UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Testes adaptativos sensíveis ao conteúdo do banco de itens: uma aplicação em exames de proficiência em inglês para programas de pós-graduação

Leandro Henrique Mendonça de Oliveira

São Carlos - SP

Testes adaptativos sensíveis ao conteúdo do banco de itens: uma aplicação em exames de proficiência em inglês para programas de pós-graduação

Leandro Henrique Mendonça de Oliveira

Orientador:

Prof. Dr. ***Sandra Maria Aluisio***

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências de Computação e Matemática Computacional.

USP – São Carlos
Maio/2002

"VERSÃO REVISADA APÓS A DEFESA"

Data da Defesa:	05/04/2002
Visto do Orientador:	

"Feliz é aquele que transfere o que sabe e aprende o que ensina."
(Cora Coralina)

Dedicatória

Aos meus pais (Eloir e Elizete) pelo exemplo de vida,
e aos irmãos (Leonardo e Delano) pelo afeto e companheirismo.

Agradecimentos

Seria ótimo se eu pudesse escrever em apenas uma única folha de papel o nome de todas as pessoas a quem devo gratidão. Mesmo se o papel fosse enorme seria impossível prestar meus agradecimentos àqueles que, desde o início, confiaram no meu esforço e dedicação, e que de variadas formas contribuíram para a realização deste trabalho. Mesmo porque quero evitar cometer injustiças a quem sempre esteve (mesmo à distância) ao meu lado. A essas pessoas resta-me dizer em letras maiúsculas o meu sincero MUITO OBRIGADO por serem exatamente o que preciso no momento em que mais preciso. Entretanto, especialmente agradeço:

- a Deus, pelo dom da vida;
- à minha família (pais e irmãos) por terem acreditado no meu desejo de uma possível realização profissional;
- a Déia, minha doce namorada, pelo amor a mim dedicado;
- aos meus avós (Armando e Narzira, *in memorian*) que mesmo sem entenderem o que significa uma pós-graduação, não relutaram;
- aos tios e tias (Aredizon, *in memorian*) pelo cuidado, presença e preocupação constantes;
- aos primos e primas pela atenção dispensada em todos os momentos. Em especial ao Juninho (Epaminondas Júnior) por ter sido muitas vezes meu amparo;
- a Marnilda, Avelino, Larissa, Avelino Júnior, Vanny, Vanessa, D. Joaninha e Sr. Cristiano por terem dedicado a mim tantos olhares;
- a Telma, pelas palavras de apoio;
- ao Prof. Paulo André pela amizade e incentivo para que eu fizesse o mestrado;
- à Profa. Sandra Maria Aluísio por ter confiado em meu trabalho e pelo privilégio da sua orientação;
- à Profa. Maria Carolina Monard por ter me aceito como aluno especial;
- aos professores Bento Carlos e Dilvan Moreira pelas contribuições no Exame de Qualificação;
- ao Prof. Dorival Leão pelos esclarecimentos em estatística;
- aos amigos do Grupo de Tecnologia em Informática (GTI) pelo estímulo constante. Em especial ao Alan, Giovanna e Rodrigo por terem sido minha família São Carlense;
- aos amigos do NILC pela amizade, e, principalmente, por terem compartilhado momentos difíceis e felizes em todo esse tempo;
- aos amigos do ICMC pelos momentos alegres e de boas conversas no "redondo";
- as amigas Rossana, Kleyssy, Lorayne, Katyana, Laríssia e Letycia pela torcida sincera;
- ao amigo Julian Stella pela amizade verdadeira e por ter tocado e criado melodias tão goianas que amenizaram minha saudade;
- à Herondina Alves Pinto pelas valiosas correções do texto;
- ao CNPq pelo apoio financeiro no segundo ano de trabalho.

A todos, meu ETERNO OBRIGADO.

Resumo

A aplicação de Testes Adaptativos Computadorizados a vários tipos de avaliações e domínios tem se tornado cada vez mais fácil devido ao suporte teórico da Teoria de Resposta de Itens (TRI), a maior disponibilidade de computadores pessoais mais potentes e ao desenvolvimento de *software* rodando nestes computadores nestes computadores para realizar as análises da TRI. Os testes adaptativos ajustam as suas questões ao nível de habilidade individual de cada estudante, adaptando-se à capacidade de cada um, gerando uma avaliação mais eficiente e de maior qualidade. A existência de um banco de itens calibrado, de acordo com os modelos estatísticos da TRI, permite a administração adequada dos testes adaptativos. As principais dificuldades de implantação dos testes adaptativos são a necessidade de um estudo empírico em larga escala, envolvendo profissionais de várias áreas, para uma boa calibração do banco de itens e a geração de avaliações multidimensionais, isto é, avaliações que atendam a diferentes conteúdos de uma área do conhecimento. O presente trabalho se propôs a popular e calibrar um banco de itens do domínio de inglês intrumental explorando técnicas sensíveis ao conteúdo do banco de itens para permitir a avaliação de várias habilidades. Atualmente, o banco possui 103 itens e estimamos que em três semestres ele virá a conter o número ideal de itens para seu uso em um teste adaptativo, um tempo razoável, dado o número de alunos do programa de pós-graduação. Adicionalmente, foram desenvolvidos um sistema de gerência desse banco e a modelagem de um sistema de avaliação adaptativa que podem ser utilizados tanto para a avaliação diagnóstica do inglês instrumental, quanto para a avaliação formal da proficiência em inglês para admissão em programas de mestrado. A modelagem foi bem detalhada para permitir uma implementação fácil do sistema, dada a dificuldade inerente à teoria estatística utilizada em testes adaptativos.

Abstract

The application of Computerized Adaptive Tests to many types of measurements and domains is becoming easier due to the framework provided by the Item Response Theory (IRT), the increase in power and availability of personal computers and software to perform IRT analyses. The items administered to each individual are selected in order to be the most appropriate for evaluating that individual, allowing a more efficient, high quality assessment. A pool of test items calibrated according to TRI models allows an adequate use of adaptive testing. The main difficulties to employ adaptive testing are the need of a large-scale empirical study for item calibration and the generation of multidimensional tests, i.e., tests that address different knowledge contents. One of the objectives of this work was to populate and calibrate a test item pool for the domain of English for academic purposes. Methods that are sensitive to the content of the pool of items were exploited to allow assessment of several abilities. Currently, the item pool has 103 itens and three semesters is an estimate of the time taken for it to be ready for use in a real application. We consider this acceptable considering the number of master students of our post-graduate program. Moreover, an item management system and the modeling of an adaptive system were developed to be used as diagnostic or final assessment of English for academic purposes in proficiency exams in post-graduate programs. We have done a detailed modeling allowing an easy implementation of the adaptive testing system as the statistical theory used in the system components is very hard.

Sumário

Lista de Figuras v

Lista de Tabelas ix

1 Introdução 1
- 1.1 Contextualização . . . 1
  - 1.1.1 Visão Geral das Avaliações . . . 1
  - 1.1.2 As Avaliações Informatizadas . . . 2
  - 1.1.3 Avaliação Informatizada do Exame de Proficiência em Inglês (EPI) do ICMC 5
- 1.2 Motivação e Relevância . . . 7
- 1.3 Objetivos . . . 8
- 1.4 Organização do Trabalho . . . 9

2 Testes Objetivos Informatizados 11
- 2.1 Histórico . . . 11
- 2.2 Tipos e Estruturas dos Testes Objetivos . . . 12
- 2.3 Abordagens . . . 15
  - 2.3.1 Métodos Convencionais Informatizados . . . 15
  - 2.3.2 Métodos Alternativos Informatizados . . . 16
    - 2.3.2.1 Medida de Probabilidade Admissível (MPA) . . . 16
    - 2.3.2.2 Testes Adaptativos . . . 19
- 2.4 Diretrizes e Princípios dos Testes Informatizados . . . 20
- 2.5 Vantagens dos Testes Informatizados . . . 21
  - 2.5.1 Vantagens Relacionadas com as Condições de Aplicação . . . 21
  - 2.5.2 Vantagens Relacionadas com o Processamento de Respostas e suas Interpretações . . . 22
- 2.6 Equivalência Entre os Testes Informatizados e de Lápis e Papel . . . 22
- 2.7 Banco de Itens - BI . . . 23
  - 2.7.1 Balanceamento de Conteúdo e Testlets . . . 24
  - 2.7.2 Construção do Banco de Itens . . . 25
  - 2.7.3 Construção Automática de Testes . . . 27
  - 2.7.4 Informações Incluídas . . . 27
  - 2.7.5 Procedimento de Construção de Testes . . . 29
  - 2.7.6 Manutenção do Banco de Itens . . . 29
    - 2.7.6.1 Atualização do Banco de Itens . . . 29
    - 2.7.6.2 Renovação do Banco de Itens . . . 30

3 Testes Adaptativos Informatizados (TAI's): Evolução e Propriedades 33
3.1 Evolução dos Testes Adaptativos Informatizados . . . 34
3.1.1 O Primeiro Teste Adaptativo - O "Binet Test" . . . 34
3.1.2 Outras Abordagens de Testes Adaptativos . . . 35
3.2 Propriedades dos Testes Adaptativos . . . 38
3.2.1 Dimensionalidade . . . 39
3.2.2 Confiabilidade . . . 39
3.2.3 Validade . . . 40
3.2.4 Estimação de Parâmetros e Calibração dos Itens . . . 40
3.2.5 Fatores Humanos . . . 41
3.3 A Lógica dos Testes Adaptativos Informatizados . . . 41
3.3.1 Potencialidades e Limitações dos TAI's . . . 42
3.3.2 Questões Técnicas . . . 43
3.3.2.1 Estudos de Simulação . . . 43
3.3.2.2 Exposição dos Itens . . . 43
3.3.2.3 Tamanho do Banco de Itens . . . 45
3.3.2.4 Critérios de Parada . . . 45
3.3.2.5 Revisão de Respostas . . . 45
3.3.2.6 Conflito e Repetição de Itens . . . 46

4 Formalização dos Testes Adaptativos: A Teoria de Resposta de Itens (TRI) 47
4.1 Conceito e Características . . . 47
4.2 Modelos de Resposta de Itens . . . 49
4.2.1 Modelo Normal de Dois Parâmetros . . . 50
4.2.2 Modelo Logístico de Dois Parâmetros . . . 51
4.2.3 Modelo Logístico de Três Parâmetros . . . 52
4.2.4 Modelo Logístico de Um Parâmetro (*Rasch Model*) . . . 54
4.3 Função de Informação de um Item . . . 54
4.4 Abordagens de Seleção de Itens Baseados na TRI . . . 57
4.4.1 Método da Máxima Informação . . . 57
4.4.2 Método Bayesiano . . . 59
4.4.3 Técnicas de Busca de Itens . . . 61
4.5 Estimação de Parâmetros . . . 62
4.5.1 Método da Estimativa de Máxima Verossimilhança (EMV) . . . 62
4.5.1.1 Estimatição dos Parâmetros dos Itens . . . 64
4.5.1.2 Estimação de Habilidade θ . . . 65
4.5.1.3 Método Bayesiano de Estimativa de Habilidade . . . 68
4.5.2 O Programa XCALIBRE . . . 69
4.5.2.1 Formatação dos Dados de Entrada . . . 69
4.5.2.2 Descrição das Interfaces do Programa XCALIBRE . . . 72
4.5.2.3 Interpretação dos Resultados da Calibração . . . 75
4.6 Procedimentos de Pontuação . . . 77

5 Programas para a construção e administração dos testes informatizados 81
5.1 Características e Tipos de Programas . . . 82
5.2 Programas Analisados . . . 83
5.2.1 O AdTest - Administração e Análise . . . 83

5.2.2 O FastTest Professional - Construção, Administração e Análise 86
5.2.3 Comparação entre os Programas AdTest e FastTest Professional sob uma Abordagem Adaptativa 89

6 O Desenvolvimento de um Teste Adaptativo Sensível ao Conteúdo 95
6.1 Criação e Calibração do Banco de Itens 96
6.1.1 Organização do Conteúdo do Banco de Itens 96
6.1.2 Coleta dos Dados de Análise 98
6.1.3 Organização das Provas do EPI 98
6.1.4 Análise e Seleção das Questões Válidas 100
6.1.5 Elaboração dos Conjuntos de Dados para Estimação 101
6.1.6 O Processo de Calibração 103
6.1.7 Resultados da Estimação dos Parâmetros 105
6.1.8 Inclusão de Novas Questões no Banco de Itens 107
6.2 Sistema de Gerenciamento do Banco de Itens (SisBI) 108
6.2.1 Os Diagramas de Modelagem da UML 108
6.2.2 Descrição das Interfaces e Funcionalidades do SisBI 109
6.3 Discussão dos Resultados 112
6.4 Modelagem do Sistema de Avaliação Diagnóstica TAEPI 113
6.4.1 O Ambiente do Administrador 115
6.4.2 O Ambiente do Professor 115
6.4.3 O Ambiente do Aluno 117
6.4.4 A Importância do TAEPI 119
6.5 Trabalhos Relacionados 119

7 Conclusão e Trabalhos Futuros 127

Bibliografia e Referências 131

A Exemplo de uma Prova do Exame de Proficiência em Inglês – EPI 135

B Modelagem e Interfaces do Sistema de Gerenciamento do Banco de Itens – SisBI 149
B.1 Interfaces de Uso 149
B.2 Diagramas de Casos de Uso 158
B.3 Descrição das Classes 159
B.4 Diagrama de Classes 160
B.5 Diagramas de Atividades 160

C Documentação e Modelagem do Sistema TAEPI 171
C.1 Documentação das Classes 171
C.1.1 Documentação das Classes e Atributos 171
C.1.2 Documentação das Operações 175
C.2 Documentação do Diagrama de Classes 176
C.3 Documentação dos Diagramas de Casos de Uso 176
C.3.1 Tarefa: CADASTRAR PROFESSOR 176
C.3.2 Tarefa: EFETUAR LOGIN 176
C.3.3 Tarefa: CRIAR EXAME 177
C.3.4 Tarefa: ESPECIFICAR CRITÉRIO DE APLICAÇÃO DO EXAME 177

C.3.5 Tarefa: CONSULTAR RESULTADO GLOBAL 177
C.3.6 Tarefa: CONSULTAR RESULTADO DETALHADO 177
C.3.7 Tarefa: CADASTRAR ALUNO 177
C.3.8 Tarefa: REALIZAR EXAME 178
C.3.9 Tarefa: CONSULTAR RESULTADO 178
C.3.10 Tarefa: CONSULTAR GABARITO 178
C.4 Descrição dos Diagramas da UML 178
C.4.1 Diagramas de Casos de Uso 178
C.4.1.1 Diagrama de Caso de Uso do Administrador 179
C.4.1.2 Diagrama de Caso de Uso do Professor 180
C.4.1.3 Diagrama de Caso de Uso do Aluno 181
C.4.2 Descrição das Classes 182
C.4.3 Diagrama de Classes 183
C.4.4 Diagramas de Atividades 184

Lista de Figuras

2.1 Partes que compõem uma questão de múltipla escolha 13
2.2 Modelo Espacial de distribuição das opções de resposta do método MPA 17
2.3 Diagrama que representa o Balanceamento de Conteúdo e Testlets em um banco de itens 26
2.4 Diagrama de passos para a construção de um banco de itens (Olea et al. 1999) . . . 28
2.5 Diagrama de passos para atualização de um banco de itens (Olea et al. 1999) 30
2.6 Diagrama dos passos para a renovação de um banco de itens (Olea et al. 1999) . . . 31

3.1 Registro de resposta de um aluno em um Teste de Binet (+ = correto, - = incorreto(Weiss, 1985) 35
3.2 Registro de resposta de um aluno em um Teste Adaptativo Estratificado (+ = correto, - = incorreto)(Weiss, 1985) 37

4.1 Passos para a estimação dos parâmetros dos itens (Hambleton and Swaminathan, 1985) 48
4.2 Curva Característica de um Item, com valores de a positivo e negativo (Baker, 1992) 51
4.3 Curvas Características de um Item baseado nos modelos Normal e Logístico com b = 0.6 e a = 1.2 (Baker, 1992) 52
4.4 Curva Característica de um Item baseado no modelo logístico de três parâmetros, com b=0.6, a=1.2 e c=0.15 (Baker, 1992) 53
4.5 Curva Característica de um Item baseado no Rasch Model, com b = 1.822 (Baker, 1992) 55
4.6 Curvas de informação de 10 itens hipotéticos (Weiss, 1985) 56
4.7 Representação das curvas de informação de 10 itens e o nível de habilidade com valor 0 (Weiss, 1985) 59
4.8 Representação das curvas de informação de 9 itens e o nível de habilidade com valor -1 (Weiss, 1985) 60
4.9 Representação das curvas de informação de 8 itens e o nível de habilidade com valor -0,5 (Weiss, 1985) 61
4.10 Um exemplo do formato do arquivo de entrada dos dados do programa XCALIBRE . 70
4.11 Primeira tela do programa XCALIBRE 72
4.12 Janela da pasta *Files* 73
4.13 Janela da pasta *Options* 74
4.14 Exemplo de valores da estimação final dos parâmetros 76
4.15 Gráfico da função de informação de um teste 78

5.1 Tela de pedido de identificação com as três iniciais 84
5.2 Tela que mostra uma questão a um estudante, com a opção 4 e o tempo restante . . 84

5.3 Tela que apresenta o resultado ao final de um teste 85
5.4 Tela que apresenta o resultado detalhado do teste, contendo informações de todos itens administrados. 85
5.5 Distribuição normal da habilidade alcançada.. 86
5.6 Tela inicial do programa FastTest Professional. 88
5.7 Tela de apresentação e operação de banco de itens 89
5.8 Tela que representa as tarefas de edição e criação de um item 90
5.9 Tela que representa a tarefa de criação e edição de um teste 91
5.10 Tela de configuração de um determinado teste 92

6.1 Organização do conteúdo (módulos e partes) do banco de itens para o Exame Diagnóstico Adaptativo 97
6.2 Exemplo de valores dos parâmetros estimados para os modelos 2P e 3P 106
6.3 Exclusão de um item durante a fase de estimação 107
6.4 Diagrama de Caso de Uso do professor 110
6.5 Diagrama de Caso de Uso geral do TAEPI 114
6.6 Diagrama de Caso de Uso do Ambiente do Administrador 115
6.7 Diagrama de Atividades que representa a função Cadastrar Professor 116
6.8 Diagrama de Caso de Uso do Ambiente do Professor 117
6.9 Digrama de Atividades da função Criar Exame 124
6.10 Diagrama de Caso de Uso do Ambiente do Aluno 125
6.11 Diagrama de Atividades da função Realizar Exame 126

