANGAS, P.; OLEA, J.; REVUELTA, J.; ABAD, F. J.; CARMEN XIMÉNEZ, C. Los tests adaptativos informatizados: investigación actual. **Metodología de las Ciencias del Comportamiento,** Suplemento 2004, p. 505–510, 2004.

PONSODA, V.; OLEA, J.; ALCÁZAR. APT-SYSTEM: Procedimientos para la evaluación psicológica adaptativa y autoadaptativa. **Capital Humano,** Nº 65, p. 28-34, 1994.

PONSODA, V.; OLEA, J.; REVUELTA, J. ADTEST: A computer adaptive test based on the maximum information principle. **Educational and Psychological Measurement,** v. 54, n. 3, p. 680-686, 1994.

PONSODA, V.; WISE, S. L.; OLEA, J.; REVUELTA, J. An Investigation of Self-Adapted Testing in a Spanish High School Population. **Educational and Psychological Measurement**, vol. 57 no. 2, p. 210-221, 1997.

POWERS, D. E. **Test Anxiety and Test Performance: Comparing Paper-Based and Computer Adaptive Versions of the GRE® General Test.** Research Report RR-99-15. Princeton, NJ, USA: ETS - Educational Testing Service, 1999.

PRIETO, G.; DELGADO, A. Construcción de ítems. In: MUÑIZ, J. (Coord.). **Psicometría**. Madrid: Universitas, 1996, p. 105-138.

PRIMI, R. Avanços na Interpretação de Escalas com a Aplicação da Teoria de Resposta ao Item. **Avaliação Psicológica,** v. 3, n. 1, p. 53-58, 2004.

PROMISSOR. **Using the CATGlobal**. Report Center. Promissor, 2003.

QUESTION MARK CORPORATION. **Question Mark for Windows. UserGuide.** Stanford, CT: Autor, 1997.

QUESTION MARK CORPORATION. **QM Perception. Gettintg started Guide**, 1998. Disponível em: <http://www.questionmark.com/perception/help/v5/product_guides/gs/Content/v5_getting_started.pdf>. Acesso em:30/06/2011.

RASCH, G. **Probabilistic Models for Some Intelligence and Attainment Tests.** Copenhagen: Danish Institute for Educational Research, 1960.

RAÎCHE, G.; BLAIS, J.-G. Practical considerations about expected a posteriori estimation in adaptive testing: Adaptive a priori, adaptive correction for bias, and adaptive integration interval. **11th International Objective Measurement Workshop**, New- Orleans, USA, 2002.

RAÎCHE, G.; BLAIS, J.-G. Considerations About Expected a Posteriori Estimation in Adaptive Testing: Adaptive a Priori, Adaptive Correction for Bias, and Adaptive Integration Interval. **Journal of Applied Measurement,** v. 10, n. 2, 2009.

RAÎCHE, G.; BLAIS, J.-G. SIMCAT 1.0: A SAS Computer Program for Simulating Computer Adaptive Testing. **Applied Psychological Measurement**, Vol. 30 No. 1, p. 60–61, 2006.

RAÎCHE G.; BLAIS J.-G.; RIOPEL M. A SAS Solution to Simulate a Rasch Computerized Adaptive Test. **Rasch Measurement Transactions, v.** 20:2 p. 1061, 2006.

REBOLLO, P.; GARCÍA-CUETO, E.; ZARDAÍN, J.C.; MARTÍNEZ, I.; ALONSO, J.; FERRER, M.; MUÑIZ, J. Desarrollo del CAT-Health, primer test adaptativo informatizado para la evaluación de la calidad de vida relacionada con la salud en España. **Medicina Clínica**, v. 133, n. 7, p. 241-251, 2009.

RECKASE, M. D. A procedure for decision making using tailored testing. In D. J. WEISS (Ed.) **New horizons in testing**, New York (USA): Academic Press, 237-255, 1983.

RECKASE, M. D. Item pool design for computerized adaptive tests. **National Council on Measurement in Education,** Chicago, IL, 2003.

RECKASE, M. D. Multidimensional item response theory. In: RAO, C. R.; SINHARAY, S. (Eds.), **Handbook of statistics**, vol. 26 (pp. 607–642). Amsterdam: Elsevier, 2007.

RECKASE, M. D. **Multidimensional Item Response Theory.** Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2009.

RECKASE, M. D. Unifactor latent trait models applied to multifactor tests: Results and implications. **Journal of Educational Statistics**, 4, 207-230, 1979.

RECKASE, M. D.; HE, W. **Ideal item pool design for the NCLEX-RN exam.** Michigan State University, East Lansing, MI, USA, 2005.

REESE, L. M. **Impact of Local Dependence on Item Response Theory Scoring in CAT.** Computerized testing Report 98-08. Law School Admission Council – LSAC: Newtown, PA, USA, 1999.

REESE, L. M.; SCHNIPKE, D. L.; LUEBKE, S. W. Incorporating content constrains into a multistage adaptive testlet design. **Annual meeting of the AERA**, Chicago, USA, 1997.

REEVE, B. B.; FAYERS, P. Applying item response theory modelling for evaluating questionnaire item and scale properties. **Oxford,** p. 55-73, 2005.

REEVE, B. B. et al. Psychometric Evaluation and Calibration of Health-Related Quality of Life Item Banks Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). **Medical Care,** V. 45, N. 5 Supl. 1, p. S22-S31, 2007.

REISE, S. P.; DUE, A. M. The influence of test characteristics on the detection of aberrant response patterns. **Applied Psychological Measurement**, 15, p. 217-226, 1991.

REISE, S. P.; HENSON, J. M. Computerization and Adaptive Administration of the NEO PI-R.. **Assessment,** v. 7, n. 4, p. 347-364, 2000.

REISE, S. P.; YU, J. Parameter recovery in the graded response model using MULTILOG. **Journal of Educational Measurement**, n. 27, p. 133-144, 1990.

REISE, S. P.; MORIZOT, J.; HAYS, R. D. The role of the bifactor model in resolving dimensionality issues in health outcomes measures **Quality of Life Research,** n. 16, p.19–31, 2007.

RENOM, J. **Tests adaptativos computerizados: Fundamentos y aplicaciones.** Barcelona: PPU, 1993.

RENOM, J.; DOVAL, E. Tests adaptativos informatizados: Estructura y desarrollo. In: OLEA, J.; PONSODA, V.; PRIETO, G (Eds.). **Tests informatizados: Fundamentos y aplicaciones.** Madrid: Pirámide, 1999, p. 127-162.

REVICKI, D. A.; CELLA, D. F. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. **Quality of Life Research**, v. 6, p. 595–600, 1997.

REVUELTA, J. Estimación de habilidad mediante ítems isomorfos. Efectos en la fiabilidad de las puntuaciones. **Psicothema**, v. 12, n. 2, p. 303-307, 2000.

REVUELTA, J. Estimating Ability and Item-Selection Strategy in Self-Adapted Testing: A Latent Class Approach. **Journal of Educational and Behavioral Statistics**, Vol. 29, No. 4, p. 379–396, 2004.

REVUELTA, J. Self Adaptive Testing. **International Encyclopedia of Education**, Third Edition, p. 148-152, 2010.

REVUELTA, J.; PONSODA, V. A comparison of item exposure control methods in computerized adaptive testing. **Journal of Educational Measurement**, 26(2), 144-163, 1998a.

REVUELTA, J.; PONSODA, V. **Fundamentos de Estadística**. Madri, Espanha: UNED, 2001.

REVUELTA, J.; PONSODA, V. Generación automática de ítems. In: OLEA, J.; PONSODA, V; PRIETO, G (Eds.). **Tests informatizados. Fundamentos y aplicaciones**. (pp. 227-250).Madri, Espanha: Pirámide, 1999.

REVUELTA, J.; PONSODA, V. Metodos sencillos para el control de las tasas de exposicion en tests adaptativos informatizados [Simple

methods for item exposure control in CATs]. **Psicologica**, 17, 161- 172, 1996.

REVUELTA, J.; PONSODA, V. Una solución a la estimación inicial en los tests adaptativos informatizados. **Revista Electrónica de Metodología Aplicada,** v. 2, n. 2, p. 1-6, 1997. Disponível em <http://www.psico.uniovi.es/REMA/v2n2/a1/>. Acesso em 21/11/2009.

REVUELTA, J.; PONSODA, V. Un test adaptativo informatizado de Análisis Lógico Basado en la Generación Automática de Ítems. **Psicothema**, v. 10, n. 03, p. 709-716, 1998b.

REVUELTA, J.; PONSODA, V.; ABAD, F. J. **Modelos Politomicos.** Madri, Espanha: La Muralla, 2006.

REVUELTA, J.; PONSODA, V.; OLEA, J. Métodos para el control de las tasas de exposición en tests adaptativos informatizados. **RELIEVE**, vol. 4, n. 2, 1998. Disponível em <http://www.uv.es/RELIEVE/v4n2/RELIEVEv4n2_4.htm>. Acesso em 10/09/2009.

RICHARDSON, R. J. **Pesquisa Social: Métodos e Técnicas.** São Paulo: Atlas, 2010.

RICHAUD, M. C. Desarrollos del analisis factorial para el estudio de item dicotomicos y ordinales. **Interdisciplinaria**, Buenos Aires, v. 22, n. 2, p. 237-251, 2005.

RILEY, B. B.; CONRAD, K. J.; BEZRUCZKO, N.; DENNIS, M. L. Relative Precision, Efficiency and Construct Validity of Different Starting and Stopping Rules for a Computerized Adaptive Test: The GAIN Substance Problem Scale. **Journal of Applied Measurement**, v. 8, n. 1, p. 48-64, 2007.

RILEY, B. B.; DENNIS, M. L.; CONRAD, K. J. A Comparison of Content-Balancing Procedures for Estimating Multiple Clinical Domains in Computerized Adaptive Testing: Relative Precision, Validity, and Detection of Persons With Misfitting Responses. **Applied Psychological Measurement** 34: p. 410-423, 2010.