B.1 Tela de entrada (acesso) do SisBI 150
B.2 Tela que representa o formulário de inclusão de uma nova questão. 151
B.3 Janela que representa a exclusão de questões. 152
B.4 Janela que representa a alteração de questões. 153
B.5 Tela de consulta de questões. 154
B.6 Janela de gráficos da TRI de uma determinada questão com valores de a=1,03 e b=0,87. 155
B.7 Janela de inclusão de Nova Parte. 156
B.8 Janela de inclusão de Novo Texto. 157
B.9 Diagrama de Caso de Uso Geral. 158
B.10 Diagrama de Caso de Uso do Professor. 159
B.11 Descrição das Classes, Atributos e Operações. 159
B.12 Diagrama de Classes. 160
B.13 Diagrama de Atividades – Efetuar Login. 161
B.14 Diagrama de Atividades – Incluir Questão. 162
B.15 Diagrama de Atividades – Excluir Questão. 163
B.16 Diagrama de Atividades – Alterar Questão. 164
B.17 Diagrama de Atividades – Consultar Questões. 165
B.18 Diagrama de Atividades – Traçar gráficos da TRI referente a uma questão. 166
B.19 Diagrama de Atividades – Incluir Nova Parte. 167
B.20 Diagrama de Atividades – Incluir Novo Texto. 168
B.21 Diagrama de Atividades - Consultar Textos. 169

C.1 Diagrama de Caso de Uso Geral do TAEPI. 179
C.2 Diagrama de Caso de Uso do Administrador. 179
C.3 Diagrama de Caso de Uso do Professor. 180

C.4 Diagrama de Caso de Uso do Aluno. 181
C.5 Descrição da Classes, Atributos e Operações. 182
C.6 Diagrama de Classes. 183
C.7 Diagrama de Atividades – Cadastrar Professor. 184
C.8 Diagrama de Atividades – Efetuar Login. 185
C.9 Diagrama de Atividades – Criar Exame. 186
C.10 Diagrama de Atividades – Consultar Resultado Global. 187
C.11 Diagrama de Atividades – Especificar Critérios de Aplicação do Exame. 188
C.12 Diagrama de Atividades – Ler Matriz de Critérios de Parada do Exame. 189
C.13 Diagrama de Atividades – Consultar Resultado Detalhado. 190
C.14 Diagrama de Atividades – Cadastrar Aluno. 191
C.15 Diagrama de Atividades – Consultar Resultado. 192
C.16 Diagrama de Atividades – Realizar Exame. 193
C.17 Diagrama de Atividades – Especificar Tetha Inicial. 194
C.18 Diagrama de Atividades – Aplicar Parte com Questões Individuais. 195
C.19 Diagrama de Atividades – Selecionar Questão mais Adequada. 196
C.20 Diagrama de Atividades – Administrar Questão. 197
C.21 Diagrama de Atividades – Aplicar Parte com TestLets. 198
C.22 Diagrama de Atividades – Selecionar TestLet mais Adequado. 199
C.23 Diagrama de Atividades – Consultar Gabarito. 200

Lista de Tabelas

1.1 Divisão dos módulos do EPI 7

2.1 Recompensa relativa às opções de escolha e tipo de classificação (Aquino, 2001) . . . 18

4.1 Principais considerações sobre modelos de resposta de itens (Hambleton and Swaminathan, 1985) 47
4.2 Modelos de resposta de um item, para respostas objetivas e subjetivas (Hambleton and Swaminathan, 1985) 49
4.3 Descrição dos caracteres da linha de controle do arquivo de entrada de dados 71

5.1 Comparação das principais características dos programas FasTest Pro e AdTest . . . 90
5.2 Dados da execução dos programas FastTest Pro e AdTest usando os métodos MLE e Bayesiano sobre um banco de itens de História Geral 93

6.1 Identificação das provas e área pertencente 98
6.2 Carta de solicitação de colaboração enviada aos alunos do ICMC 99
6.3 Identificação de Módulos e Partes 100
6.4 Distribuição dos módulos e partes em cada prova 100
6.5 Relatório das questões semelhantes entre as provas e número total de questões válidas 101
6.6 Relação dos conjuntos de análise distribuídos por módulo 103
6.7 Resultados da estimação dos conjuntos de análise 104
6.8 Resultado final da estimação dos parâmetros com os respectivos ajuste dos itens ao modelo da TRI 122
6.9 Perspectiva de tempo para o preenchimento ideal do banco de itens do EPI 123
6.10 Critérios de parada, métodos de estimação e seleção de itens nos sistemas adaptativos avaliados 123

CAPÍTULO 1

Introdução

1.1 Contextualização

1.1.1 Visão Geral das Avaliações

Praticamente em todos os processos de ensino, de treinamento, ou nos casos de admissão em programas educacionais, a avaliação é utilizada para averiguar se as metas pretendidas por estes foram alcançadas, ou se os ingressantes nesses programas possuem o nível de conhecimento e habilidade desejados. Sob esta ótica, em linhas gerais, a avaliação está inserida no contexto da educação assumindo importantes papéis que servem para corroborar o aprendizado, sendo muitas vezes um meio capaz de determinar o potencial de um indivíduo.

Considerando estes aspectos, é possível delinear diversas maneiras de avaliar um estudante, no sentido de encontrar a melhor forma de atingir os objetivos propostos. Tais maneiras variam desde a Avaliação Diagnóstica, que identifica habilidades e deficiências do estudante de forma a inseri-lo em cursos adequados, até a Avaliação Continuada que emprega testes regulares que contribuem para a nota final ao longo de um curso. Veremos as características destas duas e também das avaliações formativas e finais nos próximos parágrafos.

Na Avaliação Diagnóstica, os testes têm como principal função posicionar adequadamente os estudantes nos programas de ensino que serão disponibilizados. Por essa razão, esses testes são empregados nos estágios iniciais do processo de ensino, preferencialmente antes do início dos estudos. Dessa forma, o potencial ou a deficiência de cada estudante pode ser identificado e, dados os recursos necessários, os mesmos podem ser direcionados para programas de estudo que melhor se adaptem às suas habilidades. Uma das vantagens desse método de avaliação é que em instituições de ensino que possuem uma grande variedade de cursos, desde níveis básicos até os avançados, pode-se ganhar tempo nas alocações corretas de estudantes, de maneira a fornecer uma preparação adequada a cada grupo.

A Avaliação Formativa corresponde a exercícios, tarefas e testes progressivos distribuídos aos estudantes durante um ano ou semestre acadêmico podendo também servir como auto-avaliação.

A principal proposta desse tipo de avaliação é fornecer uma retroalimentação regular aos alunos, estimulando seu aprendizado e dando informações que os ajudem a julgar suas estratégias individuais de estudo, também servindo para alertar os professores com relação a qualquer seção do curso que mereça uma atenção especial, na qual os estudantes podem ter dificuldades de compreensão. É preferível que qualquer nota ou resultado de todas as tarefas formativas (ou tarefas de auto-avaliação) tenha apenas influência marginal sem qualquer dano no resultado final.

Na Avaliação Final, a principal proposta é fazer um julgamento, usualmente definitivo, relativo à habilidade ou conhecimento de cada estudante. É a mais comum das avaliações e pode também ser definida como um instrumento de medida de desempenho de um determinado aluno ao final do curso, ou do nível de aproveitamento ao final de uma seqüência de estudo. Segundo Miller et al. (1998), a avaliação final atende a três propósitos principais. Primeiro, os resultados da avaliação final podem comprovar as habilidades que um aluno obteve durante um curso; segundo, esse tipo de avaliação é particularmente importante para aquelas instituições em que os programas acadêmicos conduzem os estudantes a um grau escolar com méritos e honras; terceiro, a avaliação final serve para comprovar o aprendizado de um indivíduo para qua assuma um determinado grau profissional.

Avaliação Continuada usa testes e trabalhos regulares ao longo de uma unidade de estudo, na qual os resultados de cada um contribuem para o resultado final. Uma das principais dificuldades encontradas na aplicação da avaliação continuada aparece quando suas tarefas constantes se sobrepõem às atividades de ensino, aumentando a sobrecarga de trabalho e fazendo com que os estudantes não tenham tempo para estudar o assunto tratado em profundidade (devido à ocorrência constante de testes).

1.1.2 As Avaliações Informatizadas

O desenvolvimento de novas tecnologias na área da informática, por exemplo a WWW (*World Wide Web*), a necessidade da expansão da educação utilizando cursos à distância, bem como a vasta disponibilidade de computadores pessoais, têm possibilitado uma mudança enorme no campo da avaliação, pois o computador pode participar tanto da construção como da administração de testes. De maneira geral, pode-se dizer que um teste informatizado é aquele que utiliza o computador como meio principal de apresentação das questões; coleta e armazena respostas; calcula e interpreta os resultados.

Nesse contexto, existem diferentes graus de informatização das tarefas de construção e aplicação de um teste. De acordo com Olea et al., (1999), o estado de automatização de um teste pode partir de um nível baixo, consistindo na apresentação de questões de forma tradicional, em folhas de papel, no qual suas respostas são marcadas em um cartão resposta e a correção se dá a partir de leitura óptica; até um nível mais alto, no qual as questões podem assumir novos formatos com o uso de recursos multimídia, além do processamento das respostas e o cálculo dos resultados serem realizados eletronicamente.

O uso abrangente de avaliações informatizadas foi possível com o uso dos testes objetivos (Brown, 1997). Tais testes requerem que o usuário escolha ou forneça uma resposta para uma questão cuja resposta correta é pré-determinada. As questões podem se manifestar nos formatos de múltipla escolha, verdadeiro-falso, combinação, e completar, dentre outras.

Dessa maneira, quanto à administração, as avaliações informatizadas podem ser implementadas de duas formas: por meio dos métodos convencionais ou métodos alternativos informatizados. Os métodos convencionais informatizados correspondem àqueles que implementam os testes tradicionais de lápis e papel em um sistema computacional, apresentando-os de forma eletrônica (uso do monitor de vídeo). Sua principal característica é que todas as questões apresentadas aos candidatos são as mesmas e seus resultados apresentam um resumo do conhecimento do indivíduo na área avaliada, sendo taxativo na apresentação dos resultados, classificando se o conhecimento de um estudante é certo ou errado, não permitindo a expressão de conhecimento parcial.

Por outro lado, os métodos alternativos informatizados consistem na implementação de técnicas que procuram avaliar de fato o "real" conhecimento de um **indivíduo**. A técnica MPA[1] consiste na tentativa de quantificar o conhecimento parcial de um estudante por meio de associações numéricas ponderadas a cada opção de resposta de uma questão, estabelecendo um sistema de pontuação efetivo, fixando um cenário de recompensas e penalidades associadas a cada alternativa (Aquino, 2001).

Ainda dentro dos métodos alternativos, uma das implementações mais elegantes no campo da avaliação informatizada são os Testes Adaptativos Informatizados – TAI's[2] que são o foco deste trabalho de mestrado. Estes testes empregam características específicas, como, por exemplo, o nível de dificuldade, o fator de adivinhação e a discriminação de cada uma das questões do teste, bem como a habilidade individual de cada participante para estimar o nível de capacidade e rendimento de um estudante (Olea et al., 1999).

A idéia fundamental dos testes adaptativos é ajustar os itens de um teste ao nível de habilidade individual de cada participante de uma avaliação. Em outras palavras, é apresentada uma sequência de questões que mais se adaptem ao conhecimento e capacidade de cada estudante. Assim, a principal característica que o define, é a que um teste não será o mesmo para todos indivíduos, possibilitando uma medida mais fiel a respeito da competência de cada um, bem como a economia de tempo de aplicação dos testes, já que relativamente poucos itens podem ser necessários para medir um certo grau de habilidade. A Teoria de Resposta de Itens (TRI[3]) é um conjunto de modelos estatísticos que descrevem tanto os itens de um teste, quanto as suas características (índices de dificuldade e/ou discriminação e/ou adivinhação), como os estudantes, quanto às suas habilidades ou competências (Baker, 1992; Hambleton and Swaminathan, 1985; Rudner, 1998).

Para otimizar e melhorar sua execução, os TAI's mais recentes utilizam um banco de itens, o qual é armazenado de forma estruturada, facilitando a busca e recuperação de itens durante um teste.

[1] Medida de Probabilidade Admissível - técnica de avaliação que mede o conhecimento parcial de um candidato.
[2] Computer Adaptive Testing - CAT em inglês
[3] Item Response Theory – IRT em inglês

Portanto, o uso da Teoria de Resposta de Itens, em conjunto com o banco de questões, permite a administração de testes a distintos indivíduos, considerando os traços intelectuais de cada um. Um banco de itens pode ser planejado e organizado de maneira a cobrir diferentes conteúdos de um mesmo domínio de conhecimento, ou seja, existem grupos de itens que podem ser aplicados em diferentes momentos, medindo diferentes habilidades, em uma ou mais avaliações (Kingsbury and Zara, 1991; Olea et al., 1999).

Dessa maneira, é necessário realizar uma calibração do banco de itens, ajustando-o a um modelo de resposta (pertencente à TRI), ou seja, é preciso escolher e avaliar quais os modelos de resposta mais indicados para que os objetivos de uma avaliação sejam atingidos. Da mesma forma é preciso definir os procedimentos de pontuação, que são responsáveis pela conversão do desempenho observado (pontuação verdadeira) para uma escala conveniente, utilizada para classificação dos estudantes.

A seleção de itens, a calibração do banco e a estimação da habilidade de um indivíduo também são atividades inerentes aos testes adaptativos. A seleção de itens corresponde à tarefa de selecionar o próximo item a ser administrado ao estudante, dado seu nível de habilidade corrente, durante um teste. As abordagens que merecem destaque (por serem as mais atuais e produzirem melhores resultados) são a de *Máxima Informação* (Hambleton and Swaminathan, 1985; Kingsbury and Zara 1989; Olea et al., 1999), que escolhe um item do banco que fornece o maior índice de informação a um nível de habilidade individual e a *Bayesiana* (Owen, 1975; Kingsbury and Zara, 1989) que é muito parecida com a anterior, porém atribui um intervalo de confiança a cada estimação de habilidade calculada, selecionando o item do banco que mais reduz o valor desse intervalo. Tais métodos podem ser ajustados ao conteúdo do banco de itens permitindo uma avaliação unidimensional ou multidimensional. Modelos unidimensionais assumem que todos os itens do banco foram ajustados para medir um única habilidade, entretanto, com a criação de banco de itens implementando abordagens como o balanceamento de conteúdo (Content-Balanced) e *Testlets* (Kingsbury and Zara 1989; Huang 1996) é possível criar avaliações multidimensionais medindo várias habilidades.

A calibração do banco consiste na tarefa de estimar os parâmetros dos itens (a e/ou b e/ou c) identificando suas características e possibilitando seu ajuste a um modelo de resposta da TRI. A atividade de estimação pode ser realizada de várias formas. Conforme apresentado nos trabalhos de Huang (1996) e Rios et al. (1998), tal atividade pode ser realizada pela atribuição pré-determinada dos valores dos parâmetros, sendo esta conduzida por um projetista do teste baseado em seu conhecimento prévio dos itens; ou ainda, por meio de um processo puramente matemático, utilizando programas específicos para esta tarefa, baseados na implementação de algoritmos de estimação, de maneira que os valores dos parâmetros são obtidos pela inferência estatística sobre amostras de respostas fornecidas anteriormente aos itens (Bilmes, 1998; Wooddruff and Hanson, 1997; Hanson, 1996; Olea et al. 1999).

Os métodos de estimação de habilidade têm a responsabilidade de estimar a habilidade de um estudante durante um teste após o mesmo ter respondido a um determinado item, e fornecer o novo valor da habilidade para que a seleção do próximo item ocorra, e assim sucessivamente, até

que o critério de parada seja alcançado. O método mais conhecido é o da Estimação de Máxima Verosimillhança[4] (EMV), que é baseado num somatório de logarítmos naturais das probabilidades de resposta correta e incorreta dos itens administrados em um teste, para um determinado estudante (Baker, 1992; Hambleton and Swaminathan, 1985).

A teoria dos testes adaptativos, englobando seus modelos de resposta, as abordagens de seleção, métodos de estimação de habilidade, e a calibração e organização do conteúdo do banco de itens, parece ser a forma mais justa e honesta de avaliar um indivíduo num processo educativo. No entanto, a prática e a implementação desses modelos depende do acerto e cuidado de muitos detalhes inerentes a este processo. Assim, é necessário um esforço conjunto e coordenado, com o envolvimento de vários especialistas de diferentes áreas (estatísticos, analistas e educadores), além da necessidade de boas ferramentas (computadores e aplicativos), para promover uma boa avaliação. Não podemos esquecer, é claro, que todas as atividades necessárias para a criação e implantação de uma avaliação adaptativa possui um custo, relativamente alto, já que abrange várias especialidades e equipamentos.

Como exemplo de uso dessa abordagem, podemos citar o *ETS (The Educational Testing Service* – www.ets.org), organização responsável pelo exame de proficiência em inglês TOEFL[5], que a partir de 1998 teve seus módulos de avaliação *Structure* e *Listening* realizados sob a abordagem adaptativa (Zinn, 2000). Nestes módulos, não é permitido que os candidatos revisem suas respostas uma vez respondidas, já que as mesmas servirão para estimar a habilidade corrente e selecionar o próximo item a ser administrado. Outro exemplo é a existência de grandes empresas que possuem os testes informatizados, incluindo os adaptativos, como meta principal de desenvolvimento e distribuição, como a *CATGlobal Systems* (www.catinc.com), e a *Assessement Corporation Systems* (www.assess.com), que são responsáveis pelos pacotes de programas **CATGlobal** e **FastTest Professional**, respectivamente, que contém ferramentas de construção, administração e análise de testes informatizados. Estes programas são comerciais e possuem preço de aquisição e licenças de uso bastante elevados. Ainda, a empresa *Prometric Testing Center* (www.prometric.com) é responsável pelo desenvolvimento do **Prometric** que implementa os testes adaptativos na **IEEE**[6] para a avaliação dos profissionais de computação.

Entretanto, existem alguns pequenos programas que implementam, a título de experimento e avaliação, alguns dos modelos e métodos utilizados pelos testes informatizados, em especial os adaptativos. Como exemplo, existe o **AdTest** (Olea et al., 1999) que implementa o modelo de seleção de itens de Máxima Informação e o método EMV como estimação de habilidade. A maioria destes programas são de livre acesso e podem ser obtidos através de seus autores.

1.1.3 Avaliação Informatizada do Exame de Proficiência em Inglês (EPI) do ICMC

Até o início de 2000, os estudantes que ingressavam no programa de mestrado do ICMC–USP eram avaliados com relação à proficiência em inglês por meio de testes preparados a cada semestre por um

[4]Maximum Likelihood Estimation - MLE em inglês
[5]Test of English as a Foreign Language
[6]Institute of Electrical and Electronics Engineers – www.ieee.org

professor diferente. Na prática, o teste se constituía na tradução de textos tomados de revistas ou livros técnicos escritos em inglês (L2) para a língua nativa (português) (L1) visando à compreensão do conteúdo, e, em casos raros, da produção da versão em inglês de outro texto de literatura técnica, originariamente na língua nativa, ou da produção de um texto em inglês de autoria própria.

A correção dos testes de proficiência era custosa e consumia bastante tempo dos professores que já possuiam inúmeras atividades a serem cumpridas. Além disso, a configuração na qual o teste era preparado não é ideal, pois exames distintos podem ser fortemente não uniformes devido à rotatividade de professores responsáveis, e também à subjetividade inerente à construção e correção humanas. Testes que requerem somente traduções de L2 para L1, e vice-versa, são ainda mais problemáticos e limitantes, pois não são capazes de avaliar bem a competência na escrita de artigos e abstracts em inglês relatando a pesquisa desenvolvida, atividades que são essenciais para um pesquisador. A consciência do gênero, isto é, a consciência da estrutura e convenções da língua para o gênero de textos científicos, como expressa Swales (1990), não é incentivada, o que é essencial para o pesquisador novato desempenhar melhor e mais rápido as tarefas de leitura e escrita relacionadas à sua pesquisa. Artigos científicos são fortemente convencionalizados (Weissberg and Buker, 1990) e o desconhecimento das convenções pode prejudicar o relato e divulgação da pesquisa científica, bem como a leitura precisa e mais rápida das pesquisas em uma área.

Uma opção para este cenário poderia ser a exigência do TOEFL ou IELTS, que já são requeridos para os alunos de doutorado do ICMC–USP, também para os alunos de mestrado. Enquanto que estes exames são bastante completos, pois testam as habilidades de leitura/escrita, audição, conversação, e gramática eles também não tratam do tipo de texto mais importante para um futuro pesquisador – o artigo científico. Exames como o TOEFL exigem uma habilidade extra do aluno: rapidez na resposta, característica que gera a necessidade do aluno se preparar para passar neste tipo de teste. Existem cursos preparatórios para estes exames, exigindo uma grande dedicação do estudante e até total imersão nesta atividade, fazendo com que ele pare ou reduza o tempo destinado à pesquisa.

Para modificar este cenário, a orientadora deste mestrado propôs um novo tipo de exame de proficiência em inglês (instrumental)[7] que foi implantado no formato lápis e papel no segundo semestre de 2000, e no formato computadorizado no primeiro semestre de 2001. O novo exame possui dois propósitos: 1) avaliar as habilidades de: estruturar (reconhecer as partes componentes de cada seção de um artigo), compreender e reconhecer as relações entre idéias de um texto, perceber as várias estratégias utilizadas para escrever um mesmo trecho de uma seção e conhecer as convenções da língua para o gênero de textos científicos; e 2) tentar sanar os problemas de custo, tempo, falta de padronização, subjetividade e ausência de adequação dos exames tradicionais de proficiência em inglês para admissão em programas de mestrado, particularmente o do ICMC-USP. O exame proposto para o mestrado é composto de quatro módulos (M1, M2, M3 e M4) descritos na Tabela 1.1.

O trabalho de mestrado de Aquino (2001) trata da criação dos tipos de questões para o EPI do ICMC–USP (avaliação final) utilizando a Medida de Probabilidade Admissível (Bruno, 1987), e da

[7]Também chamado de Inglês para Propósitos Acadêmicos, EAP – English for Academic Purposes, em inglês.

Tabela 1.1: Divisão dos módulos do EPI

Módulos do Exame de Proficiência em Inglês (EPI)
M1: Convenções da Língua Inglesa para Textos Científicos Questões de múltipla escolha abordando morfologia, vocabulário, sintaxe, tempos verbais, marcadores de discurso, etc. utilizados nas partes de cada seção de um artigo científico em inglês.
M2: Estrutura de Textos Científicos Questões de múltipla escolha tratando das funções de cada seção de um artigo.
M3: Compreensão de Texto Questões de múltipla escolha tratando da compreensão e do reconhecimento das relações entre as idéias contidas em uma seção de um artigo.
M4: Estratégias de Escrita Questões de múltipla escolha abordando estratégias de escrita como, por exemplo, definições, descrições, classificações, argumentações.

implementação de um sistema WEB (CAPTEAP[8]) para a aplicação do exame. Neste sistema, as questões de usabilidade da interface e segurança do mesmo foram tratadas. No presente mestrado, abordamos a avaliação diagnóstica do inglês instrumental para sinalizar os problemas que alunos ingressantes no mestrado possam ter. Espera-se alertar os alunos dos benefícios que um curso de inglês, ou um estudo individualizado possa propiciar ao bom desempenho da pesquisa científica. Futuros desbobramentos desta pesquisa incluem o desenvolvimento de um *site* com a avaliação final, diagnóstica, ferramentas para a escrita de textos científcos e material *on-line* tratando do inglês para propósitos acadêmicos.

1.2 Motivação e Relevância

O advento da WWW permitiu o surgimento de novas formas de interação, utilizando recursos multimídia como áudio, vídeo, imagens gráficas e diversas maneiras de comunicação. Assim, os benefícios do uso desse ambiente puderam ser aplicados diretamente aos moldes de ensino via Internet, em que os relacionamentos interpessoais e o acesso ilimitado a variados meios de informação servem de base para redimensionar os modelos educacionais virtuais. A WWW favorece a criação e manutenção de diversos cursos "on-line" (via Internet), além de estimular o desenvolvimento de ambientes de autoria como por exemplo o AulaNet[9] e o TelEduc[10]. Tais ambientes de autoria têm como principal função a disponibilização de recursos (tarefas, atividades pedagógicas e conteúdos instrucionais) da maneira mais didática que lhes couber, possibilitando a implantação do ensino a distância. No entanto, a maioria das avaliações (quando possuem) inseridas nesses ambientes

[8]Disponível em: http://www.nilc.icmc.sc.usp.br/staff/capteap/

[9]Ambiente Virtual de ensino da Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio (http://anauel.cead.puc-rio.br/aulanet/)

[10]Ambiente Virtual de ensino da Universidade de Campinas - Unicamp (http://hera.nied.unicamp.br/teleduc/pagina_inicial/index.msql)

ainda seguem o modelo de avaliação informatizada convencional. O método adaptativo de avaliação informatizada pode ser inserido no cenário desses ambientes educacionais, com o intuito de melhor adequar o processo de avaliação ao conteúdo por eles transmitido. Também, a avaliação adaptativa pode ser inserida em diferentes meios de aprendizado, fazendo parte, tanto da avaliação diagnóstica como final da proficiência em inglês para admissão em programas de mestrado.

Sob um ângulo otimista, a avaliação adaptativa é uma boa opção para medir o alcance dos objetivos de um processo educacional ou de treinamento, principalmente quando é apoiada por um banco de itens bem elaborado. Entretanto, há dois obstáculos que fazem o emprego dos testes adaptativos informatizados ser difícil. Segundo Huang (1996), as principais dificuldades de implantar um teste adaptativo eficiente são a adoção de um estudo empírico de larga escala para uma boa calibração dos itens do banco e a geração de avaliações multidimensionais que atendam a diferentes conteúdos (módulos) de uma mesma área de conhecimento.

1.3 Objetivos

Baseado nas teorias e cenários até aqui expostos, entendemos que é possível inserir os testes adaptativos informatizados num contexto educacional, tornando mais eficiente a tarefa de avaliar, identificando as reais necessidades dos programas de ensino, bem como o perfil dos estudantes desses programas.

As propostas definidas na concepção deste projeto de mestrado eram compreender a teoria e os formalismos relacionados aos testes adaptativos e posteriormente, desenvolver uma aplicação que instituísse a avaliação adaptativa diagnóstica da proficiência em inglês no programa de mestrado do ICMC. Contudo, no decorrer dos estudos ficou claro que a qualidade de uma avaliação adaptativa dependia diretamente da existência de um banco de itens calibrado e robusto que fornecesse suporte para a implementação desta avaliação. Logo, nossa atenção se voltou para a criação e desenvolvimento de um banco de itens referente a este exame de proficiência, já que este pode ser considerado como elemento central de um sistema adaptativo.

Assim, os objetivos desta dissertação de mestrado consistem em: primeiro, popular e calibrar um banco de itens sob a abordagem adaptativa que contemple os diversos módulos do Exame de Proficiência em Inglês (EPI) do ICMC–USP; segundo, desenvolver a modelagem de um sistema de avaliação adaptativa que sirva como avaliação diagnóstica da proficiência da língua inglesa, no cenário específico do programa de mestrado deste instituto. Adicionalmente, também foi desenvolvido um ***Sistema de Gerenciamento Banco de Itens (SisBI)***, com o propósito de auxiliar a manutenção, gerência e expansão do mesmo.

Resumindo, do ponto de vista teórico este trabalho teve o propósito de estudar a teoria dos testes adaptativos, principalmente as técnicas de estimação de parâmetros voltadas para a calibração do banco de itens, os métodos de seleção de itens sob uma base sensível ao conteúdo, os procedimentos de estimação da habilidade, bem como os métodos de conversão da habilidade observada (no final do teste) para a escala de pontuação conveniente (mais comumente usada). Do ponto de vista prático,

este trabalho se propôs a popular e calibrar um banco de itens sensível ao conteúdo e modelar um sistema que implementa a avaliação diagnóstica da proficiência em inglês para programas de mestrado.

1.4 Organização do Trabalho

O presente trabalho está organizado da seguinte forma: o Capítulo 2 faz uma revisão bibliográfica dos testes objetivos informatizados, descrevendo suas abordagens, principais diretrizes, vantagens associadas, além da revisão das principais tarefas que envolvem a criação, construção e manutenção de um banco de itens.

No Capítulo 3 é realizado um estudo da evolução histórica e propriedades dos testes adaptativos informatizados, sendo ressaltadas também suas questões técnicas e limitações. O Capítulo 4 trata dos fundamentos da Teoria de Resposta de Itens, realçando seus conceitos e características, bem como os modelos de respostas, as abordagens de seleção de itens, os métodos de estimação de habilidade e os procedimentos de pontuação empregados numa avaliação. Uma descrição dos principais programas que implementam os testes adaptativos é feita no Capítulo 5, efetuando ainda a análise de dois programas abordando suas funcionalidades, desvantagens e uma comparação dos resultados de suas respectivas execuções. No Capítulo 6 são descritas as principais atividades desenvolvidas neste trabalho para a criação do banco de itens sensível ao conteúdo referente às questões do Exame de Proficiência em Inglês, a descrição da ferramenta **SisBI** de gerência do banco, bem como a modelagem do sistema adaptativo e a discussão dos trabalhos relacionados a este tema. Finalmente, no Capítulo 7 são apresentadas as conclusões e os trabalhos futuros relacionados a este trabalho.

CAPÍTULO 2

Testes Objetivos Informatizados

Um teste informatizado é aquele que utiliza o computador como meio principal de apresentação das questões, entrada das respostas e interpretação dos resultados, ou ainda, como meio de construção dos testes.

Nesse contexto, é conveniente advertir que existem diferentes graus da automatização dos processos de aplicação de um teste. São eles (Olea et al., 1999):

1. Procedem a partir de um nível mínimo de automatização, consistindo na apresentação das questões na forma tradicional, em folhas de papel, nas quais suas respostas são atribuídas mediante um cartão resposta e a correção se dá a partir de leitura óptica;

2. Consistem na apresentação das questões no formato tradicional, sendo que tanto as respostas como o processamento dos dados e até mesmo a estimativa do nível de habilidade de um estudante, são realizadas pelo computador;

3. Apresentam o nível máximo de automatização, pois incluem a apresentação das questões e suas instruções de forma eletrônica, atendendo as necessidades dos formatos que às tarefas apresentam, como recursos de áudio e vídeo.

Este capítulo estuda o impacto da informatização sobre a forma de construir, aplicar e analisar os testes objetivos. Pretende-se analisar quais os requisitos necessários que um teste informatizado deve possuir para que seja considerado adequado e destacar de que maneira suas propriedades vêm sendo afetadas. Como conclusão geral, pode-se dizer que a informatização dos testes fornece numerosas vantagens sobre os testes de lápis e papel, mas, por sua vez, não garante a sua qualidade integral, pois com ela aparecem novos problemas e situações que exigem soluções mais específicas.

2.1 Histórico

As primeiras experiências com os testes objetivos informatizados, desenvolvidas sob a expressão genérica *Computer Assisted Assessment* (CAA – em inglês), pretenderam dar suporte aos problemas

associados à correção e interpretação dos resultados de testes convencionais, baseados no formato de lápis e papel[1], e teve como principal objetivo agilizar a elaboração de relatórios informativos legíveis para o nível profissional dos organizadores dos testes.

Nos anos trinta foram realizadas as primeiras correções automatizadas mediante o uso do computador, um difícil procedimento baseado em cartões perfurados que continham as respostas das questões de um teste. Na década de quarenta, o desenvolvimento dos computadores analógicos começaram a dar suporte aos procedimentos automatizados de pontuação e correção, e na confecção de relatórios.

A partir dos anos cinqüenta, em substituição aos cartões perfurados, começaram a ser desenvolvidos sistemas que utilizavam folhas de resposta eletrônica, as quais usavam a leitura óptica para realizar as correções. Nos anos seguintes, principalmente na década de setenta, devido principalmente a capacidade e rapidez dos primeiros computadores digitais, começam a ser desenvolvidos exemplos de testes de aplicação automatizada, sendo estes apresentados no monitor de vídeo e respondidos mediante o uso do teclado.

A tecnologia digital dos computadores facilitou a criação de testes psicológicos, que requerem uma apresentação de itens mais complexos, principalmente aqueles que têm como objetivo a medida do nível de inteligência e aptidão de uma pessoa. Assim, durante a década de oitenta, foram produzidas as primeiras versões de testes informatizados sobre diversos conteúdos, aplicados principalmente nos contextos escolares, ligados intimamente à avaliação de problemas intelectuais, motivacionais ou de conduta, que podem ser a base das dificuldades de aprendizagem.

A partir do desenvolvimento das abordagens de avaliação acima descritas, iniciou-se o desenvolvimento de procedimentos de testes preocupados principalmente com a medida de diversas habilidades pessoais que até a bem pouco tempo era estudo exclusivo dos laboratórios de psicologia experimental cognitiva (Olea et. al., 1999). Tais procedimentos são a base para a criação e desenvolvimento dos testes adaptativos informatizados e da geração automática de itens (duas vertentes dos testes informatizados) que, por sua vez, incorporam novos métodos de medida de resultados, sustentados em modelos cognitivos de processamento da informação que representam uma nova concepção dos objetivos de uma avaliação automatizada.

2.2 Tipos e Estruturas dos Testes Objetivos

Os testes objetivos são os principais tipos de teste que permitem a aplicação e utilização do computador no processo de avaliação. Isso porque sua estrutura permite o uso de mecanismos lógicos para sua correta administração e manutenção.

Os testes objetivos requerem que o estudante escolha ou forneça a resposta a uma questão sendo que a(s) resposta(s) correta(s) já está(ão) previamente definida(s). Quando projetados corretamente, esses testes podem ser sensíveis, a ponto de identificar e testar a classificação de

[1] Forma tradicional de teste, na qual a apresentação das questões é em uma folha de papel e os alunos respondem de forma manuscrita.

habilidades e também reduzir a possibilidade do estudante obter sucesso por acaso e não como reflexo do seu conhecimento. Segundo Brown et. al., (1999), existem algumas diretrizes para a produção de questões eficazes, tais como:

- se os resultados do aprendizado foram identificados para o curso, então os testes devem atacá-los;
- o teste deve tentar avaliar o estudante ao longo de todo o curso, não apenas em parte dele;
- no teste, a atenção dedicada a um tópico em particular deve ser reflexo de sua importância no curso;
- o teste deve ser integrado ao projeto do curso, e não tratado a parte e posteriormente.

Em geral, as questões utilizadas nos testes objetivos apresentam os seguintes elementos, mostrados na Figura 2.1:

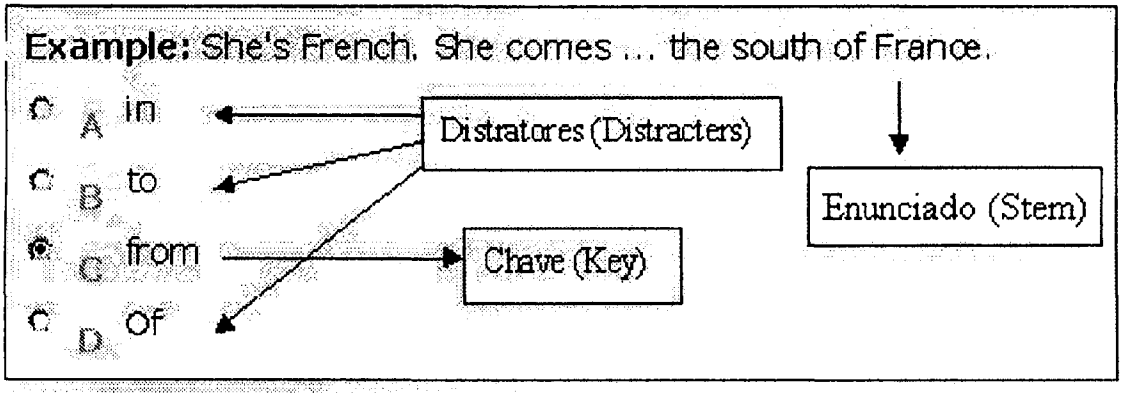

Figura 2.1: Partes que compõem uma questão de múltipla escolha

Enunciado (Stem): é o enunciado da questão, a sua requisição;

Opções (Options): é uma lista de todas as alternativas propostas para o estudante que contém a resposta correta e as incorretas;

Chave (Key): é(são) a(s) resposta(s) correta(s) pertencente(s) à lista de Opções;

Distrator (Distracter): são todas as respostas incorretas da lista de Opções.

Com base nessa forma geral, podem ser criados e estabelecidos vários tipos de questões. São eles:

1 - ***Múltipla Escolha:*** é um tipo de questão que contém um Enunciado seguido por uma lista de Opções com vários itens (geralmente cinco), sendo que um deles é a resposta correta (Chave) e os restantes são Distratores. O papel do aluno é escolher a melhor resposta, ou a que julgar ser correta;

2 - ***Resposta Múltipla:*** é uma variação do item anterior, em que existe ou pode haver mais de uma resposta correta;

3 - ***Verdadeiro/Falso:*** é outro tipo especial de questões de múltipla escolha, na qual há duas opções de escolha, verdadeiro ou falso. Tais questões têm como principal objetivo medir a habilidade do aluno em discernir se as afirmações dadas são verdadeiras ou não;

4 - ***Completar:*** é um tipo de questão que o aluno responde completando uma sentença, composta de um ou vários espaços em branco. Da forma simplificada, a questão se apresenta incompleta, de forma que a definição geral é dada e se omite a(s) palavra(s) chave(s);

5 - ***Completar Múltiplo:*** é um modelo de questão que combina as questões do tipo Verdadeiro/Falso e Completar, vistas anteriormente. O estudante deve assinalar no sentido de completar cada alternativa e logo depois classificá-la como verdadeira ou falsa. Esse tipo de questão exige maior habilidade do aluno, já que todas as alternativas exigem uma resposta;

6 - ***Seqüência:*** é um tipo de questão que oferece um conjunto de declarações e sentenças ao estudante, com o intuito de estabelecer uma seqüência lógica ou cronológica entre as mesmas. Em alguns casos, é interessante fornecer esse tipo de questão, para que se estimule a percepção do aluno para uma hierarquia de conceitos de um determinado domínio;

7 - ***Combinação/Correspondência:*** é uma questão em que a tarefa do estudante é fazer uma combinação entre os itens de duas listas fornecidas. É possível construir testes de questões de correspondência com o objetivo de obter medidas de conhecimento dos níveis mais elevados do conhecimento tendo a finalidade de testar a capacidade do estudante em recuperar e reconhecer um assunto;

8 - ***Análise de Relação/Causa-Efeito:*** é um tipo de questão que testa a capacidade de raciocínio do estudante. Consiste no fornecimento de um Enunciado e de uma Razão ao aluno, cuja tarefa é verificar a veracidade das sentenças e qual a relação entre elas.