RIZOPOULOS, D. **ltm: Latent Trait Models under IRT.** R package version 0.9-5, 2010.

ROBERTS, J. S.; DONOGHUE, J. R.; LAUGHLIN, J. E. A General model for unfolding unidimensional polytomous responses using item

response theory. **Applied Psychological Measurement,** v. 1, n. 24, p. 3-32, 2000.

ROBERTS, J. S.; LIN, Y.; LAUGHLIN, J. E. Computerized Adaptive Testing With the Generalized Graded Unfolding Model. **Applied Psychological Measurement,** v. 25, n. 2, p. 177-196, 2001.

ROCKLIN, T.; O'DONNELL, A. Self – adapted testing: A performance – improving variant of computerized adaptive testing. **Journal of Educational Psychology**, 79 (3), p. 315 – 319, 1987.

ROCKLIN, T.; O´DONNELL, A.; HOLST, P. Effects and underlying mechanisms of self-adaptive testing. **Journal of Educational Psychology,** 87 (1), 103 – 116, 1995.

RODRIGUES, M. M. M. **Avaliação Educacional Sistêmica na Perspectiva dos Testes de Desempenho e seus Resultados: Estudos do SAEB.** Tese (Doutorado em Psicologia) - Programa de Pós-Graduação, Departamento de Psicologia Social e do Trabalho, Instituto de Psicologia, Universidade de Brasília, Brasília, 2007.

ROMERO, C. et al. Herramienta Autor para la Gestión de Tests Informatizados dentro del Sistema AHA! **Revista Iberoamericana de Informática Educativa**, n. 3, p. 43-54, 2006.

ROOS, L. L.; WISE, S. L.; PLAKE, B. S. The effects of feedback in computerized adaptive and self-adapted tests. **Annual Meeting of the National Council on Measurement in Education.** San Francisco, CA, USA, 1997.

ROOS, L. L.; WISE, S. L.; PLAKE, B. S. The role of item feedback in self-adapted testing. **Educational and Psychological Measurement**, 57, 85-98, 1997.

RORIZ JUNIOR, G. S. **Da Maturidade da Gestão Industrial para o Comércio B2B em Empresas Farmacêuticas de Goiás.** 2008. 134 f. Dissertação (Mestrado em Tecnologia Farmacêutica) - Programa de Pós-Graduação em Gestão, Pesquisa e Desenvolvimento em Tecnologia Farmacêutica, Universidade Católica de Goiás, Universidade Estadual de Goiás, Centro Universitário de Anápolis, Goiânia, 2008.

ROTOU, O.; PATSULA, L.; MANFRED, S.; RIZAVI, S. Comparison of Multi-stage Tests with Computerized Adaptive and Paper and Pencil Tests. **American Educational Research Association (AERA) and the**

National Council on Measurement in Education (NCME), Chicago, IL, USA, 2003.

ROUSSOS, L.; STOUT, W. A Multidimensionality-Based DIF Analysis Paradigm. **Applied Psychological Measurement**,20: 355-371, 1996.

RUBIO, V.; SANTACREU, J. **TRASI: Test adaptativo informatizado para la evaluación del razonamiento secuencial y la inducción como factores de la habilidad intelectual general.** Madrid: TEA ediciones, 2003.

RUDNER, L. An examination of decision-theory adaptive testing procedures. In **Annual Meeting of the American Educational Research Association proceedings**, New Orleans (USA):AERA, 2002.

RUDNER, L. M. An On-line, Interactive, Computer Adaptive Testing Tutorial, 1998. Disponível em <http://EdRes.org/scripts/cat>. Acesso em: 28/07/2008.

RUDNER, L. M. Implementing the Graduate Management Admission Test Computerized Adaptive Test. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing**. Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p. 151-166.

RULISON, K. L.; LOKEN, E. I've Fallen and I Can't Get Up: Can High-Ability Students Recover From Early Mistakes in CAT? **Applied Psychological Measurement**, V. 33, N. 2, p. 83–101, 2009.

SALCEDO, P.; PINNINGHOFF, M. A.; CONTRERAS, R. Computerized Adaptive Tests and Item Response Theory on a Distance Education Platform. In: MIRA, J.; ÁLVAREZ, J. R. (Eds.):**Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach.** Lecture Notes in Computer Science, 2005, Volume 3562, p. 613-621, 2005.

SAMEJIMA, F. A comment on Birnbaum's three-parameter logistic model in latent trait theory. **Psychometrika**, 38, 221–233, 1973.

SAMEJIMA, F. A. Estimation of latent ability using a response pattern of graded scores. **Psychometric Monograph,** n. 17, 1969.

SAMEJIMA, F. Expansion of Warm's weighted likelihood estimator of ability for threeparameter logistic model to general discrete responses. **Annual Meeting of the NCME**, San Diego, USA, 1998.

SANDS, W. A.; WATERS, B. K. Introduction to ASVAB and CAT. In: SANDS, W. A; WATERS, B. K.; MCBRIDE, J. R. **Computerized Adaptive Testing: from inquiry to operation.** Washington, USA: American Psychological Association, 1997.

SANDS, W. A; WATERS, B. K.; MCBRIDE, J. R. **Computerized Adaptive Testing: from inquiry to operation.** Washington, USA: American Psychological Association, 1997.

SANTOS, P. S.; MENEZES, I. G.; BASTOS, A. V. B.; ALVES FILHO, A. P.; MENDONÇA, R. F. Comprometimento organizacional: um estudo empírico sobre a dimensionalidade do construto. In: IV Congresso Brasileiro de Avaliação Psicológica, 2009, Campinas. **Resumos - Painéis**. Campinas: IBAP - Instituto Brasileiro de Avaliação Psicológica, 2009.

SASSI, G. P.; AMARAL, L.; CURI, M. Teste Adaptativo Informatizado para Estatística. **I CONBRATRI - I Congresso Brasileiro de Teoria de Resposta ao Item**, apresentação de poster, Florianópolis, 2009.

SAWAKI, Y. Comparability of conventional and computerized tests of reading in a second language. **Language Learning & Technology**, v. 5, n. 2, p. 38-59, 2001.

SCHAEFFER, G. A.; REESE, C. M.; STEFFEN, M.; MCKINLEY, R. L.; MILLS, C. N. **Field test of a computer-based GRE General Test** (ETS Report No. RR-93-07). Princeton, NJ: Educational Testing Service, 1993.

SCHAEFFER, G. A.; STEFFEN, M.; GOLUB-SMITH, M. L.; MILLS, C. N.; DURSO, R. **The introduction and comparability of the computer-adaptive GRE General Test** (Research Rep. No. 95-20). Princeton NJ: Educational Testing Service, 1995.

SCHAEFFER, G. A. et al. **Comparability of Paper-and-Pencil and Computer Adaptive Test Scores on the GRE General Test.** (Research Rep. No. 98-38). Princeton NJ: Educational Testing Service, 1998.

SCHERBAUM, C. S; FINLINSON, S.; TAMANINI, K. Applications of item response theory to measurement issues in leadership research. **The Leadership Quarterly**, V. 17, N. 4, p. 366-386, 2006.

SCHNIPKE, D. L.; GREEN, B. F. A comparison of item selection routines in linear and adaptive tests. **Journal of Educational Measurement,** n. 3, p. 227-242, 1995.

SCHNIPKE, D. L.; SCRAMS, D. J. Modeling item response times wtih a twostate mixture model: A new method of measuring speededness. **Journal of Educational Measurement**, 34, 213-232, 1997.

SCHOONMAN, W. **An applied study on computerized adaptive testing.** Amsterdam, Holanda: Swets & Zeitlinger, 1989.

SCHWARTZ, C.; WELCH, G.; SANTIAGO-KELLEY, P.; BODE, R.; SUN, X. Computerized adaptive testing of diabetes impact: A feasibility study of Hispanics and non-Hispanics in an active clinic population. **Quality of Life Research,** 15:1503–1518, 2006.

SEGALL, D. O. A Sharing Item Response Theory Model for Computerized Adaptive Testing. **Journal of Educational and Behavioral Statistics,** v. 29, n. 4, 439-460, 2004.

SEGALL, D. O. An Adaptive Exposure Control Algorithm for Computerized Adaptive Testing using a Sharing Item Response Theory Model. **National Council on Measurement in Education**. Chicago, IL, USA, 2003.

SEGALL, D. O. Computerized Adaptive Testing. **Encyclopedia of Social Measurement**, Elsevier Inc. v. 1, p. 429-438, 2005.

SEGALL, D. O. Equating the CAT–ASVAB. In: SANDS, W. A; WATERS, B. K.; MCBRIDE, J. R. (Eds.). **Computerized adaptive testing: From inquiry to operation** (pp. 181–198). Washington, DC, USA: American Psychological Association, 1997.

SEGALL, D. O. General Ability Measurement: An application of multidimensional Item Response Theory. **Psychometrika**, v. 66, n. 1, p. 79-97, 2001.

SEGALL, D. O. Multidimensional Adaptive Testing. **Psychometrika**, 61, 331-354, 1996.

SEGALL, D. O. Principles of Multidimensional Adaptive Testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic Publishers, p. 163-182, 2000.

SEGALL, D. O. Principles of Multidimensional Adaptive Testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing.** Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p 57-76.

SEGALL, D. O. **Score equating verification analyses of the CAT-ASVAB.** Briefing presented to the Defense Advisory Committee on Military Personnel Testing. Williamsburg, VA, USA, 1993.

SEGALL, D. O.; MORENO H. E. Development of CAT-ASVAB. In: DRASGOW, F.; OLSON-BUCHANAN, J. B. (Eds.). **Innovations in computerized assessment.** Mahwah, NJ: LEA, 1999. p. 35-66.

SEGALL, D. O.; MORENO, K. E.; HETTER, R. D. Item pool development and evaluation. In: SANDS, W. A; WATERS, B. K.; MCBRIDE, J. R. (Eds.). **Computerized adaptive testing: From inquiry to operation** (pp. 117-130). Washington, DC: American Psychological Association, 1997.