De acordo com Miller et al., (1998), os testes objetivos necessitam de um julgamento prévio, antes de serem disponibilizados aos estudantes. Tal julgamento tem o propósito de:

- verificar se tanto os enunciados quanto as opções de resposta são claras e de fácil entendimento e se não existe mais de uma opção correta e se não há ambigüidades;
- analisar se a questão é importante o suficiente para ser respondida, se a informação fornecida é suficiente para obter a resposta e se a questão é precisa e gramaticalmente correta;
- verificar se todas as respostas são plausíveis e consistentes com o enunciado;
- verificar se cada item possui o mesmo número de respostas e se cada resposta tem início numa nova linha;

- observar a classificação do item de maneira a identificar o tipo de aprendizado existente, a qual seção do conteúdo pertence o teste e qual o nível de dificuldade.

2.3 Abordagens

Conforme visto na subseção anterior, existem diversas maneiras da informática ser aplicada no contexto dos testes. Desde sua aplicação na tarefa de correção automática, passando pela interpretação dos resultados, até na elaboração de relatórios específicos, a informática tem prestado um auxílio valioso.

A participação do computador no processo de criação e administração dos testes assume vários modelos de informatização. Dentre todas as modalidades da informática que podem ser introduzidas no mundo dos testes, destacam-se aqui duas linhas gerais: os testes convencionais informatizados e os testes alternativos informatizados, sendo que neste último, destacam-se a Medida de Probabilidade Admissível (MPA) e os Testes Adaptativos.

2.3.1 Métodos Convencionais Informatizados

Na implementação desses testes, o número de questões é o mesmo para todos os participantes, todos os itens têm a mesma dificuldade e são apresentados na mesma seqüência. Alguns autores consideram os testes convencionais como sendo a implementação dos testes tradicionais de lápis e papel em um sistema computacional, de maneira que se possa apresentá-los de forma eletrônica (via monitor de vídeo) e responder a eles por meio do teclado. À primeira vista, tal implementação não mudou em nada a essência dos testes tradicionais, mas sob uma análise mais detalhada, percebe-se que o uso do computador fornece uma série de vantagens instrumentais, como por exemplo, a apresentação precisa e controlada dos itens, correção automatizada das respostas, possibilidade de computar cálculos estatísticos e rapidez na divulgação dos resultados.

Em linhas gerais, o uso do computador na aplicação dos testes convencionais permite a utilização de itens de teste mais complexos e mais próximos da realidade, podendo até utilizar simulações e outros recursos como áudio e vídeo, estabelecendo possíveis melhoras na interação entre os estudantes e o teste. Enfim, quando é possível usar o computador para aplicar os testes, as vantagens instrumentais são enormes, mesmo que estas tenham um custo maior.

Não há dúvidas que existem várias vantagens operacionais no uso do computador na aplicação de testes tradicionais. Nesse cenário de convivência entre os testes convencionais informatizados e o formato de lápis e papel, a dúvida que surge é se os resultados de ambas aplicações são estritamente os mesmos, ou se a mudança introduzida pelo uso do computador altera as propriedades psicométricas dos testes.

A priori, tem-se a impressão de que a única mudança radical nos testes, dado à sua informatização, acontece na forma de apresentação das questões (via monitor de vídeo) e na maneira de respondê-las, pois a essência do conteúdo do teste continua inalterada. No entanto, o uso dos computadores na realização de um teste pode causar diferentes impactos nos estudantes.

Isso se deve ao fato de que muitas vezes os alunos não estão familiarizados com o uso do computador, fazendo com que surja uma ansiedade anormal que prejudique sua interação com o sistema e suas atitudes perante a máquina. Segundo Olea et al., (1999), um dos aspectos chave é que o *software* de avaliação utilizado deve permitir que o estudante revise e troque, caso necessário, as respostas das questões de um teste, assim como é permitido nos testes tradicionais de lápis e papel, além de permitir também que o mesmo distribua o tempo de resposta ao seu modo, evitando que exista tempo limitado para responder uma dada questão. Portanto, quando a informatização de um teste segue os rigores operacionais e é bem elaborada tecnicamente, o padrão de respostas eletrônicas não difere dos convencionais, diluindo os efeitos da mudança de apresentação. Em qualquer caso, as diferenças de nível de ambos os tipos de aplicação não seriam um problema insolúvel, pois na atualidade existe tecnologia eficiente para equiparar as pontuações recebidas em cada teste, fazendo com que a confiabilidade e validade dos testes sejam equivalentes.

A principal característica desses testes é que seus resultados apresentam um resumo do conhecimento na área avaliada, dizendo se o conhecimento de um estudante é certo ou errado. Dessa forma, ao avaliar um estudante, o teste informa se o mesmo sabe ou não um determinado conteúdo, sendo incapaz de fornecer informações de conhecimento parcial.

A maior desvantagem desse método está no fato de que a avaliação do conhecimento realizada por ele é distorcida e tendenciosa, pois existe ainda uma grande discrepância entre pontuação observada e verdadeira, pois nelas prevalecem as adivinhações e truques dos estudantes para acertar as respostas. Portanto, esse tipo de teste não pode proporcionar informação parcial dos testes, pois ele não possibilita respostas parciais.

2.3.2 Métodos Alternativos Informatizados

Existe contudo, um método alternativo de avaliação que mede os estados de informação parcial para avaliar a proficiência dos estudantes, e outro método em que as questões são selecionadas à medida que o teste procede. Tais métodos são expostos abaixo.

2.3.2.1 Medida de Probabilidade Admissível (MPA)

Dada a necessidade de uma avaliação mais útil da proficiência de um estudante, a Medida de Probabilidade Admissível procura valorizar o conhecimento parcial do mesmo, adotando estratégias que tentam perceber informações que os alunos possuem e que possam induzir a uma resposta correta ou não. Dessa forma, os cálculos envolvidos nesse método tentam quantificar o conhecimento do estudante de modo a classificá-lo de acordo com o seu nível de informação verdadeira, seja ele total ou parcial, sendo suas classes divididas em: Totalmente Informado, Quase Informado, Mal Informado, Parcialmente Desinformado ou Parcialmente Informado.

Segundo Aquino (2001), o método MPA é composto por questões objetivas, compostas de três alternativas de resposta A, B e C, sendo duas delas incorretas e um correta. Essas alternativas são combinadas de maneira a criar um espaço de representação das possíveis classificações (acima

citadas) designadas aos alunos. Dessa forma, são criadas alternativas auxiliares de resposta que representam o conhecimento parcial de um candidato. Para cada uma das alternativas auxiliares são atribuídos diferentes valores, ou pesos, de maneira a estabelecer um sistema de pontuação efetivo, impondo um cenário de recompensas e penalidades associadas a cada opção, dependendo da distância espacial entre essas e a resposta correta.

A Figura 2.2 ilustra as características espaciais das alternativas de uma questão, dado que: B e C são incorretas e A correta.

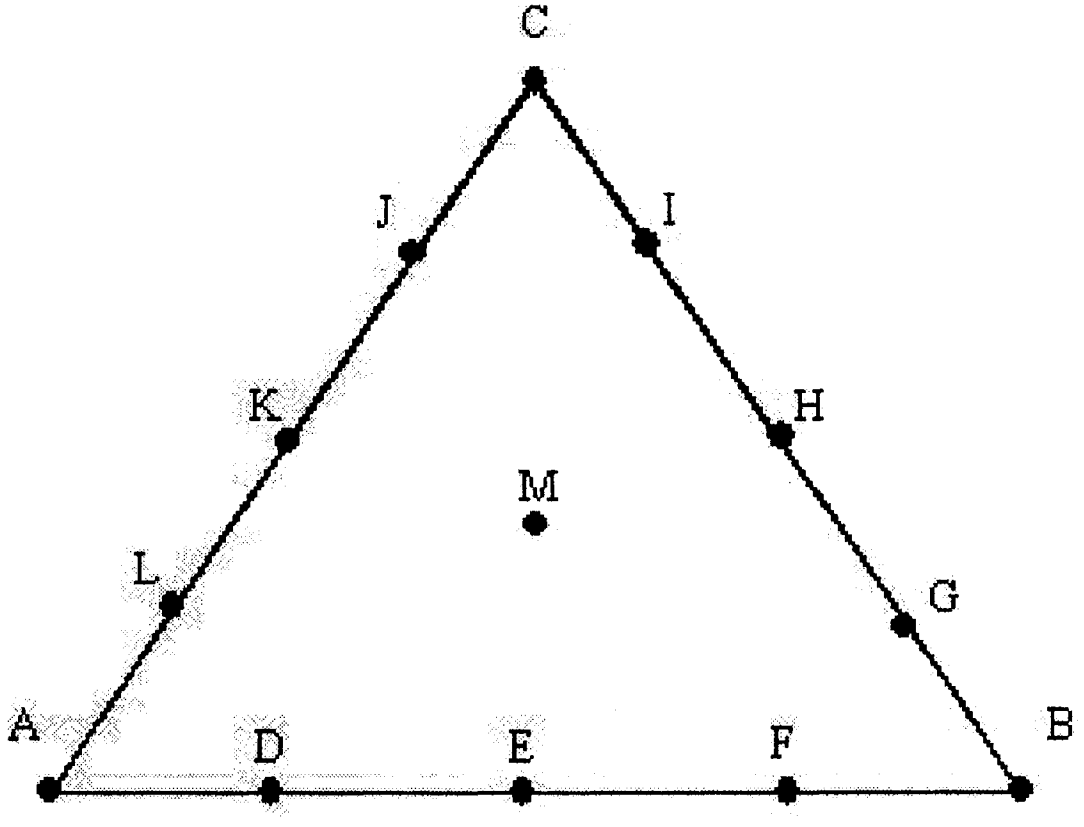

Figura 2.2: Modelo Espacial de distribuição das opções de resposta do método MPA

Cada vértice do triângulo representa uma alternativa de uma questão e corresponde à distribuição de probabilidade sobre as três escolhas possíveis de resposta em um teste particular (alternativas A, B e C). As outras opções (D, E, F, G, H, I, J, K, L e M), são as alternativas auxiliares e representam a distribuição do conhecimento parcial. O modelo da Figura 2.2 é um modelo computacional adaptado ao método MPA, proposto Bruno (1987).

Essa versão computadorizada mostra na tela a questão e o triângulo de respostas, associando as pontuações de recompensa, como mostrado na Tabela 2.1, a cada alternativa proposta. Portanto, o estudante escolhe as alternativas entre A e M que representam sua resposta e respectiva pontuação.

Tabela 2.1: Recompensa relativa às opções de escolha e tipo de classificação (Aquino, 2001)

Opção de Resposta	Pontos	Tipo de Classificação
A	+24	Totalmente Informado
D ou L	+19	Quase Informado
E ou K	+9	Parcialmente Informado
B, I, H, G e C	-76	Mal Informado
F ou J	-11	Parcialmente Desinformado
M	0	Desinformado

A principal vantagem do uso da MPA é que os professores podem colher, além das características psicométricas, uma quantidade enorme de detalhes e informações adicionais dos alunos, que são especialmente úteis para as avaliações de diagnóstico e para a articulação de currículos, a fim de melhorar programas instrucionais, bem como identificar os problemas específicos de aprendizado.

De acordo com Aquino (2001), a análise de tais informações também fornece dados importantes para professores e alunos dos sistemas de avaliação automatizada. Para os alunos a MPA fornece:

- análise de proficiência;
- crédito parcial para conhecimento parcial;
- honestidade no teste, pois minimiza a necessidade de adivinhações ou chutes, devido à inclusão de alternativas que representam o conhecimento parcial.

Para os Professores:

- informações individuais (para cada estudante) e da classe, permitindo um acompanhamento personalizado e identificando problemas específicos e gerais;
- permite uma adequação de currículos escolares dado o nível de informação obtida após a realização de um teste;
- auxilia na articulação de currículos para nutrir programas educacionais baseados no conteúdo da informação;

Baseados nos estudos de Bruno (1987) existem duas limitações quando se usa o método de avaliação MPA. Uma delas está centrada no fato de que as questões objetivas que compõem um teste contém apenas três alternativas (A, B e C), o que exige que se adapte questões de outros testes ao formato MPA. A outra diz respeito ao treinamento do corpo docente que tem função de aplicar o teste, bem como, e principalmente, dos alunos que o farão, uma vez que a MPA apresenta inovações no seu formato.

Desde a utilização dos sofisticados procedimentos da MPA, algumas pesquisas têm sido desenvolvidas na busca de soluções que tentam ajustar a quantidade de itens administrados em um

teste, dado o nível de habilidade de um estudante. Assim, alguns trabalhos envolvendo Testes Adaptativos sugerem que menos tempo instrucional seja necessário para atingir um determinado grau de proficiência individualizado. Tais testes serão objetos de estudo da próxima subseção.

2.3.2.2 Testes Adaptativos

Os Testes Adaptativos Informatizados – TAI – (Computer Adaptive Testing – CAT – em inglês) são o resultado da simbiose entre os avanços da informática no campo dos testes e as contribuições dos modelos estatísticos e psicométricos da Teoria de Resposta de Itens – TRI – que será abordada com mais detalhes no Capítulo 4.

A noção fundamental dos testes adaptativos vem do desejo de adaptar os itens de um teste ao nível de habilidade individual de cada estudante. Idealmente, para cada indivíduo, seria dado um teste cuja apresentação das questões é realizada de acordo com o seu nível de competência. A característica principal que o define é que um teste não será o mesmo para todas as pessoas, adaptando-se ao nível de habilidade pessoal em um determinado domínio de conhecimento.

A grande vantagem da estratégia de adaptar os testes ao nível do indivíduo examinado está no fato de conseguir uma melhor precisão de medida a respeito da capacidade de cada um, bem como na economia de tempo de aplicação de testes, já que é necessária a administração de poucos itens para alcançar a proficiência do indivíduo.

De acordo com Olea et al., (1999), pode-se dizer que os TAI's constituem a aplicação mais poderosa da informática no âmbito dos testes. Mesmo que outras estratégias de avaliação automática sejam possíveis, a execução dos testes adaptativos, em geral, se dá a partir de um banco de itens no qual o computador acessa e vai aplicando ao indivíduo um item de cada vez, elegendo aqueles que melhor se ajustam ao nível de habilidade que vem demonstrando a pessoa avaliada. Quais itens eleger, quando parar a exposição e como estimar a pontuação dos indivíduos nos testes são os detalhes técnicos dos TAI's, inerentes à Teoria de Resposta de Itens, e serão abordados no Capítulo 4.

Entretanto, o uso dos testes adaptativos informatizados apresenta três limitações a considerar. A primeira delas diz respeito à impossibilidade dos alunos de reavaliar, e se for o caso, trocar suas respostas uma vez respondidas, visto que existe uma resistência por parte dos alunos a respeito dos TAI's, pois sempre fica a sensação de que estes não são tão válidos quanto os convencionais.

A segunda diz respeito ao tamanho do banco de itens e ao balanceamento do conteúdo do banco. Alguns trabalhos sobre testes adaptativos demonstram que a segurança e a confiabilidade dos TAI's dependem diretamente da existência de grandes bancos de itens. Segundo Olea et al., (1999), um banco de itens adequado deve ter, no mínimo, dez vezes mais itens do que um teste utiliza. Outro problema associado ao banco está relacionado à freqüência de exposição dos itens, pois pode ocorrer que um item com elevado poder discriminativo que seja de boa qualidade tende a ser aplicado mais vezes do que itens não tão discriminativos.

Em terceiro lugar, existe uma preocupação na preparação do conteúdo do banco de itens, principalmente relacionado à especificação de seus parâmetros. Isso porque dependendo do âmbito da aplicação do teste, pode haver dezenas de restrições de conteúdo, enquanto que não parece ser razoável sobrepor certos limites do número de variáveis de descrição. Por exemplo, podem existir algumas restrições entre itens que não podem aparecer juntos no mesmo teste, devido ao fato de que um deles pode conter a resposta do outro. Contudo, todos os efeitos desses problemas podem ser diminuídos com a realização de uma calibração estruturada e consciente dos itens de teste que compõem o banco. Entretanto, essa calibração exige um estudo empírico bastante caro, o que pode ser proibitivo para pequenas instituições. Tais limitações serão estudadas com mais detalhes no Capítulo 3.

2.4 Diretrizes e Princípios dos Testes Informatizados

A informatização dos testes passou nos últimos anos por um crescimento considerável. Agregados a esse desenvolvimento, cresceram também questões que abarcam aspectos importantes da utilização dos testes. Assim, faz-se necessário destacar as principais diretrizes específicas que tratam de regular toda a interação entre os usuários e os ambientes de testes informatizados.

A partir dos estudos de Olea et al., (1999), tem-se as seguintes diretrizes:

Aplicação do teste – a principal e essencial característica que todo teste informatizado deve seguir é a de garantir a padronização do procedimento de aplicação, ou seja, todos os participantes do teste devem ter igualdade de condições de maneira que não se interfira no processo de resposta. As condições internas do ambiente devem ser confortáveis e todos os equipamentos devem funcionar corretamente, visto que são tecnicamente mais complexos. É especialmente importante assegurar também que todos os indivíduos saibam perfeitamente manusear todos os equipamentos envolvidos e entendam a forma de responder às questões, tratando de evitar qualquer erro devido à carência de familiaridade com os computadores. Em suma, é preciso assegurar que o modo de aplicação informatizado não interfira significativamente na relação aluno-computador.

Interpretação dos relatórios – a maior limitação da interpretação automática de relatórios está no fato de que a mesma não incorpora informação individual, situacional ou contextual, o que pode ser útil para diagnosticar problemas ou melhorar a aplicação de um teste. É comum que nenhum sistema de interpretação automática seja capaz de levar em conta todas as características individuais e contextuais, e por isso mesmo é de suma importância que exista um profissional que seja capaz de identificar singularidades que influenciam no resultado final da interpretação dos relatórios. Isso porque podem existir diferentes particularidades pessoais, ou ainda determinadas situações de aplicação dos testes, que deturpam o resultado final e que sejam razões suficientes para invalidá-lo, caso não sejam consideradas. Em resumo, nada impede que ocorra a elaboração e interpretação automática de relatórios, desde que estas sejam

realizadas sob a supervisão de um profissional qualificado que decidirá sobre sua pertinência ou não.

Propriedades de pontuação – não se pode garantir que a partir da informatização de um teste, antes realizado no formato de lápis e papel, todas as medidas e propriedades se mantenham inalteradas. Os estudos de Olea et al., (1999) dizem que a equivalência de pontuações em função do modo de resposta (informatizado/tradicional) varia segundo o tipo de teste. Para que as respostas sejam similares e os resultados equivalentes, é preciso garantir, no mínimo, duas condições: em primeiro lugar, deve haver uma grande correlação entre as pontuações obtidas em ambos modelos; e em segundo, dada a correlação das respostas, deve-se investigar os desvios típicos na distribuição de pontuação relacionada com cada questão proposta.