SEGALL, D. O.; MORENO, K. E.; KIECKHAEFER, W. F.; VICINO, F. L.; MCBRIDE, J. R. Validation of the experimental CAT-ASVAB system. In: SANDS, W. A; WATERS, B. K.; MCBRIDE, J. R. (Eds.). **Computerized Adaptive Testing: from inquiry to operation.** Washington, USA: American Psychological Association, 1997.

SHANNON, C. E. A mathematical theory of communication. Part III. **Bell System Thechnical Journal**, n. 28, p. 623-656, 1949.

SHEEHAN, K.; LEWIS, C. Computerized mastery testing with nonequivalent testlets. **Applied Psychological Measurement**, 16, 65-76, 1992.

SHERMIS, M. D.; MZUMARA, H.; BROW, M.; LILLING, C. Computerized adaptive testing though the World Wide Web. **Annual meeting of AERA**, Chicago, USA, 1997.

SIERRA-MATAMOROS, F. A. et al. Test Adaptativos Informatizados. **Avances en Medición**, 5, p. 157–162, 2007.

SIJTSMA, K.; MOLENAAR, I. W. **Introduction to nonparametric item response theory.** Volume 5, Sage Publications, 2002.

SILVA, H. **Modelo Logístico Multidimensional da Teoria da Resposta ao Item.** Universidade da Beira Interior, Covilhã, Portugal, 2005.

SILVA, H. T. A. O.; CURI, M. Comparação de Dois Testes Adaptativos Informatizados para Depressão. In: **XI EMR – Escola de Modelos de Regressão.** Recife, 2009.

SIMMS, L. J., CLARK, L. J. Validation of a computerized adaptive version of the Schedule of Non-Adaptive and Adaptive Personality (SNAP). **Psychological Assessment**, 17(1), 28-43, 2005.

SMITH, R. M. Reporting candidate performance on computer-adaptive tests. **Rasch Measurement Transactions,** v. 8:1 p.344-5, 1994.

SMITS, N.; CUIJPERS, P.; VAN STRATEN, A. Applying computerized adaptive testing to the CES-D scale: A simulation study. **Psychiatry Research**, v. 188, n. 1, p. 147-155, 2011.

SOARES, T. M. Utilização da teoria da resposta ao item na produção de indicadores sócio-econômicos. **Pesquisa Operacional**. Rio de Janeiro, v. 25, n. 1, p. 83-112, 2005.

SOUZA, S. Z. 40 Anos de Contribuição à Avaliação Educacional. **Estudos em Avaliação Educacional**, v. 16, n. 31, jan./jun. 2005.

SPRAY, J. A.; ABDEL-FATTAH, A. A.; HUANG, C.; LAU, C. A. **Unidimensional approximations for a computerized test when the item pool and latent space are multidimensional** (Research Report 97-5). Iowa City, Iowa: ACT, Inc, 1997.

SPRAY, J. A.; PARSHALL, C. G.; THOMAS, L. Calibration of CAT items administered online for classification: assumption of local independence. **Annual meeting of the Psychometric Society,** Gatlinburg, TN, USA, 1997.

SPRAY, J. A.; RECKASE, M. D. The selection of test items for decision making with a computer adaptive test. **Annual meeting of the National Council on Measurement in Education**, New Orleans LA, USA, 1994.

SPRAY, J. A.; RECKASE, M. D. Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. **Journal of Educational and Behavioral Statistics**, 21, p. 405–414, 1996.

STARK, S.; CHERNYSHENKO, O. S.; GUENOLE, N. Can Subject Matter Experts' Ratings of Statement Extremity Be Used to Streamline the Development of Unidimensional Pairwise Preference Scales? **Organizational Research Methods**, v. 14 n. 2, p. 256-278, 2011.

STICHA, P. J.; BARBER, G. **CAT-ASVAB Prototype Internet Delivery System: Final Report** (FR-03-46). Alexandria, VA: Human Resources Research Organization, 2003.

STOCKING, M. L. An alternative method for scoring adaptive tests. **Journal of Educational and Behavioral Statistics**, 21, 365-389, 1996.

STOCKING, M. L. **Controlling item exposure rates in a realistic adaptive testing paradigm** (Tech. Rep. No. 93-2). Princeton, NJ, USA: Educational Testing Service, 1993.

STOCKING, M. L. Revising item responses in computerized adaptive tests: A comparison of three models. **Applied psychological measurement**, 21 (2), 129-142, 1997.

STOCKING, M. L. **Three practical issues for modern adaptive testing item pools** (No. ETSRR-94-5): Educational Testing Service, Princeton, NJ, USA, 1994.

STOCKING, M. L.; LEWIS, C. H. **A new method for controlling item exposure in computer adaptive testing** (Research Report 95-25). Princeton, NJ, USA: Educational Testing Service, 1995.

STOCKING, M. L.; LEWIS, C. H. Controlling item exposure conditional on ability in computerized adaptive testing, **Journal of Educational and Behavioral Statistics**, 23, 57-75, 1998

STOCKING, M. L.; LEWIS, C. H. Methods of controlling the exposure of items in CAT. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic, 2000, p 163-182.

STOCKING, M. L.; SMITH, R.; SWANSON, L. **An Investigation of Approaches to Computerizing the GRE Subject Tests.** GRE Board Professional Report No. 93-08P. Princeton, USA: ETS - Educational Testing Service, 2000.

STOCKING, M. L.; SWANSON, L. A method for severely constrained item selection in adaptive testing. **Applied Psychological Measurement,** n. 17, p. 277-292, 1993.

STOCKING, M. L.; SWANSON, D. Optimal design of item banks for computerized adaptive testing. **Applied Psychological Measurement**, 22, 271–279, 1998.

STONE, G. E.; LUNZ, M. E. The effect of review on the psychometric characteristics of computerized adaptive tests. **Applied Measurement in Education**, 7, 211-222, 1994.

STOUT, W.; HABING, B.; DOUGLAS, J.; KIM, H. R.; ROUSSOS, L.; ZHANG, J. Conditional Covariance-Based Nonparametric Multidimensionality Assessment. **Applied Psychological Measurement**, 20: 331-354, 1996.

STRICKER, L. J.; WILDER, G. Z. **Examinees' Attitudes About the TOEFL-CBT, Possible Determinants, and Relationships With Test Performance.** TOEFL – Technical Report. New Jersey: ETS, 2001.

SUÁREZ, J. Student Evaluation through Membership Functions in CAT Systems. Revista Mexicana de Física, 49 (4) p. 371-378, 2003.

SUMBLING, M.; SANZ, P.; VILADRICH, M. C.; DOVAL, E.; RIERA, L. Development of a Multiple-Component CAT for Measuring Foreign Language Proficiency (SIMTEST). In WEISS, D. J. (Ed.). **Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing**, 2007.

SUKAMOLSON, S. Computerized Test/Item Banking and Computerized Adaptive Testing for Teachers and Lecturers. **Information Technology and Universities in Asia – ITUA,** 2002.

SWANSON, D.; STOCKING, M. L. A method for severely constrained item selection in adaptive testing. **Applied Psychological Measurement**, 17, 277–292, 1993a.

SWANSON, D.; STOCKING, M. L. A model and heuristic for solving very large item selection problems. **Applied Psychological Measurement**, 17(2), 151-166, 1993b.

SYMPSON, J. B.; HETTER, R. D. Controlling item exposure rates in computerized adaptive testing In **Proceedings of the 27th annual meeting of the military testing association** (pp. 973-977). San Diego, CA, USA: Navy Personnel Research and Development Center, 1985.

TANG, K. L.; EIGNOR, D. R. **Concurrent Calibration of Dichotomously and Polytomously Scored TOEFL® Items Using IRT Models**. TOEFL – Technical Report TR-5. New Jersey: ETS, 1997.

TANG, K. L.; WAY, W. D.; CAREY, P. A. **The Effect of Small Calibration Sample Sizes on TOEFL® IRT-Based Equating.** TOEFL – Technical Report TR-7. New Jersey: ETS, 1993.

TAO, Y.-H.; WU, Y.-L.; CHANG, H.-Y. A Practical Computer Adaptive Testing Model for Small-Scale Scenarios.**Educational Technology & Society**, 11(3), p. 259–274, 2008.

TEJADA, A. J. R. Pasado, presente y futuro de los Tests Adaptativos Informatizados: entrevista con Isaac I. Bejar. **Psicothema**, v. 13, n. 04, p. 685-690, 2001.

TEZZA, R. **Proposta de um construto para medir usabilidade em site de e-commerce utilizando a Teoria da Resposta ao Item.** 2009. 139 f. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2009.

TEZZA, R.; BORNIA, A. C. Teoria da Resposta ao Item: Vantagens e Oportunidades para a Engenharia de Produção. In: XXIX Encontro Nacional de Engenharia de Produção – ENEGEP, 2009, Salvador. **Anais...** Salvador: ABEPRO, 2009.

TEZZA, R.; BORNIA, A. C.; ANDRADE, D. F. Measuring web usability using item response theory: Principles, features and opportunities. **Interacting with Computers**. Volume 23, Issue 2, p. 167-175, 2011.

TEZZA, R.; BORNIA, A. C.; MOREIRA JUNIOR, F. J. Avaliação de usabilidade em sites de e-commerce: Uma aplicação da Teoria da Resposta ao Item. IX USIHC - Congresso Internacional de Ergonomia e Usabilidade de Interfaces Humano-Computador, **Anais...** Curitiba - PR, 2009.

THEUNISSEN, J. J. J. M. Binary programming and test design. **Psychometrika**, 50, 411-420, 1985.

THEUNISSEN, J. J. J. M. Optimization algorithms in test design. **Applied Psychological Measurement,** 10, 381-389, 1986.

THISSEN, D. Measurement precision and "reliability": Some considerations of metrics and stopping rules in CAT. **Proceedings of the 27th Annual Conference of the Military Testing Association**. San Diego: NPRDC, 1986.