Classificação – a pontuação obtida em um teste, seja ele tradicional ou informatizado, serve freqüentemente para classificar um indivíduo. Para que tais classificações sejam justas e significativas, deve-se eleger pontos de corte, que por sua vez dependem de vários fatores relacionados com os objetivos gerais da avaliação, tais como: a importância do teste, qualidade das questões, condições de aplicação, confiabilidade do equipamento utilizado, segurança no processamento das respostas e validade dos dados obtidos. Quanto mais relevantes sejam as conseqüências de uma classificação para os estudantes, maior cuidado deve-se ter no estabelecimento dos pontos de corte, evitando injustiças.

Revisão – é interessante que os professores dos testes permitam que especialistas, tanto do domínio do conhecimento abordado quanto técnicos em informática, tenham acesso ao modelo do teste informatizado, com o intuito de avaliar a qualidade do mesmo. É aconselhável também que os professores realizem testes de simulação de respostas, observando o comportamento do sistema e comprovando os resultados simulados com dados reais que, por sua vez, são executados pelos alunos (estes diferentemente dos especialistas, pois não terão acesso a dados técnicos), averiguando a validade da avaliação.

2.5 Vantagens dos Testes Informatizados

Nesta seção, analisam-se as principais vantagens que o uso dos testes informatizados oferece.

2.5.1 Vantagens Relacionadas com as Condições de Aplicação

Sem dúvida, os testes informatizados, apesar da dificuldade de se estabelecer e garantir os adequados critérios de aplicação, otimizam determinadas condições de aplicação dos testes. São elas:

1. *Requerem menos tempo* – os estudantes gastam menos tempo respondendo às questões com o uso de um teclado (ou *mouse*), do que na forma tradicional de lápis e papel. Isso faz com que o tempo de aplicação do teste informatizado seja menor do que seu equivalente teste tradicional;

2. *Redução na possibilidade de cópia* – a aplicação informatizada dos testes, com o emprego de determinados procedimentos de segurança, pode evitar a invasão de pessoas não autorizadas, minimizando a possibilidade de que os estudantes conheçam o conteúdo das questões e suas respostas antes do acontecimento do teste. Isso é particularmente importante quando há aplicações massivas de testes, em que muitos estudantes participam;

3. *Possibilidade de que as condições sejam iguais para todos* – com o emprego do computador, é possível dedicar condições e instruções similares aos estudantes, e controlar, quando for necessário, o tempo de resposta e exposição de cada item.

2.5.2 Vantagens Relacionadas com o Processamento de Respostas e suas Interpretações

O uso do computador trouxe uma grande vantagem no que diz respeito à correção dos testes. A correção automática das questões reduz de forma considerável o trabalho dos professores, pois o trabalho que eles deveriam fazer durante horas e às vezes até dias, pode ser feito de maneira eletrônica, diminuindo, ainda, a possibilidade de ocorrência de erros nesse processo.

Os sistemas informatizados de testes (estes principalmente relacionados aos testes adaptativos) também podem ser programados para que forneçam informação de maneira imediata (*FeedBack* imediato) a respeito do nível de habilidade individual em uma escala pré-determinada e que seja de fácil entendimento.

A respeito da elaboração automática de relatórios, todos os procedimentos que realizam essa tarefa se baseiam nas seguintes premissas: possuem um algoritmo que abriga um conjunto de regras programadas de geração de relatórios; e possuem uma base de dados (obtida com a aplicação e resultados dos testes) a partir da qual se constroem os mesmos.

2.6 Equivalência Entre os Testes Informatizados e de Lápis e Papel

A equivalência de resultados ou de pontuação entre a versão de um teste informatizado e o mesmo teste tradicional tem sido uma das questões mais discutidas no campo da informatização dos testes. A ausência de semelhança ou mesmo de igualdade de pontuação de ambas abordagens tem gerado grandes preocupações no que diz respeito à validade e utilidade do teste informatizado. Uma solução simples seria a comparação de pontuação de ambas abordagens, visto que sua validade ficaria condicionada à semelhança dos resultados, ou seja, a qualidade será garantida se ambas pontuações forem análogas, e as médias e a distribuição de pontuação de cada questão forem aproximadamente as mesmas.

Mas, conforme foi visto anteriormente neste capítulo, o simples fato de alterar o modelo de resposta do teste faz com que o resultado do teste também se altere. Assim, uma das primeiras revisões realizadas por Mazzeo e Harvey em 1988, citados por Olea et al., (1999) diz que a equivalência entre testes tradicionais e informatizados teve os seguintes resultados: a maioria dos resultados alcançou pontuações diferentes, mesmo que pequenas e de pouco significado prático;

entre as diferenças encontradas, a maioria teve médias mais altas nas abordagens tradicionais; nos testes nas quais a velocidade de resposta era relevante, as versões informatizadas se comportaram mais adequadamente.

Dessa forma, é importante observar o nível de informatização que um determinado teste pode alcançar e, assim, planejar, de forma estruturada e consciente, a melhor maneira de aplicação do teste, observando todas as condições ideais (homogeneidade do grau de informatização, procedimentos de acesso a informações úteis, treinamento dos alunos com os equipamentos, etc.) que a versão informatizada deve contemplar para diminuir os efeitos colaterais que venham afetar o resultado final.

2.7 Banco de Itens - BI

No início da informatização de testes, o uso do computador se deu fundamentalmente nas formas de apresentação e correção automática das respostas. Nesse princípio, as respostas dos estudantes eram introduzidas de forma manual, via teclado, ou por meio de cartões perfurados ou folhas de leitura óptica.

Nos anos setenta e oitenta, com o advento dos microcomputadores e computadores pessoais, desenvolveram-se numerosas pesquisas no campo da avaliação informatizada (Olea et al, 1999). Devido a grande capacidade de processamento desses computadores, foram possíveis a criação e implantação de diversas melhorias, tais como o uso de gráficos, geração de sons, alta resolução de vídeo e grande capacidade de armazenamento de informação. Essas características fizeram dos computadores ferramentas extremamente úteis, não só para o processamento dos resultados dos testes (pontuação, análise e elaboração de relatórios), mas para a construção, administração e armazenamento de itens.

Podemos definir um banco de itens como sendo um grande conjunto de itens (questões de teste), pertencentes ao mesmo domínio de conhecimento, que armazenados de tal maneira, facilitam a sua recuperação em um determinado momento.

Os primeiros bancos de itens consistiam em um conjunto de cartões, cada um deles contendo o enunciado da questão e suas possíveis respostas. No momento de construir um teste adaptativo, eram eleitos aqueles itens que melhor se adaptavam à necessidade de uso, sendo esses, os itens mais adequados para o propósito da avaliação em evidência.

O desenvolvimento da Teoria de Resposta de Itens (TRI) (Hambleton and Swaminathan, 1985) contribuiu para a criação e utilização de bancos de itens mais sofisticados, devido à inclusão de diversas informações (entre elas as variáveis psicométricas[2]) que descrevem mais precisamente as características de um determinado item. Com isso, juntamente com a utilização massiva dos computadores, se assiste a uma mudança na concepção acerca dos bancos de itens, principalmente na maneira de construção dos testes.

[2]Conjunto de parâmetros (a e/ou b e/ou c) da Teoria de Resposta de Itens que formaliza, mediante modelos matemáticos, a relação entre as respostas dos itens e o nível de habilidade de um aluno.

Esta mudança de concepção fez surgir definições variadas a respeito dos bancos de itens ao longo dos anos. De acordo com Olea et al. (1999), da análise das mesmas se depreende que um banco de itens possui duas características principais:

- Consiste da coleção relativamente grande e estruturada de itens que medem um domínio de conhecimento bem definido;
- Os itens se encontram armazenados junto com suas informações, tanto de conteúdo como psicométricas, estas últimas obtidas pelo processo de estimação dos parâmetros baseado nos modelos de resposta de itens da TRI.

Esta nova concepção dos bancos de itens é que o faz ser considerado como elemento central de um sistema de avaliação informatizada.

Quando se trabalha com banco de itens, é preciso se preocupar com aspectos fundamentais relacionados com sua gestão. Estes por sua vez, envolvem tarefas de construção e aplicação de testes a partir do banco, e principalmente a sua manutenção.

2.7.1 Balanceamento de Conteúdo e Testlets

Conforme dito anteriormente, um banco de itens consiste numa coleção de questões organizadas de tal maneira que facilita sua recuperação no momento da execução de um teste. Porém, esse aspecto do banco de itens exige que os itens armazenados se refiram a apenas um único conteúdo do domínio de conhecimento que o banco implementa. Entretanto, em uma dada área de conhecimento podem existir diversos conteúdos que necessitam ser avaliados, pois o professor pode querer avaliar diferentes habilidades que dependem de conteúdos variados.

Dessa forma, ainda sob o mesmo aspecto do banco de itens, para atender os diversos conteúdos de uma determinada área do conhecimento, seria necessário criar vários bancos de itens, um para cada conteúdo que se queira avaliar. Como exemplo podemos citar o Exame de Proficiência em Inglês (EPI)[3] do ICMC. Dentro da proficiência do inglês, alguns dos conhecimentos que o exame avalia são o da estrutura e compreensão de textos científicos escritos em inglês. Nesse contexto, considerando cada "conhecimento" subordinado a um conteúdo, seria necessária a criação de dois bancos de itens, um que medisse o conhecimento de estrutura de textos e outro para as habilidades de compreensão.

Entretanto, existe uma técnica aplicada na construção do banco de itens que é identificada como Balanceamento de Conteúdo[4] (Olea et al. 1999; KingsBury and Zara 1989; Wainer and Kiely, 1987). Essa técnica permite a divisão do banco de itens em várias secções, cada uma delas representando um conteúdo específico que se deseja avaliar. O uso dessa técnica evita a "quebra" do banco de itens em vários pequenos bancos e permite o desenvolvimento de um único banco volumoso contendo todos os conteúdos de domínio de conhecimento.

[3] http://www.nilc.icmc.sc.usp.br/staff/capteap/

[4] Content-Balanced em inglês.

A grande vantagem do uso do Balanceamento de Conteúdo é a possibilidade de um único exame avaliar várias habilidades dos estudantes sobre um mesmo banco de itens. Para isso, o professor especifica os conteúdos que o exame irá contemplar, garantindo que os mesmos serão cobertos pelo exame. Essa possibilidade é também muito útil para as áreas de conhecimento em que o conteúdo total é muito extenso e o professor deseja avaliar apenas uma única secção ou apenas algumas delas.

Outra técnica que pode ser implementada na construção do banco de itens é o uso de *Testlets*. Segundo Wainer and Kiely (1987) um *Testlet* é um grupo de itens relacionado a um determinado conteúdo que é desenvolvido como uma "unidade de teste" contendo um número de itens predeterminado, e que pode ser fornecido a um aluno durante uma avaliação. Os *Testlets* são usados nos testes adaptativos em que os procedimentos de seleção de itens individuais não são adequados, de maneira que cada *Testlet* pode ser selecionado separadamente e os alunos respondem todas as questões pertencentes a ele.

Um bom exemplo do uso de *Testlets* são as questões que fazem referência a um determinado texto, muito comum em exames em que a compreensão de textos é avaliada. Nesse exemplo, cada Testlet poderia ser formado pelo conjunto de questões que fazem alusão a um texto fornecido ao aluno.

A construção híbrida de um banco de itens que implementa o Balanceamento de Conteúdo e os *Testlets* pode ser muito proveitosa, já que é possível garantir que um determinado exame avalie as habilidades desejadas pelo professor, assegurando a vigência das particularidades de cada conteúdo coberto pelo exame com o uso de *Testlets*. Tal característica do banco de itens é que permite a implementação de testes adaptativos sensíveis ao conteúdo. Um bom exemplo do uso de um banco de itens com balanceamento de conteúdo é o sistema CBAT-2 (Huang, 1996) o qual implementa um banco de itens com hierarquia de conceitos de um determinado curso, associando diversos conceitos a diferentes módulos.

A Figura 2.3 mostra o esquema de um banco de itens implementando o balanceamento de conteúdo e os *Testlets*.

2.7.2 Construção do Banco de Itens

Segundo Olea et al. (1999), é preciso seguir os seguintes passos para a construção do banco de itens:

1. **Definição da estrutura do banco** – é a fase fundamental da criação do banco de itens, pois nela definem-se as tarefas a seguir durante todo o processo de criação do banco. Deve-se definir o nível de conhecimento que será medido, as pessoas a quem se destinam os testes, os tipos de itens que serão incluídos no banco e os diversos formatos dos itens de acordo com as diferentes áreas de conteúdo.

2. **Desenvolvimento dos itens** – compreende basicamente a eleição e criação dos itens que irão formar o banco. Podem seguir dois caminhos: aproveitar itens existentes ou construir novos itens, ambos adequados às especificações feitas na primeira fase. Seja qual for o caminho

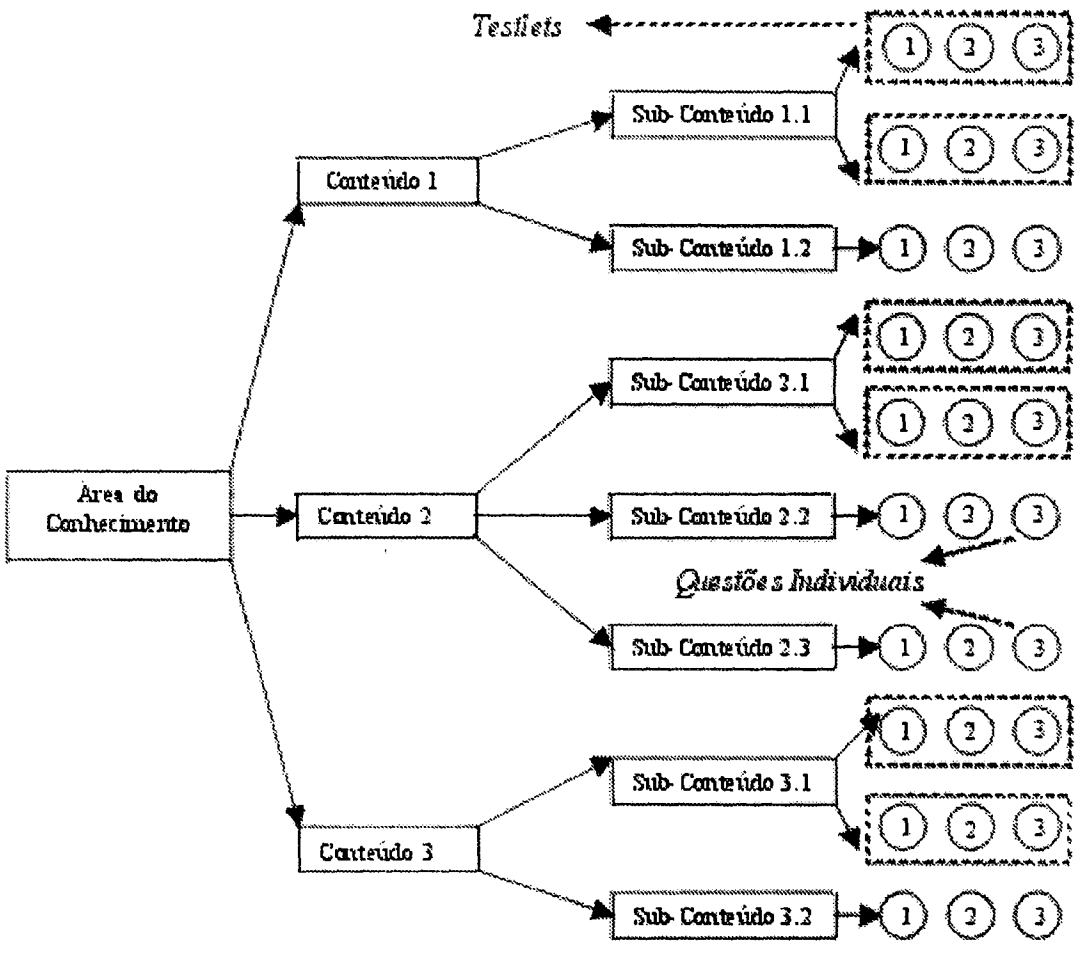

Figura 2.3: Diagrama que representa o Balanceamento de Conteúdo e Testlets em um banco de itens

escolhido para o desenvolvimento dos itens, é preciso realizar uma análise, tanto de conteúdo quanto da forma de apresentação dos itens, pois problemas relacionados a isso devem ser identificados antes da sua aplicação.

3. **Coleta de dados** – diz respeito ao processo de determinação dos dados que vão ser utilizados para as fases posteriores (principalmente a calibração de parâmetros) da construção do banco de itens. No caso dos testes adaptativos, os dados a serem coletados para a construção do banco são principalmente: as respostas fornecidas pelos alunos em testes pré-organizados, a omissão de resposta (caso ocorra), a quantidade e o tipo de questões fornecidas. Dentre os modelos de coleta de dados mais utilizados está aquele em que se aplicam vários conjuntos de itens a diferentes grupos de alunos, de maneira a analisar as respostas por eles fornecidas, e sobre esses dados realizar a estimação dos parâmetros de cada um dos itens administrados.

4. **Administração dos Itens** – diz respeito à administração do conjunto de itens definido na fase anterior em relação aos grupos de alunos pré-especificados, de maneira a coletar os dados necessários para a fase de calibração. A apresentação dos itens pode se dar tanto na forma eletrônica quanto na de lápis e papel, sendo que nesta última, a tarefa de coleta e reunião dos dados é mais trabalhosa.

5. **Análise dos itens** – está relacionada com o passo anterior, pois depois dos itens serem administrados, dispõe-se de dados para primeira análise dos mesmos, sendo essa tarefa responsável pela obtenção da informação acerca do comportamento dos itens e, posteriormente, pelo ajuste a um modelo de respostas, como por exemplo, os modelos da Teoria de Resposta de Itens (TRI).

6. **Calibração dos itens** – corresponde à especificação e inclusão de informações correspondentes às características dos itens que compõem o banco. Envolve as tarefas de estimação automática dos parâmetros dos itens de acordo com o modelo escolhido, o que finalmente irá caracterizar seu comportamento durante um teste.

7. **Armazenamento de informação** – corresponde ao armazenamento, de forma estruturada, de todos os dados disponíveis, de maneira que possam ser utilizados em função de necessidades específicas.

A Figura 2.4 mostra a representação gráfica dos passos necessários para a construção de um banco de itens.

2.7.3 Construção Automática de Testes

Efetivamente, os progressos da informática têm auxiliado na construção de testes e resultam de uma grande utilidade na montagem e elaboração dos itens, facilitando a especificação das variáveis e a confecção de testes com as propriedades necessárias ao alcance dos objetivos de uma determinada avaliação. Por exemplo, um teste de diagnóstico pode ser mais curto se implementado com testes adaptativos, já que a habilidade do aluno nessa abordagem é um bom critério de diagnóstico.

Adicionado à possibilidade de construção de testes, há um interesse no desenvolvimento de métodos que possam realizar a geração automática de itens de um teste, não se limitando apenas à mera montagem, ou ainda, em assuntos mais complexos como a correção automatizada de provas com respostas discursivas de caráter aberto e subjetivo.