THISSEN, D. M.; MISLEVY, R. J. Testing algorithms. In H. WAINER (Ed.) **Computerized adaptive testing: a primer** (second edition), Mahwah, New Jersey (USA): Lawrence Erlbaum Associates, 101-132, 2000.

THISSEN, D.; REEVE, B. B.; BJORNER, J. B.; CHANG, C.-H. Methodological issues for building item banks and computerized adaptive scales. **Quality of Life Research**, 16, 109-116, 2007.

THISSEN, D.; STERNBERG, L.; MOONEY, J. A. Trace lines for testlets: A use of multiplecategorical response models. **Journal of Educational Measurement**, 26, 247-260, 1989.

THISSEN, D.; WAINER, H. **Test Scoring.** NJ: Lawrence Erlbaum, 2001.

THOMAS, T. J. Item-presentation controls for multidimensional item pools in computerized adaptive testing. **Behavioral Research and Methods**, 22, 247-252, 1990.

THOMASSON, G. L. CAT Item exposure control: New evaluation tools, alternate methods and integration into a total CAT program. **Annual meeting of the National Council of Measurement in Education,** San Diego, CA, USA, 1998.

THOMASSON, G. L. New item exposure control algorithms for computerized adaptive testing. **Annual meeting of the Psychometric Society**, Minneapolis, MN, USA, 1995.

THOMASSON G. L. The goal of equity within and between computerized adaptive tests and paper and pencil forms. **Annual Meeting of the National Council on Measurement in Education.** Chicago, IL, USA, 1997.

THOMPSON, T. D. **Growth, Precision, and CAT: An Examination of Gain Score Conditional SEM.** Research Report, Pearson, 2008.

THOMPSON, N. A. Item Selection in Computerized Classification Testing. **Educational and Psychological Measurement**, V. 69, N. 5, p. 778-793, 2009a.

THOMPSON, N. A. Termination Criteria for Computerized Classification Testing. **Practical Assessment, Research & Evaluation**, Vol 16, No 4, 2011. Disponível em: <http://pareonline.net/getvn.asp?v=16&n=4>. Acesso em: 10/10/2011.

THOMPSON, N. A. The MEDPRO Project: An SBIR Project for a Comprehensive IRT and CAT Software System — CAT Software. **Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing**, 2009b.

THOMPSON, N. A. Variable-length computerized classification testing with item response theory. **Clear Exam Review**, 17(2), 21-26, 2006.

THOMPSON, N. A.; GUYER, R. **User's Manual for Iteman 4.1.** St. Paul MN: AssessmentSystems Corporation, 2010.

THOMPSON, N. A.; WEISS, D. J. A Framework for the Development of Computerized Adaptive Tests. **Practical Assessment, Research & Evaluation**, 16(1), 2011. Disponível em: <http://pareonline.net/getvn.asp?v=16&n=1>. Acesso em 10/10/2011.

THOMPSON, N. A.; PROMETRIC, T. A Practitioner's Guide for Variable-length Computerized Classification Testing. **Practical Assessment Research & Evaluation**, Vol 12, No 1, 2007.

THOMPSON, T .D.; DAVEY, T. CAT procedures for passage-based tests. **Annual meeting of the NCME**, Montreal, USA, 1999.

THOMPSON, T.; WAY, D. Investigating CAT designs to achieve comparability with a paper test. In: Weiss, D. J. (Ed.). **Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing**, 2007.

TIAN, J-Q.; MIAO, D-M.; ZHU, X.; GONG, J-J. An Introduction to the Computerized Adaptive Testing. **US-China Education Review**, v. 4, n. 1, p. 72-81, 2007.

TOIT, M. **IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT.** Scientific Software International, 2003.

TRIANTAFILLOU, E.; GEORGIADOU, E.; ECONOMIDES, A. A. CAT-MD: Computerized adaptive testing on mobile devices. **International Journal of Web-Based Learning and Teaching Technologies.** Vol. 3, No. 1, pp. 13-20, 2008.

TRUELL, A. D.; ZHAO, J. J.; ALEXANDER, M. W. The Impact of Settable Test Item Exposure Control Interface Format on Postsecondary Business Student Test Performance. **Journal of Career and Technical Education**, Vol. 22, No. 1, p. 31-41, 2005.

TSENG, F.-L.; HSU, T.-C. Multidimensional Adaptive Testing Using the Weighted Likelihood Estimation: A Comparison of Estimation Methods. **Annual meeting of NCME**, Seattle, 2001.

TSUTAKAWA, R. K.; JOHNSON, C. The effect of uncertainty on item parameter estimation on ability estimates. **Psychometrika**, 55, 371–390, 1990.

TUTZ, G. Sequential item response models with an ordered response. **British Journal of Mathematical and Statistical Psychology,** n. 43, p. 39–55, 1990.

UNIVERSITY OF MINNESOTA. Department of Psychology. CAT Central. **Bibliography on Computerized Adaptive Testing (CAT).** Disponível em <http://www.psych.umn.edu/psylabs/catcentral/>. Acesso em: 10 de novembro de 2010.

VALLE, R. C. A Construção e a Interpretação de Escalas de Conhecimento – Considerações Gerais e uma Visão do que vem sendo feito no SARESP. **Estudos em Avaliação Educacional,** n. 23, p. 71-92, 2001.

VAN DER LINDEN, W. J. A comparison of item-selection methods for adaptive tests with content constraints**. Journal of Educational Measurement**, 42, 283–302, 2005a.

VAN DER LINDEN, W. J. A lognormal model for response times on test items. **Journal of Educational and Behavioral Statistics**, 31, 181–204, 2006a.

VAN DER LINDEN, W. J. Assembling Tests for the Measurement of Multiple Traits. **Applied Psychological Measurement**, 20: 373-388, 1996.

VAN DER LINDEN, W.J. Bayesian item selection criteria for adaptive testing. **Psychometrika**, 63, 201-216, 1998a.

VAN DER LINDEN, W. J. Bayesian item selection in adaptive testing. **Annual Meeting of the Psychometric Society**, Minneapolis, MN, USA, 1995.

VAN DER LINDEN, W. J. Computerized Adaptive Testing With Equated Number-Correct Scoring. **Applied Psychological Measurement,** Vol. 25 No. 4, p. 343–355, 2001.

VAN DER LINDEN, W. J. Constrained adaptive testing with shadow tests. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic Publishers, 2000, p. 27-52.

VAN DER LINDEN, W. J. Constrained adaptive testing with shadow tests. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing.** Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010a, p. 31-56.

VAN DER LINDEN, W. J. Empirical initialization of the trait estimator in adaptive testing. **Applied Psychological Measurement**, 23, 21-29, 1999a.

VAN DER LINDEN, W. J. Equating Scores From Adaptive to Linear Tests. **Applied Psychological Measurement,** Vol. 30 No. 6, p. 493–508, 2006b.

VAN DER LINDEN, W. J. **Linear models for optimal test design.** New York: Springer-Verlag, 2005b.

VAN DER LINDEN, W. J. Multidimensional adaptive testing with a minimum error-variance criterion. **Journal of Educational and Behavioral Statistics**, 24, 398-412, 1999b.

VAN DER LINDEN, W. J. Predictive Control of Speededness in Adaptive Testing. **Applied Psychological Measurement**, Vol. 33 No. 1, p. 25–41, 2009.

VAN DER LINDEN, W. J. Sequencing an Adaptive Test Battery. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing.** Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010b, p 103-119.

VAN DER LINDEN, W. J. Some alternatives to Sympson–Hetter item-exposure control in computerized adaptive testing. **Journal of Educational and Behavioral Statistics,** 28, p. 249–265, 2003.

VAN DER LINDEN, W. J. Some new developments in adaptive testing technology. **Zeitschrift für Psychologie,** v. 216, n. 1, p. 22-28, 2008a.

VAN DER LINDEN, W. J. Stochastic order in dichotomous item response models for fixed, adaptive, and multidimensional tests. **Psychometrika**, 63, 211-226, 1998b.

VAN DER LINDEN, W. J. Using Response Times for Item Selection in Adaptive Testing. **Journal of Educational and Behavioral Statistics,** v. 33, N. 1, p. 5–20, 2008b.

VAN DER LINDEN, W. J.; ADEMA, J. J. Simultaneous assembly of multiple test forms. **Journal of Educational Measurement,** 35, 185–198 [Addendum in Vol. 36, 90–91], 1998.

VAN DER LINDEN, W. J.; ARIEL, A.; VELDKAMP, B. P. Assembling a CAT item pool as a set of linear tests. **Journal of Educational and Behavioral Statistics**, 31, 81–100, 2006.

VAN DER LINDEN, W. J.; BOEKKOOI-TIMMINGA, E. A maximin model for test design with practical constraints. **Psychometrika**, 54, 237-247, 1989.

VAN DER LINDEN, W. J.; CHANG, H.-H. Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. **Applied Psychological Measurement**, 27, 107–120, 2003.

VAN DER LINDEN, W. J.; GLAS, C. A. W. Capitalization on Item Calibration Error in Adaptive Testing. **Applied Measurement in Education**, 13 (1). pp. 35-53, 2000a.

VAN DER LINDEN, W. J.; GLAS, C. A. W. **Computerized Adaptive Testing: Theory and Practice.** Dordrecht, Netherlands: Kluwer Academic, 2000b.

VAN DER LINDEN, W. J.; GLAS, C. A. W. Cross-validating item parameter estimation in computerized adaptive testing. In: BOOMSMA, A.; VAN DUIJN, M. A. J.; SNIJDERS, T. A. M. (Eds.), **Essays on item response theory** (pp. 205–219). New York: Springer-Verlag, 2001.

VAN DER LINDEN, W. J.; GLAS, C. A. W. **Elements of Adaptive Testing.** Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010.

VAN DER LINDEN, W. J.; GLAS, C. A. W. Statistical aspects of adaptive testing. In: RAO, C. R.; Sinharay, S. (Eds.), **Handbook of statistics** (Vol. 27: Psychometrics) (pp. 801–838). Amsterdam: North-Holland, 2007.

VAN DER LINDEN, W. J.; GUO, F. Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. **Psychometrika**, V. 73, N. 3, p. 365–384, 2008.

VAN DER LINDEN, W. J.; HAMBLETON, R. K. **Handbook of Modern Item response Theory.** New York: Springer-Verlag, 1997.

VAN DER LINDEN, W. J.; PASHLEY, P. J. Item selection and ability estimation in adaptive testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic, 2000, p 1-25.

VAN DER LINDEN, W. J.; PASHLEY, P. J. Item selection and ability estimation in adaptive testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing.** Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p 3-30.

VAN DER LINDEN, W. J.; REESE, L. M. A model for optimal constrained adaptive testing. **Applied Psychological Measurement**, 22, 259-270, 1998.

VAN DER LINDEN, W. J.; SCRAMS, D. J.; SCHNIPKE, D. L. Using Response-Time Constraints to Control for Differential Speededness in Computerized Adaptive testing. **Applied Psychological Measurement**, Vol. 23 No. 3, p. 195–210, 1999.

VAN DER LINDEN, W. J.; VAN KRIMPEN-STOOP, E. M. L. A. Using Response Times To Detect Aberrant Responses In Computerized Adaptive Testing. **Psychometrika**, V. 68, N. 2, p. 251-265, 2003.

VAN DER LINDEN, W. J.; VELDKAMP, B. P. Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. **Journal of Educational and Behavioral Statistics,** 32, 398-418, 2007.

VAN DER LINDEN, W. J.; VELDKAMP, B. P. Constraining Item Exposure in Computerized Adaptive Testing With Shadow Tests. **Journal of Educational and Behavioral Statistics**, Vol. 29, No. 3, pp. 273–291, 2004.

VAN DER LINDEN, W. J.; VELDKAMP, B. P.; REESE, L. M. An Integer programming approach to item bank design. **Applied Psychological Measurement**, v. 24, n. 2, p 139–150, 2000.

VAN KRIMPEN-STOOP, E. M. L. A.; MEIJER, R. R. CUSUM-Based Person-Fit Statistics for Adaptive Testing. **Journal of Educational and Behavioral Statistics**, Vol. 26, No. 2, p. 199-218, 2001.

VAN KRIMPEN-STOOP, E. M. L. A.; MEIJER, R. R. Detecting person misfit in adaptive testing using statistical process control techniques. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic, 2000, p 201-219.

VAN KRIMPEN-STOOP, E. M. L. A.; MEIJER, R. R. Detection of Person Misfit in Computerized Adaptive Tests With Polytomous Items. **Applied Psychological Measurement**, V. 26 N. 2, p. 164–180, 2002.

VAN KRIMPEN-STOOP, E. M. L. A.; MEIJER, R. R. The null distribution of person-fit statistics for conventional and adaptive tests. **Applied Psychological Measurement,** 23, p. 327-345, 1999.

VAN RIJN, P. W.; EGGEN, T. J. H. M.; HEMKER, B. T.; SANDERS, P. F. Evaluation of selection procedures for computerized adaptive testing with polytomous items. **Applied Psychological Measurement,** 26, p. 393-411, 2002.

VARGAS, V. C. C. **Medida Padronizada para Avaliação de Intangíveis Organizacionais por Meio da Teoria da Resposta ao Item.** 2007. 207 f. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2007.

VARGAS, V. C. C.; SELIG, P. M.; ANDRADE, D. F.; RIBEIRO, J. L. D. Avaliação dos intangíveis: uma aplicação em capital humano. **Gestão da Produção,** São Carlos, v. 15, n. 3, p. 619-634, set.-dez. 2008.

VAS, R. "Educational Ontology and Knowledge Testing" **The Electronic Journal of Knowledge Management,** V. 5, N. 1, p. 123 - 130, 2007.

VASCONCELOS, A. P.; BATISTA, M. J.; ALEXANDRE, J. W. C.; ANDRADE, D. F.; ARAUJO, A. M. S. Teoria da Reposta ao Item Aplicada à Gestão pela Qualidade Total: proposta do modelo de escala gradual. In: 34ª Reunião Regional da Associação Brasileira de Estatística, 2002, Fortaleza - CE. **Resumos**. São Paulo: ABE - Associação Brasileira de Estatística, 2002.

VEERKAMP, W. J. J. Taylor approximations to logistic IRT models and their use in adaptive testing. **Journal of Educational and Behavioural Statistics**, 25, 307-343, 2000.

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

VEERKAMP, W.; GLAS, C. A. W. Detection of known items in adaptive testing with a statistical quality control method. **Journal of Behavioral and Educational Statistics**, v. 25, n. 4, p. 373-389, 2000.

VELDKAMP, B. P. Item selection in polytomous CAT. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), **New developments in psychometrics** (pp. 207–214). Tokyo, Japan: Springer-Verlag, 2003.

VELDKAMP, B. P.; VAN DER LINDEN, W. J. Designing item pools for adaptive testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing**. Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p. 231-246.

VELDKAMP, B. P.; VAN DER LINDEN, W. J. Designing item pools for computerized adaptive testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic, 2000, p 149-162.

VELDKAMP, B. P.; VAN DER LINDEN, W. J. Implementing Sympson-Hetter Item-Exposure Control in a Shadow-Test Approach to Constrained Adaptive Testing. **International Journal of Testing,** 8: 272–289, 2008.

VELDKAMP, B. P.; VAN DER LINDEN, W. J. Multidimensional adaptive testing with constraints on test content. **Psychometrika**, v. 67, n. 4, p. 575-588, 2002.

VELDKAMP, B. P.; VERSCHOOR, A. J.; EGGEN, T. J. H. M. A multiple objective test assembly approach for exposure control problems in Computerized Adaptive Testing. **Psicológica**, 31, p. 335-355, 2010.

VENDRAMINI, C. M. M.; SILVA, M. C.; CANEL, M. Análise de Itens de uma Prova de Raciocínio Estatístico. **Psicologia em Estudo, Maringá,** v. 9, n. 3, p. 487-498, set./dez, 2004.

VERGARA, L. G. L. **Avaliação do Ensino de Ergonomia para o Design Aplicando a Teoria da Resposta ao Item (TRI).** 2005. 186 f. Tese (Doutorado Engenharia de Produção) – Programa de Pós-

Graduação em Engenharia de Produção, Centro Tecnológico, Universidade Federal de Santa Catarina,. Florianópolis, 2005.

VERSCHOOR, A. J.; STRAETMANS, J. J. M. MATHCAT: A Flexible Testing System in Mathematics Education for Adults. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic, 2000, p 101-116.

VERSCHOOR, A. J.; STRAETMANS, J. J. M. MATHCAT: A Flexible Testing System in Mathematics Education for Adults. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing**. Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p. 137-150.

VIANNA, H. M. **Testes em Educação.** São Paulo: IBRASA, 1987.

VISPOEL, W. P. Computerized adaptive and fixed item versions of the ITED vocabulary subtest. **Educational and Psychological Measurement**, 53, 779-788, 1993.

VISPOEL, W. P.Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and text anxiety. **Journal of educational measurement,** 35 (2), 155-167, 1998a.

VISPOEL, W. P. Review and changing answers on computerized adaptive and self-adaptive vocabulary tests. **Journal of educational measurement**, 35 (4), 328-347, 1998b.

VISPOEL, W. P.; HENDERICKSON, A. B.; BLEILER, T. Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. **Journal of Educational Measurement**, 37(1), 21-38, 2000.

VISPOEL, W. P.; ROCKLIN, T. R.; WANG, T. Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. **Applied Measurement in Education**, 53, 53-79, 1994.

VISPOEL, W. P.; ROCKLIN, T. R.; WANG, T.; BLEILER, T. Can examinees use a review option to obtain positively biased estimates on a computerized adaptive test? **Journal of Educational Measurement**, 36 (2), 141-157, 1999.

VISPOEL W. P.; WANG T.; BLEILER T. Computerized adaptive and fixed-item testing of music listening skill: A comparison of efficiency,

precision, and concurrent validity. **Journal of Educational Measurement**. 34, 43–63, 1997.

VITÓRIA, F.; ALMEIDA, L. S.; PRIMI, R. Unidimensionalidade em testes psicológicos: conceito, estratégias e dificuldades na sua avaliação. **PSIC - Revista de Psicologia da Vetor Editora**, v. 7, n. 1, p. 1-7, Jan./Jun. 2006.

VOGELS, A. G. C.; JACOBUSSE, G. W.; REIJNEVELD. Item Response Theory based Computerized Adaptive Testing can provide an accurate and efficient identification of children with psychosocial problems. In: VOGELS, A. G. C. **The identification by Dutch preventive child health care of children with psychosocial problems: do short questionnaires help?** Proefschrift Rijksuniversiteit Groningen, Nederlands, 2008.

VOMLEL, J. Building adaptive tests using bayesian networкs. **Kybernetika,** v. 40, n. 3, p. 333-348, 2004.

VOS, H. J. A Bayesian Procedure in the Context of Sequential Mastery Testing. **Psicológica**, 21, p. 191-211, 2000.

VOS, H. J.; GLAS, C. A. W. Testlet-based adaptive mastery testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic, 2000, p 289-309.

VOS, H. J.; GLAS, C. A. W. Testlet-based adaptive mastery testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing.** Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p 409-431.

WAINER, H. CATs: Whither and whence. **Psicológica,** v. 21, n. 1, p. 121-133, 2000a.

WAINER, H. **Computerized Adaptive Testing: A Primer.** New Jersey: Lawrence Erlbaum Associates, 2000b.

WAINER, H. Rescuing Computerized Testing by Breaking Zipf's Law. **Journal of Educational and Behavioral Statistics,** v. 25, N. 2, p. 203-224, 2000c.

WAINER, H. Some practical considerations when converting a linearly administered test to an adaptive format. **Educational measurement: Issues and practice**, 12, 15-20, 1993.