2.7.4 Informações Incluídas

Dada a construção de um Banco de Itens, três tipos de informação podem ser armazenadas. São elas (Olea et al. 1999):

1. **Informações gerais** – correspondem às informações básicas a respeito de um determinado item. Dentre elas destacam-se: a palavra chave que descreve o conteúdo de um item, um

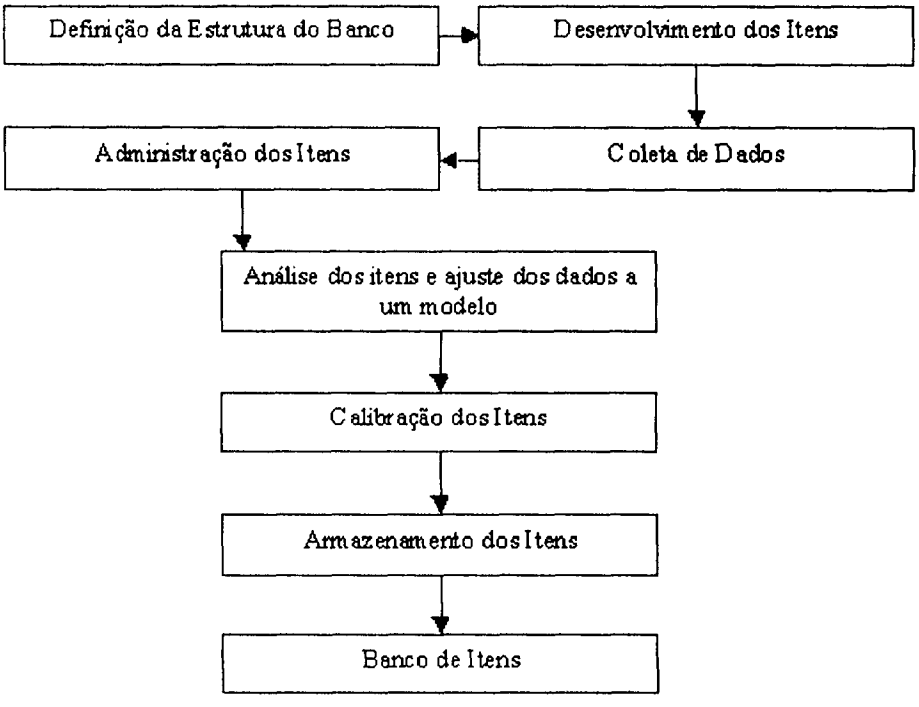

Figura 2.4: Diagrama de passos para a construção de um banco de itens (Olea et al. 1999)

número para sua identificação, o tipo de formato do item (resposta múltipla, completar, etc.), data de criação, autor, etc.

2. **Informações de conteúdo dos itens** – corresponde ao conteúdo do item propriamente dito, incluindo o enunciado, as opções de resposta, e a resposta correta.

3. **Informações de caráter psicométrico** – correspondem às informações formadas por exemplo, pelos parâmetros de Discriminação (a), Dificuldade (b) e Adivinhação (c) de cada item do banco. Tais parâmetros são indispensáveis quando se aplica o modelo da Teoria de Resposta de Itens para a seleção de itens nos testes adaptativos.

Além das informações acima citadas, ainda pode-se incluir, quando o banco de itens for usado para construção de testes, dados adicionais como a freqüência que um item vem sendo utilizado, a data em que foi usado pela última vez ou a identificação do teste no qual foi aplicado.

Dessa forma, toda e qualquer informação, que caracteriza seus itens, incluída no banco serve para realizar uma gestão mais eficaz do mesmo, fornecendo uma maior flexibilidade no processo de construção e aplicação dos testes.

2.7.5 Procedimento de Construção de Testes

Mesmo que na atualidade existam programas de autoria que permitem a criação de testes a partir de um Banco de Itens, como é o caso FastTest (www.assess.com) (analisado no Capítulo 5), a tarefa de composição de testes exige cuidados especiais e um acompanhamento profissional por parte dos professores.

Uma vez disponível um banco de itens, devidamente calibrado, o procedimento de construção de um teste segue os seguintes passos:

1. Definição do objetivo do teste: auto-avaliação, avaliação final, diagnóstica ou continuada;
2. Especificação das características do teste, definindo o tipo de formato dos itens, o conteúdo coberto dentro de uma área de conhecimento e as variáveis psicométricas (quando for o caso) que os descrevem;
3. Edição destas características em formato computacional e eleição do algoritmo de seleção de itens que melhor se ajusta às características definidas.

As estratégias de seleção de itens para a construção de um teste a partir de um Banco de Itens dependem diretamente do tipo de teste que se deseja criar e quais fins educacionais ele venha cobrir. Assim os dois tipos básicos são (Olea et al. 1999):

Testes Convencionais – os itens que farão parte do teste são fixados de antemão. Os testes gerados por esta estratégia são testes administrados igualmente a todos os sujeitos, sem levar em consideração qualquer característica individual;

Testes Adaptativos – esta estratégia se baseia na seleção de itens que compõem um teste que mais se adaptam ao nível de habilidade de um sujeito. Assim, cada indivíduo que participa da avaliação pode ter um teste diferente, dependendo da competência de cada um. Subjacente ao modelo adotado, o item selecionado será aquele que fornecer mais informação, dada a habilidade de um indivíduo. Tal procedimento será visto com detalhes no Capítulo 4;

2.7.6 Manutenção do Banco de Itens

Uma vez construído o banco de itens, temos ainda que tratar da manutenção das informações correspondentes aos itens. A correta manutenção do banco de itens ocorre basicamente com o emprego de duas atividades: a atualização e a renovação dos itens do banco.

2.7.6.1 Atualização do Banco de Itens

A atualização do banco de itens se dá a partir dos resultados da administração de testes realizados sobre o banco, mais especificamente sobre as repostas fornecidas pelos indivíduos que realizaram o teste, pois podem servir para atualizar a informação armazenada no banco a respeito das propriedades psicométricas dos itens (nível de dificuldade dos itens, por exemplo). Para isso é necessário

repetir todo o processo de construção do banco, desde a análise dos itens até o ajuste ao modelo teórico utilizado, descrito na subseção anterior. Isso porque em determinadas circunstâncias um certo item pode ser considerado adequado e em outros pode ser que não seja. A Figura 2.5 mostra os passos para a atualização de um banco de itens.

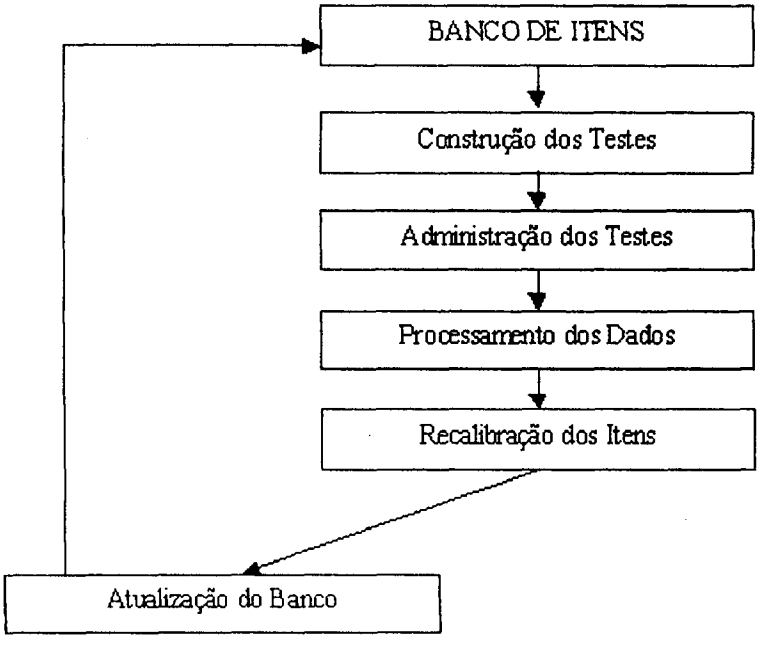

Figura 2.5: Diagrama de passos para atualização de um banco de itens (Olea et al. 1999)

2.7.6.2 Renovação do Banco de Itens

O processo de renovação de um Banco de Itens ocorre tanto com a eliminação quanto com a inclusão de novos itens. Quando um banco de itens é usado muito freqüentemente, pode ser necessária a eliminação definitiva ou temporária de alguns deles, pois há à possibilidade de seus itens serem usados constantemente em diferentes testes. Outro fator que leva a exclusão de itens é a constatação que seu comportamento não está mais adequado, ou seja, não atende às especificações e objetivos da avaliação.

Dada à criação de novos itens (questões), a tarefa de inclusão no banco pode ser realizada de duas maneiras: a primeira delas é a administração dos novos itens a um grupo de alunos, levando em consideração todas as tarefas envolvidas na construção de um banco, desde seu desenvolvimento até a análise qualitativa e o armazenamento propriamente dito; a segunda envolve a criação de um teste contendo itens novos e itens antigos (já cadastrados no banco) que é aplicado a um grupo de alunos, e a partir dos dados de resposta, os itens são calibrados (itens novos) ou recalibrados (itens antigos) e armazenados no banco.

A adição de itens antigos a um teste e sua posterior recalibração são muito importantes para a vida e para manutenção do banco de itens, pois é possível comparar os valores antigos (calibração anterior) com os valores da nova calibração, melhorando cada vez mais a métrica do banco. A Figura 2.6 mostra os possíveis passos para a renovação de um banco de itens.

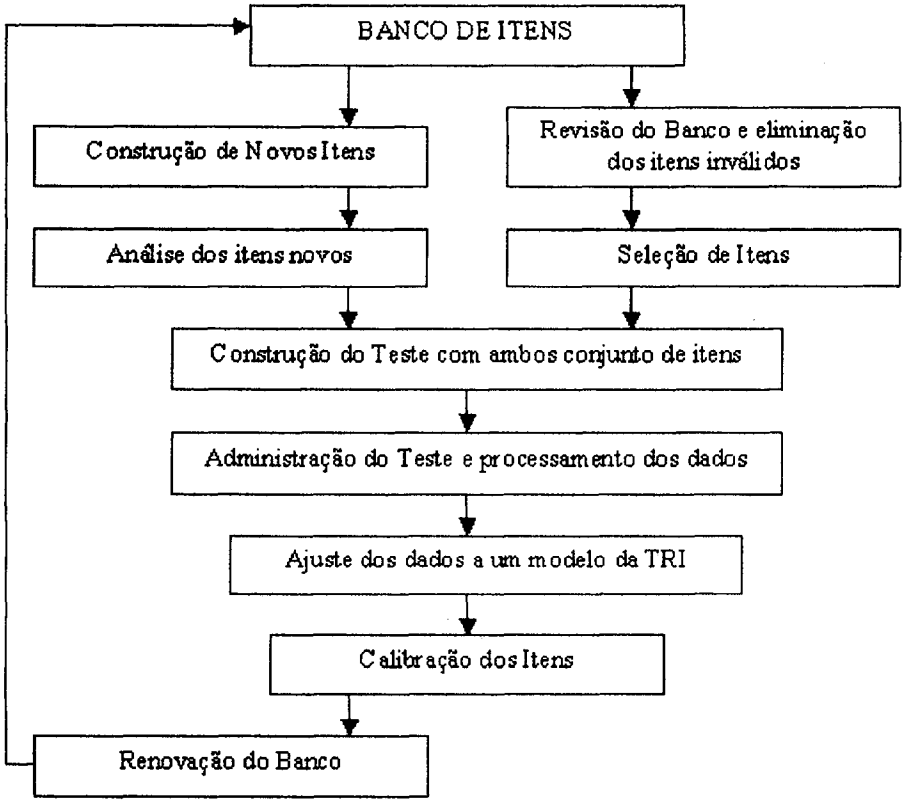

Figura 2.6: Diagrama dos passos para a renovação de um banco de itens (Olea et al. 1999)

A informatização dos testes e sua aplicação mediante o uso do computador oferecem maiores possibilidades frente à administração tradicional, visto que é possível aproveitar as vantagens tecnológicas que estes oferecem. O emprego da informática pode se dar desde a aplicação convencional, que é a forma tradicional de lápis e papel usando o computador, até a gestão dos testes adaptativos informatizados, que permitem uma maior flexibilidade no processo de avaliação, dado que se podem utilizar de maneira mais eficiente as propostas dos bancos de itens, adaptando-se ao nível de habilidade de cada indivíduo.

CAPÍTULO 3

Testes Adaptativos Informatizados (TAI's): Evolução e Propriedades

O Teste Adaptativo é um método de teste no qual diferentes conjuntos de questões (itens) são administrados para diferentes indivíduos, dependendo do nível de habilidade de cada um. Os testes adaptativos contrastam com os testes convencionais informatizados na seqüência de apresentação e número dos itens.

Em um teste adaptativo, vários itens são apresentados ao estudante, que os respondem e podem ser pontuados como corretos ou incorretos. Baseado nestas respostas, itens adicionais, cujo grau de dificuldade é conhecido, são selecionados a partir de um banco de itens e apresentados ao estudante. Por exemplo, se uma resposta a uma questão for incorreta, a próxima questão selecionada será mais fácil; se a resposta for correta, a próxima questão será mais difícil, assim novos itens são administrados e respondidos, sendo os itens subsequentes selecionados baseados na resposta anterior.

Alguns testes adaptativos têm sido desenvolvidos e administrados via lápis e papel. Outras abordagens exigem que os itens sejam administrados por meio do computador. Esta última abordagem traz grandes vantagens e é referida como Testes Adaptativos Informatizados – TAI's – (*Computer Adaptive Testing* – CAT em inglês).

A maior vantagem dos testes adaptativos informatizados em relação aos convencionais é a eficiência. É possível, com o uso dos testes adaptativos, chegar a um resultado equivalente ou superior ao dos testes tradicionais com a administração de menos itens em menos tempo. Junto com a economia de tempo, os testes adaptativos fornecem uma melhor identificação e medida dos traços individuais envolvidos num processo de avaliação.

Este capítulo pretende dar uma visão geral dos testes adaptativos, destacando os principais conceitos, identificando suas diretrizes e propriedades. A lógica de trabalho, as limitações e potencialidades, bem como as questões técnicas dos testes adaptativos também são ressaltadas.

3.1 Evolução dos Testes Adaptativos Informatizados

Conforme exposto anteriormente, existem diferentes abordagens dos testes adaptativos, baseadas principalmente no método de seleção de itens a partir de um banco. Como complemento, servindo como característica essencial do funcionamento dos testes, deve-se considerar a habilidade inicial de cada participante, para que se possa definir o ponto de início do teste. Outra preocupação inerente aos TAI's está centrada no(s) critério(s) de parada, visto que o desenvolvimento de cada indivíduo durante o teste será diferente.

3.1.1 O Primeiro Teste Adaptativo - O "Binet Test"

Os princípios gerais dos testes adaptativos foram implementados pela primeira vez na França, no início do século passado, por Alfred Binet (por isso o nome de Binet Test) em seu laboratório, e resultaram no que foi chamado Teste de Inteligência de Binet (Weiss, 1985). A estratégia usada por Binet foi baseada na idéia de ramificação fixa, usando uma regra mecânica de ramificação na qual os pontos de início e critérios de parada dos testes eram variáveis.

O Teste de Inteligência de Binet foi criado para o diagnóstico do nível de inteligência de uma criança em comparação com a idade cronológica. Assim, o administrador do teste de Binet, baseado em qualquer informação disponível (idade da criança, por exemplo), estimava o nível inicial para o início do teste. Por exemplo, se uma criança de oito anos de idade está sendo testada, mas o administrador sabe que a criança é esperta, o teste poderá iniciar com itens destinados a uma criança de nove anos. Similarmente, se o administrador sabe que a criança possui certas dificuldades, o teste poderá iniciar com um nível mais baixo, para uma criança de sete anos.

Uma vez identificado o nível da criança, os itens correspondentes a tal nível são fornecidos e pontuados seqüencialmente. Depois de todos os itens terem sido administrados, o próximo nível, mais alto ou mais baixo, é escolhido. Se todos os itens propostos do primeiro nível forem respondidos corretamente, o ***Nível Inferior*** é identificado e é assumido que todos itens abaixo desse nível seriam respondidos corretamente caso fossem disponibilizados. Nesse caso são fornecidos os itens de maior dificuldade (nível mais alto), os quais novamente são administrados e pontuados.

Os itens continuam a ser disponibilizados de acordo com o aumento do nível de dificuldade, até que o ***Nível Superior*** seja alcançado, ou seja, até que todos os itens em um determinado nível sejam respondidos incorretamente. Quando esse nível é identificado, é assumido que todos itens acima deste seriam respondidos incorretamente caso fossem aplicados, e assim o teste é finalizado. Se de outra forma, a criança não responde corretamente todos os itens do nível inicial, itens de grau de menor dificuldade são fornecidos, até que seja encontrado o **Nível Inferior**. Após a identificação desse nível, itens mais difíceis são aplicados até o **Nível Superior** ser encontrado.

Assim, a regra de ramificação do Teste de Binet é simples: se todos os itens de um determinado nível forem respondidos corretamente, são fornecidos itens de um nível mais alto até que todos eles sejam respondidos incorretamente (Nível Superior). Caso contrário, se todos os itens de um certo

nível forem respondidos incorretamente, são disponibilizados itens de um nível mais baixo até que todos eles sejam respondidos corretamente (Nível Inferior).

O critério de parada do Teste de Binet ocorre quando o Nível Inferior e Nível Superior para um dado indivíduo são identificados. Entre esses níveis, são respondidas várias questões, correta e incorretamente, formando assim um conjunto de itens aproximadamente adaptado ao nível de habilidade individual.

Na Figura 3.1 é ilustrado o procedimento de seleção de itens do Teste de Binet.

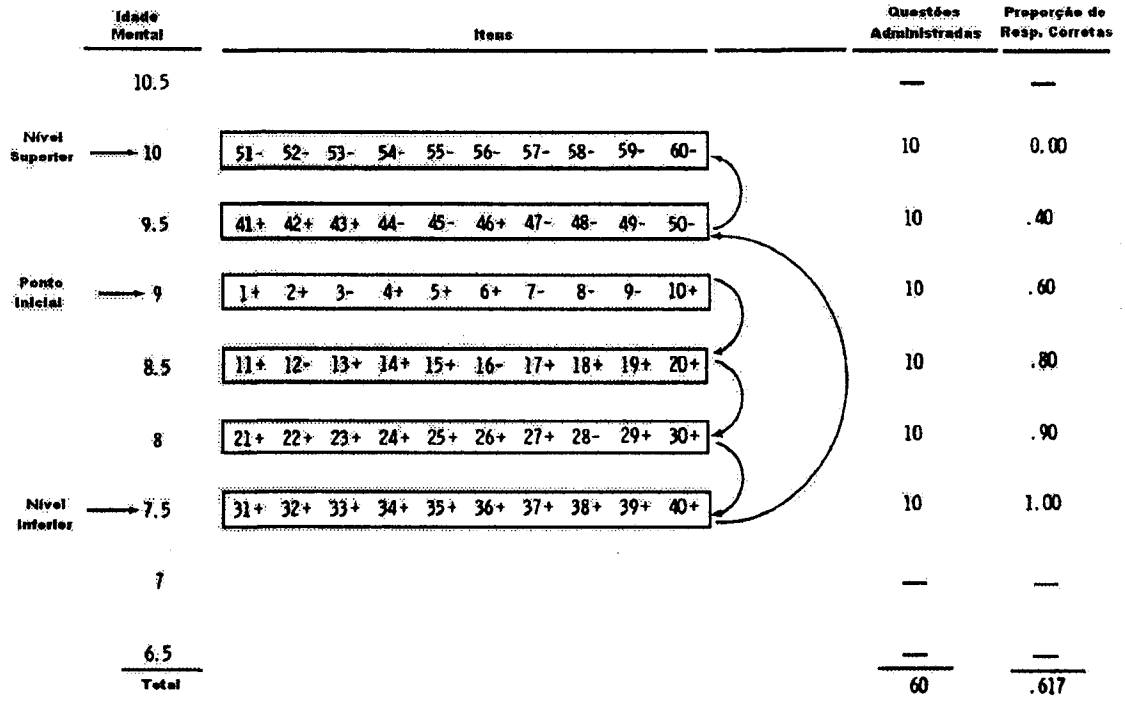

Figura 3.1: Registro de resposta de um aluno em um Teste de Binet (+ = correto, - = incorreto(Weiss, 1985)

3.1.2 Outras Abordagens de Testes Adaptativos

Seguindo a introdução dos testes adaptativos realizada pelo Teste de Binet, outros exemplos surgiram apenas no final dos anos 50. Tais testes foram experimentais e usados apenas para estudar as diversas características dos testes adaptativos (Weiss, 1985).