WAINER, H.; BRADLOW, E.; DU, Z. Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic, 2000, p 245-269.

WAINER, H.; BRADLOW, B.; WANG, X. **Testlet response theory and its application.** Cambridge, UK: Cambridge University Press, 2007.

WAINER, H.; EIGNOR, D. Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In: WAINER, H. (Ed.). **Computerized adaptive testing: A primer**. 2. ed. Hillsdale, New Jersey, USA: LEA, 2000, p. 271-300.

WAINER, H.; KAPLAN, B.; LEWIS, C. A comparison of the performance of simulated hierarchical and linear testlets. **Journal of Educational Measurement**,29, 243-251, 1992.

WAINER, H.; KIELY, G. Item clusters in computerized adaptive testing: A case for testlets, **Journal of Educational Measurement,** n. 24, p. 185-202, 1987.

WAINER, H.; LEWIS, C.; KAPLAN, B.; BRASWELL, J. Building Algebra Testlets: a comparison of hierarchical and linear structures. **Journal of Educational Measurement**, 28, 311-323, 1991.

WAINER, H.; MISLEVY, R. J. Item response theory, calibration, and estimation. In WAINER, H. (Ed.) **Computerized Adaptive Testing: A Primer.** Mahwah, NJ: Lawrence Erlbaum Associates, p. 61-100, 2000.

WAINER, H.; WANG, X. Using a new statistical model for testlets to score TOEFL. **Journal of Educational Measurement**, 37, 203-220, 2000.

WALKER, C. M.; BERETVAS, S. N.; ACKERMAN, T. An Examination of Conditioning Variables Used in Computer Adaptive Testing for DIF Analyses. **Applied Measurement in Education**, V. 14, N. 1, p. 3 – 16, 2001.

WALLER, N. G.; REISE, S. P. Computerized Adaptive Personality Assessment. **Journal of Personality and Social Psychology**, v. 57, n. 6, p. 1051-1058, 1989.

WALTER, O. B. et al. Development and evaluation of a computer adaptive test for ‘Anxiety’ (Anxiety-CAT). **Qual Life Res**, 16, p. 143–155, 2007.

WALTER, O. B. Adaptive Tests for Measuring Anxiety and Depression. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing**. Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p. 123-136.

WALTER, O. B.; HOLLING, H. Transitioning from fixed-length questionnaires to computer-adaptive versions. **Zeitschrift für Psychologie,** v. 216, n. 1, p. 22-28, 2008.

WANG, T. Essentially unbiased estimates in computerized adaptive testing. **Annual meeting of the AERA**, Chicago, USA, 1997.

WANG, C.; CHANG, H-H. Item Selection in Multidimensional Computerized Adaptive Testing - Gaining Information from Different Angles. **Psychometrika**, v. 76, n. 3, p. 363-384, 2011a.

WANG, C.; CHANG, H-H. Restrictive Stochastic Item Selection Methods in Cognitive Diagnostic Computerized Adaptive Testing. **Journal of Educational Measurement,** Vol. 48, No. 3, p. 255–273, 2011b.

WANG, C.; CHANG, H-H.; BOUGHTON, K. A. Kullback–Leibler Information and its Applications in Multi-Dimensional Adaptive Testing. **Psychometrika**, v. 76, n. 1, p. 13-39, 2011.

WANG, W.-C.; LIU, C.-W. Computerized Classification Testing Under the Generalized Graded Unfolding Model. **Educational and Psychological Measurement,** 71(1), p. 114–128, 2011.

WANG, W.-C.; CHEN, P.-H. Implementation and Measurement Efficiency of Multidimensional Computerized Adaptive Testing. **Applied Psychological Measurement,** v. 28, n. 5, p. 295–316, 2004.

WANG, T. H.; HANSON, B. A. **Development and calibration of an item response model that incorporates response time.** American Educational Research Association in Seattle, 2001.

WANG, S.; JIAO, H.; YOUNG, M. J.; BROOKS, T. E.; OLSON, J. Comparability of computer-based and paper-and-pencil testing in K-12 assessment: A meta-analysis of testing mode effects. **Educational and Psychological Measurement**, 68, 5-24, 2008.

WANG, T.; KOLEN, M. J. Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. **Journal of Educational Measurement**, 38 (1), 19-49, 2001.

WANG, H.; SHIN, C. D. Comparability of Computerized Adaptive and Paper-Pencil Tests. **Test, Measurement & Research Services**, v. 13, p. 1-7, 2010.

WANG, S.; WANG, T. Precision of Warm's weighted likelihood estimates for a politomous model in computerized adaptive testing. **Applied Psychological Measurement**, v. 25, n. 4, p. 317-331, 2001.

WANG, T.; HANSON, B. A.; LAU, C.-M. A. Reducing bias in CAT trait estimation: A comparison of approaches. **Applied Psychological Measurement**, 23, 263–278, 1999.

WANG, T.; VISPOEL, W. P. Properties of ability estimation methods in computerized adaptive testing. **Journal of Educational Measurement**, 35, 109–135, 1998.

WANG, Y.-C. et al. Clinical Interpretation of Computerized Adaptive Test–Generated Outcome Measures in Patients With Knee Impairments. **Archives of Physical Medicine and Rehabilitation**, v. 90, n. 8, p. 1340-1348, 2009.

WANG, Y.-C. et al. Translating Shoulder Computerized Adaptive Testing Generated Outcome Measures into Clinical Practice. **Journal of Hand Therapy**, v. 23, n. 4, p. 372-383, 2010.

WARE, J. E. et al. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. **Quality of Life Research** 12: p. 935–952, 2003.

WARE, J. E. et al. Item Response Theory and Computerized Adaptive Testing: Implications for Outcomes Measurement in Rehabilitation. **Rehabilitation Psychology**, v. 50, n. 1, p. 71-78, 2005.

WARM, A. W. Weighted likelihood estimation of ability in item response theory with tests of finite length. **Psychometrika**, 54, 427-450, 1989.

WAY, W. D. Protecting the integrity of computerized testing item pools. **Educational Measurement: Issues and Practice,** n. 17, p. 17-26, 1998.

WAY, W. D.; DAVIS, L. L.; FITZPATRICK, S. **Practical questions in introducing computerized adaptive testing for K12 assessments** (Pearson Educational Measurement Research Report 05-03). Iowa City, IA, USA, 2006.

WAY, W.; ZARA, A.; LEAHY, J. Strategies for managing item pools to maximize item security. **Annual meeting of the National Council on Measurement in Education,** San Diego, USA, 1996.

WEEKS, J. P. **plink: IRT Separate Calibration Linking Method.** R package version 1.3-1, 2011.

WEISS, D. J. Adaptive Testing by Computer. **Journal of Consulting and Clinical Psychology,** v. 53, n. 6, p. 774-789, 1985.

WEISS, D. J. **Ability Measurement: Conventional or Adaptive.** Research Report 73-1. Minneapolis, USA: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973.

WEISS, D. J. Better Data From Better Measurements Using Computerized Adaptive Testing. **Journal of Methods and Measurement in the Social Sciences**, Vol. 2, No. 1, p. 1-23, 2011.

WEISS, D. J. Improving measurement quality and efficiency with adaptive testing. **Applied Psychological Measurement**, 6, 473-492, 1982.

WEISS, D. J. **Manual for POSTSIM 3: Post-hoc simulation of computerized adaptive testing**. Version 3.0. St. Paul MN: Assessment Systems Corporation, 2008.

WEISS, D. J. **Manual for the FastTEST Professional Testing System,** Version 2. St. Paul MN: Assessment Systems Corporation, 2006.

WEISS, D. J. **New horizons in testing: Latent trait test theory and computerized adaptive testing.** New York: Academic Press, 1983.

WEISS, D. J. **Strategies of adaptive ability measurement.** Research Report 74-5. Minneapolis, USA: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

WEISS, D. J. **The stratified adaptive computerized ability test** Research Rep. No. 73-3. Minneapolis, USA: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973.

WEISS, D. J.; GUYER, R. **Manual for CATSim: Comprehensive simulation of computerized adaptive testing**. St. Paul MN: Assessment Systems Corporation, 2010.

WEISS, D. J.; KINGSBURY, G. G. Application of Computerized Adaptive Testing to Educational Problems. **Journal of Educational Measurement,** n. 21, p. 361-375, 1984.

WEISS, D. J.; MCBRIDE, J. R. Bias and information of Bayesian adaptive testing. **Applied Psychological Measurement**, 8, 273-285, 1984.

WEISS, D.; SCHLEISMAN, J. Adaptive testing. In: MASTERS, G.; KEEVES, J. (Eds.) **Advances in measurement in educational research and assessment** p. 129-137. Elsevier Science, the Netherlands, 1999.

WEISSMAN, A. A Feedback Control Strategy for Enhancing Item Selection Efficiency in Computerized Adaptive Testing. **Applied Psychological Measurement**, Vol. 30 No. 2, p.84–99, 2006.

WEISSMAN, A. Assessing the efficiency of item selection in computerized adaptive testing. **Annual Meeting of the American Educational Research Association,** Chicago, IL, USA, 2003.

WEISSMAN, A.. Mutual information item selection in adaptive classification testing. **Educational and Psychological Measurement**, 67, 41-58, 2007.

WEISSMAN, A. Mutual information item selection in multiple-category classification CAT. **Annual meeting of the National Council on Measurement in Education**, San Diego CA, USA, 2004.

WEITZMAN, R. A. Sequential testing for selection. **Applied Psychological Measurement**, 6, 337-351, 1982.

WEN, J.-B.; CHANG, H. H.; HAU, K.-T. Adaptation of the a-stratified method in variable length computerized adaptive testing. **American Educational Research Association Annual Meeting,** 2000.

WIBERG, M. An Optimal Design Approach to Criterion-Referenced Computerized Testing. **Journal of Educational and Behavioral Statistics.** Vol. 28, No. 2, p. 97–110, 2003.

WILSON, M. **Constructing Measures: An Item Response Modeling Approach.** Nova Jersey, USA: Lawrence Erlbaum Associates, 2005.

WILSON, D. T.; WOOD, R.; DOWNS, P. K; GIBBONS, R. **TESTFACT: Test Scoring, Item Statistics and Item Factor Analysis.** Chicago: Scientific Software, Inc, 1991.

WISE, S. L. A critical analysis of the arguments for and against item review in computerized adaptive testing. **Annual meeting of the National Council on Measurement in Education**, New York. (ERIC document reproduction service No. ED 400 267), 1996.

WISE, S. L. Comparison of stratum scored and maximum-likelihood scored CATs. **Annual, conference of the NCME**, Montreal, USA, 1999a.

WISE, S. L. Tests autoadaptados informatizados: Fundamentos, resultados de investigación e implicaciones para la práctica. En J. OLEA, V. PONSODA Y G. PRIETO (eds.), **Tests informatizados: Fundamentos y aplicaciones**. Madrid: Pirámide, 1999b.

WISE, S. L.; KINGSBURY, G. Practical issues in developing and maintaining a computerized adaptive testing program. **Psicológica**, n. 21, p. 135-155, 2000.

WISE, S. L.; PONSODA, V.; OLEA, J. Self-adapted testing: an overview. **International Journal of Continuing Engineering Education and Lifelong Learning**, 12 (1-4), 107-122, 2002.

WISE, S. L.; ROOS, L. R.; PLAKE, B. S.; NEBELSICK-GULLETT, L. J. The relationship between examinee anxiety and preference for self-adapted testing. **Applied measurement in education**, 7(1), 81-91, 1994.

WISNIEWSKI, D. R. An application of the Rasch model to computerized adaptive testing: The binary search method (Doctoral dissertation, Wayne State University, 1985). **Dissertation Abstracts International**, 47(1-A), 159, 1986.

WOLFE, J. H.; MCBRIDE, J. R.; SYMPSON, J. B. Development of the experimental CAT-ASVAB system. In: SANDS, W. A; WATERS, B. K.; MCBRIDE, J. R. (Eds). **Computerized Adaptive Testing: from inquiry to operation.** Washington, USA: American Psychological Association, 1997.

WOLFE, J. H.; MORENO, K. E.; SEGALL, D. O. Evaluating the Predictive Validity of CAT-ASVAB. In: SANDS, W. A; WATERS, B. K.; MCBRIDE, J. R. (Eds). **Computerized adaptive testing: From**

inquiry to operation (pp. 175-180). Washington, DC: American Psychological Association, 1997.

WRIGHT, B. D. Practical adaptive testing. **Rasch Measurement Transactions,** v. 2:2, p.24, 1988.

WRIGHT, B. D. **Sample-free test calibration and person measurement.** Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: ETS - Educational Testing Service, 1968.

WRIGHT, B. D.; BELL, S. R. Item Banks: What, Why, How. **Journal of Educational Measurement**, 21, p. 331–345, 1984.

XIAO, B. Strategies for computerized adaptive grading testing. **Applied Psychological Measurement**, 23, 136-146, 1999.

XING, D.; HAMBLETON, R. K. Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. **Educational and Psychological Measurement**, V. 64, N. 1, p. 5-21, 2004.

XU, X.; DOUGLAS, J. Computerized Adaptive Testing Under Nonparametric IRT Models. **Psychometrika**. v. 71, n. 1, p. 121–137, 2006.

YAMAMOTO, K. **Estimating the Effects of Test Length and Test Time on Parameter Estimation Using the HYBRID Model.** TOEFL – Technical Report TR-10. New Jersey: ETS, 1995.

YAN, D.; LEWIS, C.; STOCKING, M. Adaptive Testing With Regression Trees in the Presence of Multidimensionality. **Journal of Educational and Behavioral Statistics**. v. 29, n. 3, p. 293-316, 2004.

YANG, X.; POGGIO, J. C.; GLASNAPP, D. R. Effects of Estimation Bias on Multiple-Category Classification with an IRT-Based Adaptive Classification Procedure. **Educational and Psychological Measurement**, 66, p. 545-564, 2006.

YAO, T. CAT with a Poorly Calibrated Item Bank. **Rasch Measurement Transactions,** v. 5:2, p. 141, 1991.

YEH, S. S. Reforming Federal Testing Policy to Support Teaching and Learning. **Educational Policy**, V. 20, N. 3, p. 495-524, 2006.

YI, Q. Incorporating the Sympson-Hetter exposure control method into the a-stratified method with content blocking. **Annual meeting of AERA,** New Orleans, USA, 2002.

YI, Q.; NERING, M. L. **Simulating nonmodel-fitting responses in a CAT environment** (Research report). Iowa City, IA, USA: ACT, 1998.

YI, Q.; WANG, T.; BAN, J.C. **Effects of scale transformation and test termination rule on the precision of ability estimates in CAT**. (Research report). Iowa City, IA, USA: ACT, 2000.

YI, Q.; ZHANG, J.; CHANG, H-H. Assessing CAT Test Security Severity. **Applied Psychological Measurement**, V. 30 N. 1, p. 62-63, 2006.

YI, Q.; ZHANG, J.; CHANG, H-H. Severity of Organized Item Theft in Computerized Adaptive Testing: A Simulation Study. **Applied Psychological Measurement**, V. 32 N. 7, p. 543–558, 2008.

YOUNT, S. E. et al. Brief, Valid Measures of Dyspnea and Related Functional Limitations in Chronic Obstructive Pulmonary Disease (COPD). **Value in Health**, v. 14, n. 2, p. 307-315, 2011.

YU, C. H.; POPP, S. O.; DIGANGI, S.; JANNASCH-PENNELL, A. Assessing unidimensionality: A comparison of Rasch Modeling, Parallel Analysis, and TETRAD. **Practical Assessment, Research & Evaluation.** v. 12, n. 14, Out 2007.

ZARA, A. R. Using Computerized Adaptive Testing to Evaluate Nurse Competence for Licensure: Some History and Forward Look. **Advances in Health Sciences Education** 4: 39–48, 1999.

ZENISKY, A.; HAMBLETON, R. K.; LUECHT, R. M. Multistage Testing: Issues, Designs, and Research. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing.** Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p 355-372.

ZHANG, Q. Towards cognitive response theory in diagnostic language assessment. In: CHAPELLE, C. A.; CHUNG, Y.-R.; XU, J. (Eds.). **Towards adaptive CALL: Natural language processing for diagnostic language assessment** (p. 40-61). Ames, IA: Iowa State University, 2008.

ZHAO, Y.; HAMBLETON, R. **Software for IRT Analyses: Descriptions and Features.** 2009. Disponível em

<http://www.umass.edu/remp/software/CEA-652.ZH-IRTSoftware.pdf>. Acesso em: 31 de agosto de 2011.

ZHU, D.; FAN, M. Adjusting computer adaptive test starting points to conserve item pool. **Annual meeting of the AERA**, Montreal, USA, 1999.

ZICKAR, M. J.; OVERTON, R. C.; TAYLOR, R; HARMS, H. J. The development of a computerized selection system for computer programmers in a financial services company. In: DRASGOW, F.; OLSON-BUCHANAN, J. B. (Eds.). **Innovations in computerized assessment.** Mahwah, NJ, USA: LEA, 1999.

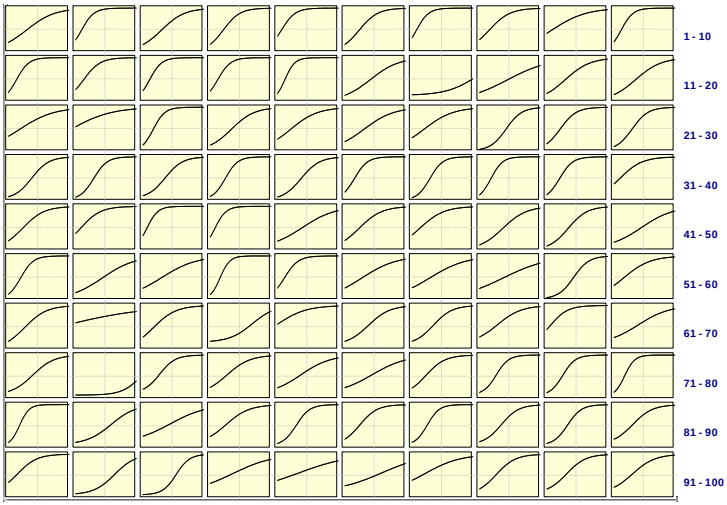
ZOPLUOGLU, C. **EstCRM: Calibrating Parameters for the Samejima's Continuous IRT Model.** R package version 1.1, 2011.

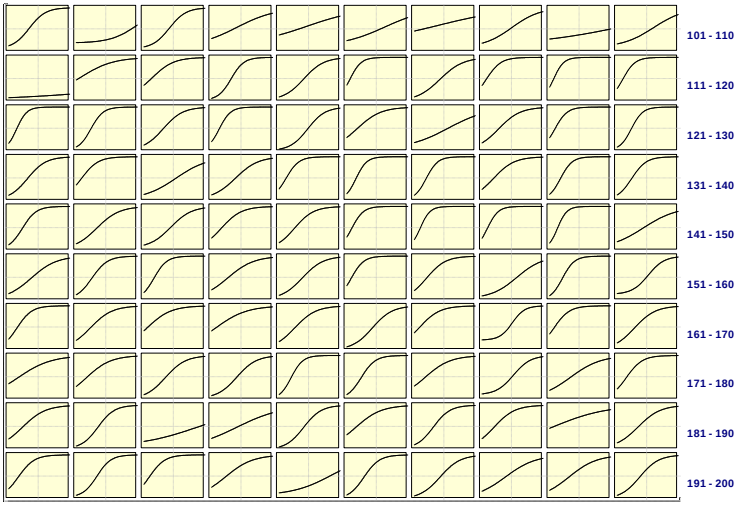
ZWICK, R. The Assessment of differential item functioning in computer adaptive tests. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Computerized adaptive testing: Theory and practice.** Dordrecht, Netherlands: Kluwer Academic, 2000, p 221-244.