Testes Adaptativos de dois estágios (Two-Stage Adaptive Testing)

Os testes adaptativos de dois estágios foram construídos para a administração na forma de lápis e papel, utilizando regras de ramificação bastante simples e alcançando pouca capacidade adaptativa. Sua estratégia consiste no fornecimento de dois testes ao aluno, um em cada estágio. Todos os

estudantes recebem um primeiro teste, chamado de *Routing Test*, apresentando itens de dificuldade média. Baseado nas respostas do *Routing Test*, os alunos recebem um segundo teste, chamado de *Measurement Test*, o qual é aproximadamente mais adaptado à habilidade dos mesmos. Indivíduos que respondem corretamente todos ou a maioria dos itens do *Routing Test* recebem o *Measurement Test*, com itens de maior dificuldade. Por sua vez, se o indivíduo responde cerca da metade dos itens corretamente, serão fornecidos no segundo estágio *(Measurement Test)* itens de dificuldade média; e caso responda poucos itens corretos, receberá itens de menor dificuldade.

Testes Adaptativos Piramidais (Pyramidal Adaptive Test)

Os testes adaptativos piramidais consistem de um conjunto de itens organizado pelo grau de dificuldade, que estruturalmente lembra a forma de uma pirâmide. O topo da pirâmide possui um único item de dificuldade média, que é fornecido a todos os indivíduos. No próximo nível da pirâmide, existem dois itens, um levemente mais difícil que o primeiro e outro levemente mais fácil. Assim, em cada nível da pirâmide, são acrescentados mais dois novos itens, um com grau maior de dificuldade e outro com grau menor. Dessa forma, são administrados os itens levemente mais difíceis de cada nível da pirâmide, caso o estudante responda corretamente o item do nível superior; ou, em caso de erro na resposta, o item levemente mais fácil é fornecido. Portanto, à medida que um aluno segue respondendo todas ou a maioria dos itens corretamente, os itens fornecidos aumentam em grau de dificuldade, enquanto que um aluno que responde incorretamente, receberá itens de dificuldade decrescente (Hambleton and Swaminathan, 1985; Weiss, 1985).

Testes Adaptativos Estratificados (Stratified Adaptive Test)

O teste adaptativo estratificado é uma variante melhorada do Teste de Binet, e também incorpora pontos de início variáveis (dependendo da habilidade inicial do indivíduo) e diferentes pontos de parada do teste (baseado no desempenho de cada participante). A principal diferença dessa estratégia em relação ao Teste de Binet é que ela seleciona um item após o outro, ou seja, a eleição do próximo item a ser administrado é realizada logo depois de cada item ter sido respondido.

Quando um estudante responde corretamente um determinado item, a ele é fornecido um item ainda não administrado, situado em um nível maior de dificuldade. Em contrapartida, uma resposta incorreta conduz o estudante a receber um item pertencente a um nível menor de dificuldade. Similarmente ao Teste de Binet, o processo de teste continua fornecendo item após item, alguns mais difíceis, outros mais fáceis, até que seja identificado o **Nível Superior**, que é um nível de dificuldade no qual nenhum item foi respondido corretamente. Embora os testes adaptativos estratificados não requeiram a identificação do **Nível Inferior**, um dos critérios de parada pode ser a sua identificação, que, análogo ao Teste de Binet, é um nível de dificuldade no qual todos os itens foram respondidos corretamente (Weiss, 1985).

Na Figura 3.2, é exibido um exemplo (o mesmo do Teste de Binet, Figura 3.1) da estratégia de seleção de itens do teste adaptativo estratificado. Nesse exemplo, viu-se um registro de resposta de um aluno, em que cada item foi administrado e respondido. Baseado na resposta do aluno, o próximo item será escolhido, podendo haver aumento ou diminuição do grau de dificuldade.

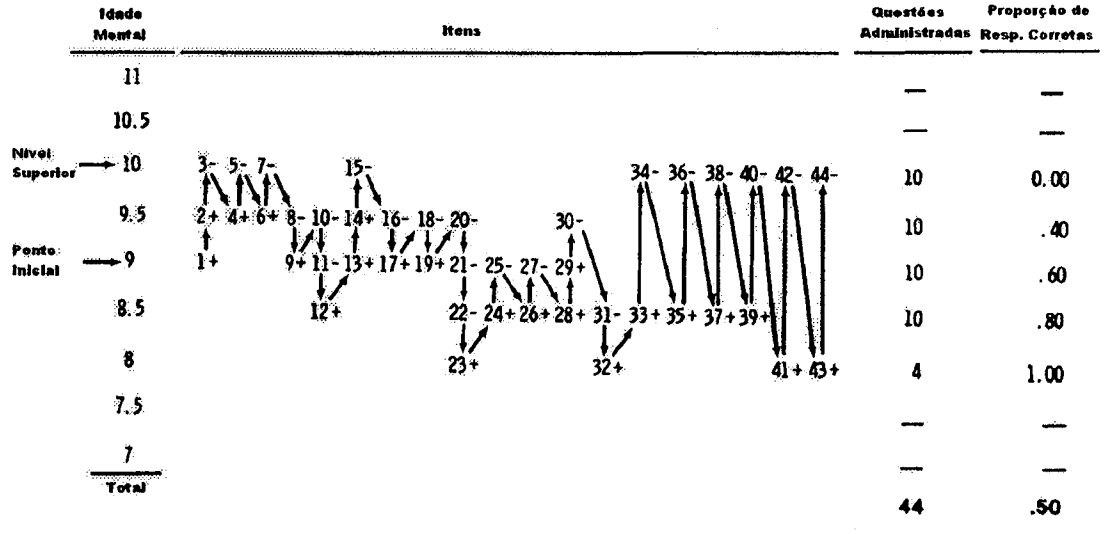

Figura 3.2: Registro de resposta de um aluno em um Teste Adaptativo Estratificado (+ = correto, - = incorreto)(Weiss, 1985)

Conforme mostrado, o primeiro item administrado foi o primeiro item disponível no nível de dificuldade[1] 9 (equivalente a idade mental de 9 anos). Esse item foi respondido corretamente e, seguindo a regra de ramificação, o próximo item a ser escolhido será o primeiro item do próximo nível de maior dificuldade, nesse caso 9,5. Esse item (2) também foi respondido corretamente e por isso é oferecido ao estudante o primeiro item do nível de dificuldade 10. Devido ao fato do item (3) ter sido respondido incorretamente, o segundo item do nível de dificuldade 9.5 é administrado. Os itens de 5 a 7 são alternativamente respondidos como incorreta e corretamente, até o item 8 (quarto item do nível de dificuldade 9.5) ter sido respondido incorretamente. Como conseqüência, o segundo item do nível de dificuldade 9 foi fornecido. Com poucas exceções, os itens de 10 a 20 também foram alternativamente sendo respondidos correta e incorretamente, concentrando o processo de teste entre os níveis de dificuldade 9 e 9.5. Três erros consecutivos dos itens 20, 21 e 22 levaram a administrar o primeiro item (23) do nível de dificuldade 8. Seguindo o processo do teste, usando o esquema de seleção de item após item, o item 30 (pertencente ao nível de dificuldade 9.5) foi respondido incorretamente, o que levaria à seleção do próximo item do nível de dificuldade 9. Mas, visto que todos os itens desse nível já foram administrados, o seguinte item será o pertencente ao nível de dificuldade 8.5. Analogamente, devido ao fato de que os itens do nível 9 e 9.5 foram

[1]Nesta subseção, Nível de Dificuldade será tratado como Idade Mental, tendo o mesmo sentido aplicado contexto da Figura 3.1

todos administrados, a seleção de itens após o item 33 adiantou vários níveis até terminar. Em nosso exemplo, o teste terminou quando o ***Nível Superior*** foi identificado, (nível de dificuldade 10), pois todos os itens foram respondidos incorretamente.

Em comparação com o Teste de Binet, o Teste Adaptativo Estratificado apresentou redução no tamanho do teste de 27%, pois foram necessários 44 itens para o alcance do nível superior, visto que, na primeira estratégia, foi necessário a administração de 60 itens para o mesmo objetivo. Os dois testes resultaram, porém, na proporção de acertos diferentes, sendo 0.50 (50%) nos testes adaptativos estratificados e 0.617 (61.7%) no Teste de Binet.

Testes Adaptativos baseados na Teoria de Resposta de Itens (TRI)

Recentemente, têm surgido novas abordagens de testes adaptativos, principalmente aquelas baseadas no uso da Teoria de Resposta de Itens. Tal teoria permite em conjunto com o banco de itens, o desenvolvimento e calibração de itens aos quais podem ser adicionados os parâmetros de *discriminação* (a), *dificuldade* (b) e *adivinhação* (c). A adição desses novos parâmetros permite obter maior flexibilidade a respeito da identificação mais precisa da habilidade, ou seja, dos diversos níveis de competência dos participantes de um teste.

Dessa forma, pode-se dizer que o desenvolvimento da TRI foi um marco importante na aplicação dos testes adaptativos, já que tornou possível a administração de diferentes itens com diversas características a distintos indivíduos, respeitando e considerando os traços intelectuais de cada um. O progresso da TRI também propiciou o uso mais intenso do computador nos testes adaptativos, já que a complexidade de seus modelos matemáticos de estimação de parâmetros e a tarefa de seleção de itens (do banco) seriam impossíveis sem o uso dessa ferramenta.

No próximo capítulo, serão vistas com mais detalhes as principais abordagens dos testes adaptativos sob à luz da TRI.

3.2 Propriedades dos Testes Adaptativos

Nesta subseção, serão expostas as principais propriedades relacionadas com os testes adaptativos. Essas propriedades contemplam os aspectos fundamentais implicados na construção e aplicação dos TAI's, que são as seguintes: Dimensionalidade, Confiabilidade, Validade, Estimação de parâmetros, Calibração dos Itens e os Fatores humanos envolvidos. Tais propriedades são bastante importantes para todos os testes e dependem diretamente dos procedimentos de administração: a estimação de parâmetros, seleção de itens, pontuação dos testes e os critérios de parada (Green et al., 1984; Olea et. al., 1999).

3.2.1 Dimensionalidade

O conceito de dimensionalidade está intimamente ligado ao número de habilidades[2] ou competências individuais que podem ser avaliadas em um teste, dadas todas as características pessoais. O termo ***habilidades*** pode ser entendido como um conjunto de capacidades que podem ser "medidas" por meio de uma avaliação, por exemplo, o conhecimento em determinado domínio, velocidade de raciocínio, percepção de detalhes, aptidão para relacionar conceitos de uma área de conhecimento e etc. É comumente assumido que apenas uma habilidade seja necessária para demonstrar a performance (boa ou ruim) de um estudante em um teste.

Os modelos de resposta de um teste que assumem e medem apenas uma habilidade são referidos como ***unidimensionais***. A maior exigência requerida nesses modelos é a especificação mais precisa possível dos parâmetros do teste, pois a partir deles é que se pode dimensionar e identificar o que se vai medir, ou qual componente do teste vai influenciar na performance do indivíduo (Hambleton and Swaminathan, 1985). É preciso levar em consideração, também nesse modelo, que a existência dos fatores cognitivos e de personalidade, como o grau de motivação e ansiedade, influenciam no processo de avaliação. Modelos de resposta que assumem a medida de mais de uma habilidade a respeito dos estudantes são chamados de modelos ***multidimensionais***.

Os atuais modelos da TRI utilizados para o testes adaptativos assumem que um teste é unidimensional, pois os testes realizados sob esse modelo, com fins aplicados a uma determinada área do conhecimento, devem utilizar um banco de itens calibrado, com dados projetados para a medida da habilidade desejada. Se o teste a ser administrado possuir característica altamente multidimensional, uma alternativa de uso da TRI seria a divisão do banco de itens em diversos conteúdos, cada qual medindo uma habilidade desejada, conforme dito na seção 2.7.

3.2.2 Confiabilidade

Assim como qualquer outro método de avaliação, os testes adaptativos devem ser precisos e medir as habilidades dos estudantes com poucas variações de erro, ou seja, têm que ser confiáveis. Uma das grandes vantagens dos TAI's, verificado seu correto funcionamento, é que seu uso supõe uma melhora na precisão das medições realizadas, já que ajustam as características das questões do teste (itens) ao nível de cada pessoa. A forma de expressar a confiabilidade no contexto dos testes adaptativos também se baseia na TRI, pois esta permite enunciar a medida de erro em função do nível da habilidade da pessoa avaliada.

Dessa forma, a medida de erro depende diretamente do critério de parada adotado. Se a própria medida de erro for o critério de parada, todos os indivíduos são avaliados sob o mesmo valor (tão ampla quanto a escala de habilidade θ[3]), sendo que o erro padrão de medida é constante. Se o banco de itens não é grande o bastante para suportar uma variância de erro uniforme ou se outro

[2]Nesse contexto, também pode ser referenciado como Traço Intelectual ou Nível Intelectual de um estudante.
[3]TETHA (θ) variável que representa a habilidade de um indivíduo.

critério de parada for adotado (número de itens administrados por exemplo), a medida de erro irá depender do nível de habilidade θ (Green et al., 1984).

3.2.3 Validade

Quando se utilizam os testes adaptativos, assim como se utiliza qualquer outro método de teste, em todos os contextos, a validade é o objetivo fundamental, ou seja, além de medir com confiabilidade, é preciso alcançar aquilo que foi proposto para o teste, e não outra coisa.

Quando os testes adaptativos estão sendo implantados como uma nova alternativa sobre uma antiga metodologia de teste (no lugar lápis e papel, por exemplo), a validade do TAI deve ser avaliada. Qualquer mudança nas variáveis que indica a qualidade do teste, ou que aponta qualquer alteração na natureza do mesmo, é necessário realizar uma reavaliação. Isso também é útil para avaliar as relações entre testes supostamente equivalentes.

3.2.4 Estimação de Parâmetros e Calibração dos Itens

A calibração dos itens consiste na especificação dos parâmetros (**a**, e/ou **b** e/ou **c**) dos itens que compõem um banco, e que futuramente serão selecionados para uso em um determinado teste. A quantidade de parâmetros a especificar depende diretamente do modelo subjacente à TRI que será adotado, e esse pode ser de um, dois ou três parâmetros. Quando o modelo contempla apenas um parâmetro[4], este será o argumento de **Dificuldade (b)**; no modelo de dois parâmetros, adiciona-se o argumento de **Discriminalidade (a)**, no modelo de três parâmetros, além dos dois anteriormente citados, é adicionado o parâmetro de **Adivinhação (c)**.

Ambas as atividades de estimação e calibração dos itens constituem uma fase essencial para que um teste adaptativo funcione corretamente. Se os itens estiverem mal calibrados e os parâmetros forem mal estimados, todo o processo de seleção de itens, inerente ao processo de "adaptação", torna-se viciado, pois causa a seleção freqüente de alguns itens e a não seleção de outros.

A estimação dos parâmetros pode acontecer de várias formas. Uma delas é utilizar os modelos da TRI, que são a base da calibração dos itens, e exigir uma amostra ampla e representativa de itens para que as funções matemáticas sejam consistentes (Olea et. at. 1999). Outra forma consiste na divisão dos itens em subtestes e a administração dos mesmos a uma amostra de estudantes, e, baseado na análise e observação do resultado destes, especificar os parâmetros. Uma maneira alternativa ainda possível é contar com o auxílio de especialistas e professores do domínio do conhecimento implementado, promovendo uma estimação dos parâmetros apoiado no conhecimento empírico destes indivíduos. Essa tarefa de estimação pode ser alterada após vários testes, com a medição do número de vezes em que os estudantes erraram determinados itens (Huang, 1996).

[4]Modelo chamado de Rasch Model que usa apenas o parâmetro de Dificuldade (b) para estimação de habilidade e seleção dos itens.

3.2.5 Fatores Humanos

Visto que a forma (ordem e número de itens) de apresentação dos testes adaptativos diferem acentuadamente dos testes tradicionais, é preciso ter um cuidado especial com os equipamentos envolvidos (mouse, teclado, monitor etc.), com o ambiente (ventilação, temperatura, iluminação etc.) onde o teste será aplicado e com a explicação clara dos procedimentos (como responder?) de resposta, para que tais fatores não influam no resultado do teste. É necessário que o ambiente seja bem iluminado, quieto, confortável e livre de distrações, de maneira que qualquer variável externa não atrapalhe o desempenho do indivíduo.

3.3 A Lógica dos Testes Adaptativos Informatizados

O progresso dos testes adaptativos, juntamente com o aperfeiçoamento da Teoria de Resposta de Itens (TRI), têm promovido o desenvolvimento de modelos matemáticos que têm levado a resultados muito importantes no que diz respeito à descrição das características pessoais no contexto de uma avaliação. Isso porque tais modelos permitem que a pontuação dos testes de todos os indivíduos seja expressada em uma escala comum, sem levar em consideração que cada estudante pode ter respondido diferentes números de itens, permitindo que os classifique a respeito da habilidade que está sendo medida. Este ponto será detalhado na subseção 4.6 do Capítulo 4.

Conforme dito na subseção anterior, um item de teste (questão) pode ser descrito por um, dois e até três parâmetros. A capacidade que um item possui de discriminar entre vários indivíduos de diferentes níveis de habilidade (também pode ser entendida como característica pessoal) é chamada de "poder de discriminação" de um item e é baseada no parâmetro ***Discriminação (a)***. O grau de dificuldade de um item é dado pelo parâmetro ***Dificuldade (b)***. Os modelos da TRI incluem ainda um parâmetro que representa a probabilidade de se ter uma resposta correta dada a baixa habilidade de um indivíduo, e é referido como ***Adivinhação (c)***. Assim, os testes adaptativos que seguem os modelos da TRI são fundamentados na descrição dos itens, dado os parâmetros de dificuldade, poder de discriminação e suscetibilidade de adivinhação (Kreitzberg et at., 1978).