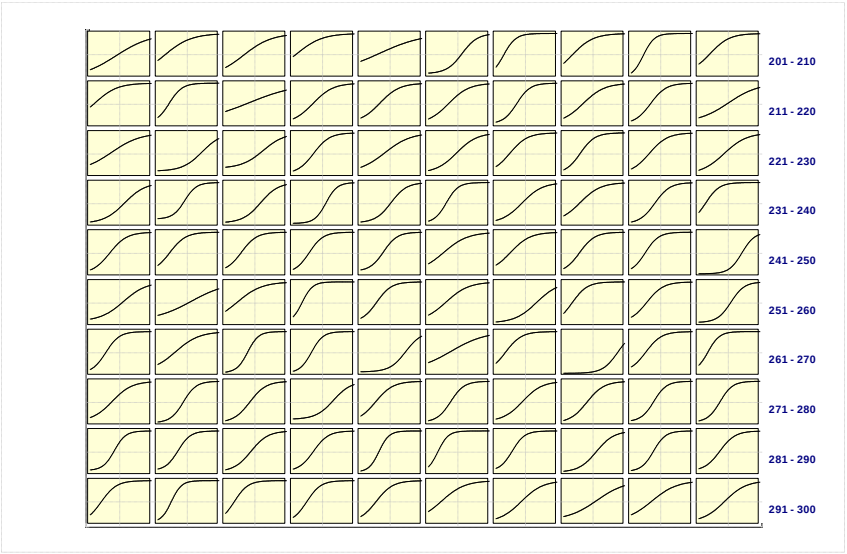
ZWICK, R. The Investigation of Differential Item Functioning in Adaptive Tests. In: VAN DER LINDEN, W. J; GLAS, C. A. W. (Eds.). **Elements of Adaptive Testing**. Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010, p. 331-352.

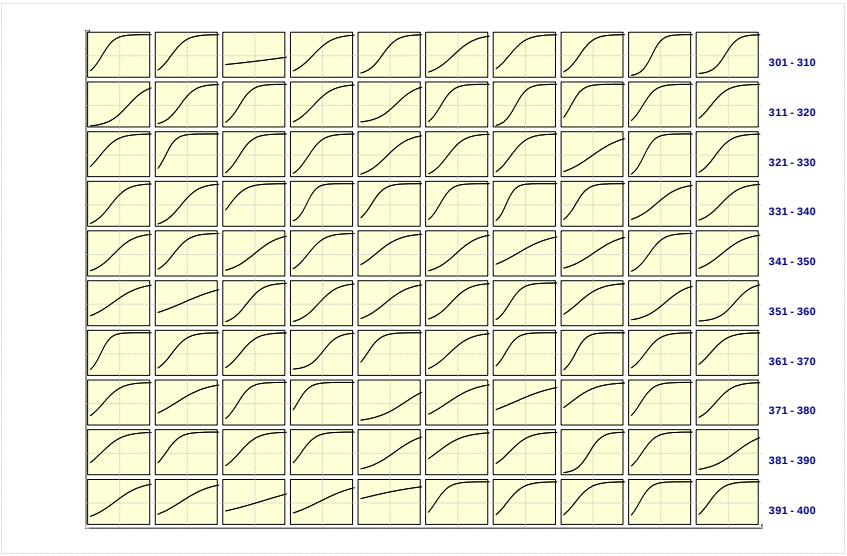
ZWICK, R.; THAYER, D. T. Application of an Empirical Bayes Enhancement of Mantel-Haenszel Differential Item Functioning Analysis to a Computerized Adaptive Test. **Applied Psychological Measurement**, Vol. 26 No. 1, p. 57–76, 2002.

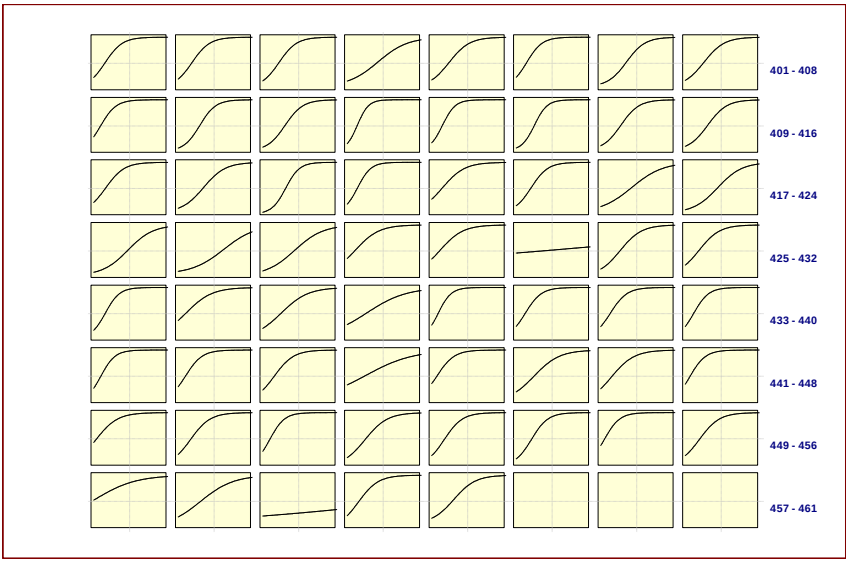
APÊNDICE A – CCI DOS ITENS APÓS A PRIMEIRA CALIBRAÇÃO



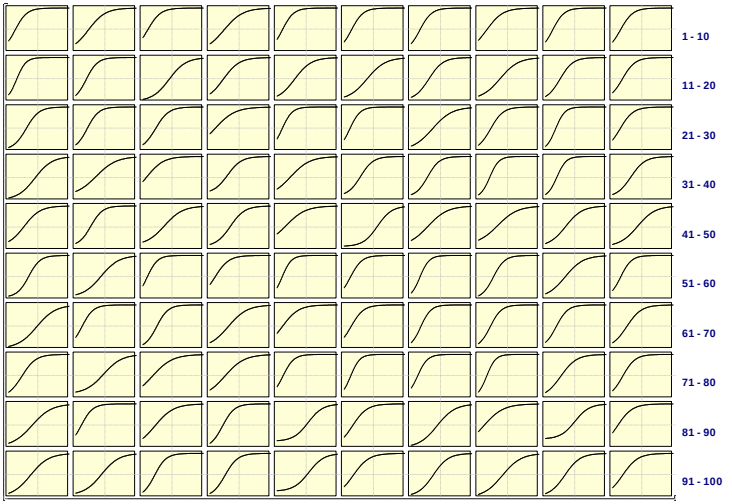


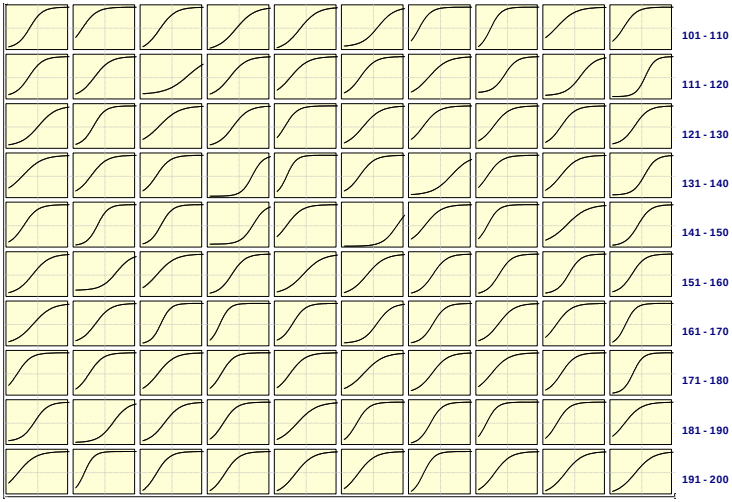

201 - 210
211 - 220
221 - 230
231 - 240
241 - 250
251 - 260
261 - 270
271 - 280
281 - 290
291 - 300

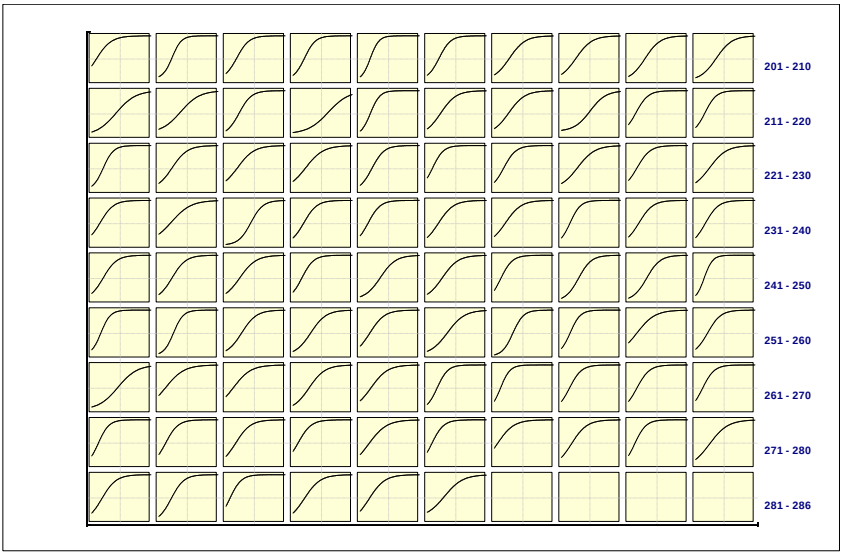

301 - 310
311 - 320
321 - 330
331 - 340
341 - 350
351 - 360
361 - 370
371 - 380
381 - 390
391 - 400


401 - 408
409 - 416
417 - 424
425 - 432
433 - 440
441 - 448
449 - 456
457 - 461

APÊNDICE B – CCI DOS ITENS APÓS A QUINTA CALIBRAÇÃO (A > 1)






201 - 210
211 - 220
221 - 230
231 - 240
241 - 250
251 - 260
261 - 270
271 - 280
281 - 286