Portanto, um teste adaptativo está pronto para iniciar quando existe um banco de itens, com os parâmetros da TRI especificados; quando um procedimento para estimar a habilidade dos participantes tenha sido escolhido (Procedimento da Máxima Informação ou Modelo Bayesiano, por exemplo), quando existe um programa que gerencia a administração de itens aos candidatos e quando os critérios de parada já tenham sido decididos. Dessa maneira, a execução de um TAI é um processo interativo que segue os seguintes passos:

1. Todos os itens do banco que ainda não foram administrados são identificados e avaliados para determinar qual será o "melhor" item a ser apresentado, dado o corrente nível de habilidade estimado;

2. O item selecionado na tarefa acima é administrado e o estudante o responde;

3. A nova estimativa de habilidade é calculada baseada na resposta do estudante;

4. Os passos de 1 a 3 são repetidos até que o critério de parada seja encontrado.

Diversos métodos podem ser usados para calcular as estatísticas necessárias para cada um desses quatro passos, principalmente os métodos adequados à TRI (Hambleton and Swaminathan, 1985; Baker, 1992). Considerando que os parâmetros dos itens estão corretamente especificados, a estimação de habilidade é o valor de Tetha (θ) que melhor se ajusta ao modelo de aluno. O critério de parada poderá ser o número de itens administrados, a alteração dinâmica do nível de habilidade, a cobertura de conteúdo (se o teste cobriu todo o conteúdo ministrado), um indicador de precisão como o erro padrão ou uma combinação destes (Rudner, 1998).

3.3.1 Potencialidades e Limitações dos TAI's

De acordo com Olea et al., (1999); Hambleton and Swaminathan, (1984); Kreitzberg et al., (1978); Weiss, (1985), os testes adaptativos fornecem as seguintes potencialidades:

- As questões do teste são disponibilizadas de acordo com a demanda (se ainda for preciso que novos itens sejam administrados e o critério de parada não for encontrado a disponibilização acontecerá) e as pontuações em cada uma delas são calculadas imediatamente;
- Os itens são individualmente passados ao estudantes, de acordo com sua competência;
- A segurança do teste é melhorada porque não existem cadernos de prova, nem testes impressos em folhas de resposta, evitando o extravio. Também, medidas de segurança realizadas sobre o banco de itens evitam o acesso não autorizado;
- Os novos conteúdos e diferentes formatos (com uso de recursos multimídia, por exemplo) de itens podem ser adicionados ao teste;
- O tempo necessário para a administração do teste é menor. Poucos itens são exigidos para alcançar uma precisão aceitável pré-definida. Os testes adaptativos reduzem o tempo de teste, mantendo o mesmo nível de confiança promovendo também redução da fadiga dos estudantes, que é um fator que pode afetar significativamente os resultados do teste;
- Os padrões de respostas podem ser identificados, úteis principalmente nos testes de diagnóstico, permitindo o ajuste aos moldes dos participantes.

Embora os TAI's forneçam várias potencialidades, como as descritas acima, os testes adaptativos possuem diversas limitações, principalmente aquelas relacionadas com questões técnicas e procedimentais. São elas:

- O *Hardware* (memória e CPU) pode restringir a apresentação de alguns tipos de itens. Itens envolvendo muitos detalhes gráficos ou tamanhos longos (de leitura demorada) podem ser difíceis de apresentar;

- Os bancos de itens requerem cuidado na calibração. Os parâmetros dos itens do banco devem ser cuidadosamente dimensionados, sob pena de mau funcionamento dos testes;
- Os estudantes recebem um conjunto diferente de questões (de maneira individual), o que pode causar um sentimento de injustiça e desigualdade.

3.3.2 Questões Técnicas

No âmbito dos testes adaptativos, existe na literatura uma considerável quantidade de questões técnicas e procedimentais que regem o seu uso e que são responsáveis por uma melhor aplicação e implementação. Nesta subseção, veremos algumas a seguir.

3.3.2.1 Estudos de Simulação

É possível utilizar estudos de simulação para averiguar as propriedades das diferentes abordagens dos testes adaptativos, além de comparar os resultados destes com a aplicação dos testes convencionais, avaliando os efeitos das diferentes variáveis (parâmetros dos itens e nível de habilidade) que influenciam nos testes adaptativos.

Tais estudos são realizados por meio da implementação de modelos matemáticos de simulação e consistem na atribuição fictícia de respostas que pessoas reais dariam aos itens de um teste. Devido ao fato de tais respostas serem geradas por um programa de computador, é possível simular situações reais de um teste, conferindo uma distribuição diversificada de características pessoais, que servem para avaliar o comportamento do modelo implementado. A aplicação dos estudos de simulação pode ter duas vantagens: primeiro, permite a rápida avaliação dos efeitos de vários testes e as características dos estudantes; e segundo, é possível conhecer a conduta de uma avaliação, dispostos os vários tipos de dados simulados (Olea et al. 1999; Weiss, 1985).

Para simular as respostas de um indivíduo, a primeira tarefa consiste em determinar seu nível hipotético de habilidade. Esse valor pode ser tomado aleatoriamente, dado um limite de variação de valores (por exemplo entre -3 e +3). O segundo passo consiste no cálculo da probabilidade de uma resposta correta, dado o nível de habilidade encontrado de forma randômica. Na continuação, é escolhido um item para administrar e é obtido um valor aleatório uniforme (sendo 0 ou 1) que representa se tal item foi respondido corretamente ou não (0 = incorreto, 1 = correto) e, baseado neste valor, calcula-se a nova habilidade e o processo se repete.

Segundo Weiss (1985), diversos estudos de simulação têm comparado vários tipos de testes convencionais e adaptativos. Os resultados desses estudos mostram que os TAI's fornecem um desempenho superior, comprovando as hipóteses que motivam a sua aplicação.

3.3.2.2 Exposição dos Itens

Os testes adaptativos se baseiam no princípio de ajustar, da melhor maneira possível, o conteúdo dos testes às características de cada indivíduo sendo avaliado. Para isso, a partir de um banco de

itens, selecionam-se aqueles cujo nível de dificuldade mais se aproxima do nível de competência do indivíduo. Na prática, entretanto, é freqüente encontrar itens que são escolhidos repetidas vezes em diferentes testes, enquanto que outros nunca são utilizados. Dessa forma, a quantidade com que se utiliza um determinado item em distintas aplicações de um teste denomina-se *taxa de exposição de itens.*

Assim, é importante observar que há a necessidade de controlar a taxa de exposição dos itens, prevenindo a sobreposição de alguns deles, o que pode representar uma ameaça para a validade e a segurança do teste. Em geral, os métodos de controle de exposição dos itens podem ser classificados em dois grandes grupos (Olea et al., 1999; Rudner, 1998):

Métodos de seleção aleatória

Em um TAI, se todos os indivíduos possuem a mesma habilidade inicial, todos eles recebem o mesmo primeiro item. O segundo será o mesmo para aqueles que responderem corretamente o primeiro, e do mesmo modo, o segundo item também será o mesmo para aqueles que responderem incorretamente. O mesmo ocorrerá para o terceiro e sucessivos itens a administrar.

Para evitar este problema, o *método de seleção aleatória* consiste em selecionar os primeiros itens de maneira randômica. No primeiro item, aplica-se o procedimento de seleção adotado (Máxima Informação, Bayesiano por exemplo) e entre uma certa quantidade de itens escolhidos (cinco, por exemplo), elege-se um deles de forma aleatória. Aplica-se o método novamente (baseado no nível de habilidade do aluno), e elege-se o segundo item entre os quatro selecionados; o processo se repete e o terceiro item é escolhido entre os três melhores e o quarto entre os dois, sendo que o quinto, e o restante dos itens, se administram aqueles que o método de seleção eleger. Dessa forma, após o quinto item os alunos serão suficientemente diferenciados e receberão itens distintos.

Método do controle direto da taxa de exposição

Atualmente, sendo o mais utilizado em testes adaptativos reais, esse método consiste na atribuição de variáveis a cada um dos itens para controlar de forma direta a sua taxa de exposição. Nessa abordagem, assume-se que a probabilidade de um item ser administrado P(A) (também chamado de valor k) é menor que o valor da taxa máxima de exposição permitida (valor r). Se esse valor for menor, o item é administrado, caso contrário, não.

Nesse método, é necessário fixar o parâmetro k de todos os itens, antes do teste iniciar. Para a especificação deste, é preciso considerar as características que envolvem a aplicação dos testes adaptativos em seu contexto geral, desde o tamanho do banco de itens, a distribuição e habilidade, até a quantidade de indivíduos que serão avaliados.

3.3.2.3 Tamanho do Banco de Itens

A dimensão do Banco de Itens[5] depende diretamente dos objetivos da avaliação na qual os testes adaptativos estão inseridos e as características em que estes estão sendo construídos. Segundo Weiss (1985), uma implementação satisfatória de um TAI[6] tem sido obtida com uma quantidade de apenas 100 itens de boa qualidade (bem especificados, com parâmetros bem definidos), mas aponta que banco de itens com 150 ou 200 itens bem calibrados é preferido. Se a aplicação dos testes adaptativos assume grandes decisões importantes e tem a participação massiva de vários candidatos, é importante ditar um conjunto de restrições, como Balanceamento de Conteúdo e Seleção Randômica e, para isso, será necessário Banco de Itens com grandes dimensões.

De acordo com Olea et al., (1999), um banco de itens deve ter dez vezes mais itens do que a quantidade de questões de um teste qualquer. Por exemplo, se um teste possui trinta questões, a quantidade de itens do banco deve ser no mínimo de trezentos.

3.3.2.4 Critérios de Parada

Uma das principais vantagens dos testes adaptativos é a sua flexibilidade, assumindo extensões variáveis e podendo ser continuado até que um nível satisfatório (ou desejado) de precisão de medida seja alcançado. Os critérios de parada de um TAI consistem em um conjunto de regras ou condições pré-estabelecidas que indicam quando o processo de um teste deve terminar. Assim, vários critérios de parada podem ser adotados, como por exemplo, o nível de habilidade alcançado pelo indivíduo (sendo este inferior ou superior), quantidade de itens administrados, medida do erro padrão do ajuste de habilidade, índice de acertos ou de erros ou uma combinação de todos estes.

A definição do(s) critério(s) de parada também depende dos objetivos e propostas da avaliação na qual os TAI's estão sendo aplicados, pois sua escolha está vinculada ao propósito dos testes e aos resultados a que se querem chegar.

3.3.2.5 Revisão de Respostas

Em um teste tradicional de lápis e papel e nos testes convencionais informatizados, o estudante pode normalmente revisar suas respostas e modificar as que considerar conveniente, considerando que a revisão pode aumentar as chances de melhorar seus resultados. Entretanto, nos testes adaptativos, não é possível revisar e nem sequer modificar os erros acidentais associados ao procedimento de resposta, mesmo porque, tal fato iria contra aos princípios dos TAI's (já que a habilidade do indivíduo é estimada a cada item respondido, e os próximos itens a serem oferecidos dependem da resposta anterior).

Se a revisão e a mudança das respostas fossem permitidas, é possível que os estudantes alterassem o resultado do teste, pois intencionalmente um estudante poderia deixar de responder os primeiros itens, fazendo com que o próximos itens a serem fornecidos sejam mais fáceis (devido predição de

[5]Banco de Itens também pode ser entendido como Pool de Itens
[6]Unidimensional e sem balanceamento de conteúdo

uma baixa habilidade). Dessa forma, após um tempo, o estudante poderia voltar e responder os primeiros itens, conseguindo assim um teste "fácil", alcançando alta taxa de respostas corretas, o que resultaria num resultado artificial e deturpado da habilidade real do indivíduo.

Algumas abordagens de testes adaptativos admitem a revisão de respostas. Tais abordagens permitem que um indivíduo omita a resposta de um item, fazendo com que seja apresentado outro item de dificuldade semelhante; ou ainda, permitem que a resposta seja trocada, mas o nível de habilidade é novamente estimado e todas as questões posteriores já respondidas podem ser perdidas (Olea et al., 1999). Outros métodos, principalmente aqueles que usam do Balanceamento de Conteúdo, concedem o direito de revisão de respostas, mas apenas dentro de um grupo de itens (*Testlets*), transferindo a função de "adaptação" do teste ao fornecimento de grupos de itens, de acordo com o nível do candidato, pois uma vez que todos os itens de um grupo forem respondidos, sua revisão é desautorizada.

3.3.2.6 Conflito e Repetição de Itens

A maioria dos métodos de seleção de itens empregados nos testes adaptativos não possuem um procedimento para evitar a administração de itens semelhantes a uma pessoa que realizou o teste mais de uma vez. Isso significa que um estudante pode receber as mesmas questões se realizar uma avaliação várias vezes.

Para evitar um resultado inválido ou pouco significativo, devido à repetição de itens em diferentes testes, algumas restrições no algoritmo de seleção, bem como no banco de itens devem ser implementadas. Por exemplo: criar um registro de itens já administrados ao estudante em uma sessão de teste; verificar se o estudante já participou de outras sessões e quais itens foram apresentados a ele; e eleger um item não previamente administrado para apresentar ao estudante (Kingsbury and Zara, 1989).

Embora as restrições acima descritas possam evitar alterações na validade dos testes devido a repetição de itens, ainda há outros problemas que precisam ser considerados. Itens conflitantes são aqueles que não podem ser administrados no mesmo teste, pois pode ocorrer que o conteúdo de um contenha a resposta de outro, ou vice–versa. O conflito de itens pode ser evitado com a adição de uma lista (para cada item), contendo todos os itens que lhes são incompatíveis. Esta lista de conflitos deve ser criada antes dos testes adaptativos serem administrados.

Portanto, com a implementação de tais restrições, o procedimento de seleção de itens escolheria uma determinada questão que melhor se ajusta ao nível de habilidade do indivíduo, sem que esta já tivesse sido aplicada em outras ocasiões, e sem conflitos com quaisquer outras questões administradas anteriormente.

È importante ressaltar que quanto maior forem as restrições impostas ao banco de itens e, consequentemente, à execução do teste adaptativo, maior deve ser o volume do banco, já que tais restrições inibem as opções de seleção de itens.

CAPÍTULO 4

Formalização dos Testes Adaptativos: A Teoria de Resposta de Itens (TRI)

4.1 Conceito e Características

A Teoria de Resposta de Itens (TRI) é uma reunião de modelos estatísticos usados para fazer predições, estimativas ou inferências sobre as habilidades (ou competências) medidas em um teste. Através dos modelos estatísticos inerentes ao seu contexto, é possível predizer tais habilidades por meio da correspondência entre a pontuação obtida por um estudante em uma situação de teste e os itens a ele fornecido (Hambleton and Swaminathan, 1985; Rudner, 1998).

Um modelo de resposta de itens especifica a relação entre a pontuação (desempenho) *observada* e ***não observada*** de um estudante em um teste. Essa relação é descrita por uma função matemática baseada nos parâmetros dos itens do teste e nas respostas dos indivíduos envolvidos na avaliação. Assim, diferentes modelos estatísticos são formados pela especificação dos parâmetros dos itens e das habilidades dos estudantes (o que se deseja avaliar) que influenciam seus desempenhos.

A Tabela 4.1 especifica as principais considerações sobre os Modelos de Resposta de Itens, quando ocorre o adequado ajuste entre um modelo de resposta escolhido e os parâmetros dos itens.

Tabela 4.1: Principais considerações sobre modelos de resposta de itens (Hambleton and Swaminathan, 1985)

Considerações sobre os modelos de resposta
São modelos que supõem que o desempenho de um estudante em um teste pode ser predito em termos de uma ou mais habilidades pessoais;
O modelo de resposta de um item especifica a relação entre a pontuação observada de um estudante e as habilidades esperadas ou assumidas que fundamentam os testes;
O desempenho de um estudante em um teste deve ser estimado a partir da pontuação observada em um conjunto de itens de teste.

A Figura 4.1 mostra uma seqüência de passos para estimar os parâmetros dos itens de um banco e obter a escala de pontuação da habilidade.

Primeiro, para estimar os parâmetros dos itens o projetista do teste analisa um conjunto de respostas de itens já administrados a um grupo de estudantes, que o conduz a levantar várias características preliminares sobre seu comportamento, por exemplo, como quais itens foram respondidos freqüentemente de forma correta ou incorreta. Depois da análise desse conjunto de respostas, é necessário selecionar o modelo de resposta que se ajusta da melhor forma possível aos itens observados. O próximo passo é estimar os parâmetros dos itens, combinando com o melhor modelo de resposta escolhido sob os dados analisados. Uma vez realizada a estimação dos parâmetros é definido um procedimento que efetua a conversão da escala de habilidade observada para uma escala conveniente de pontuação real (utilizada nos testes tradicionais de lápis e papel).

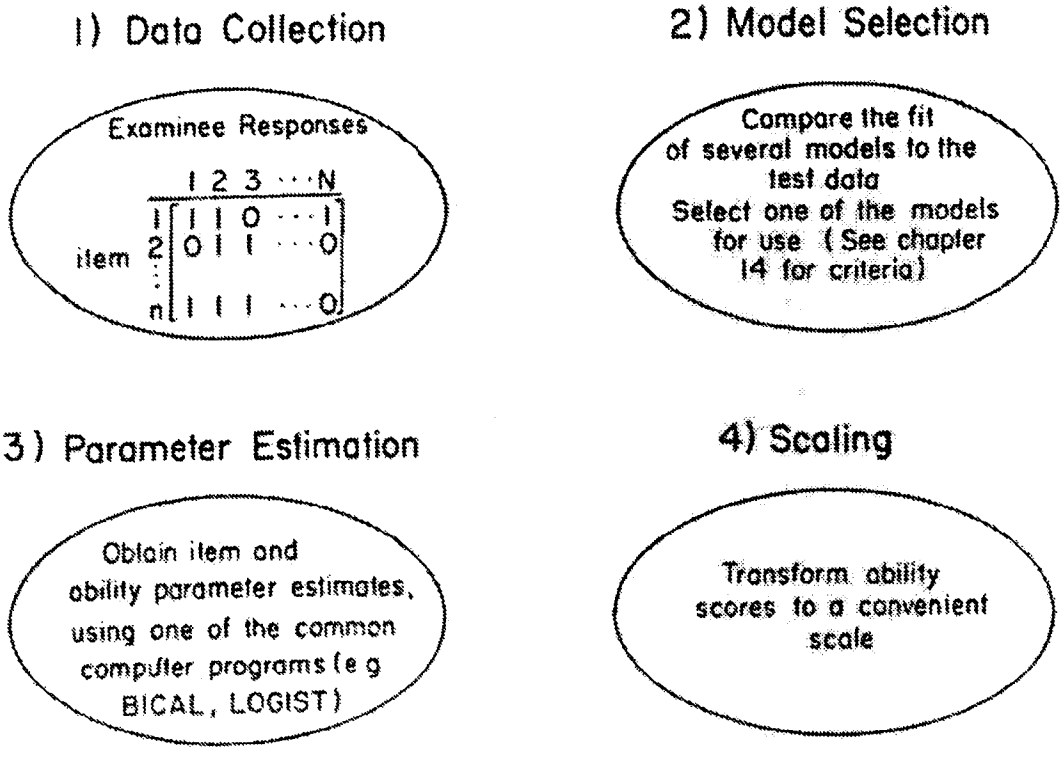

Figura 4.1: Passos para a estimação dos parâmetros dos itens (Hambleton and Swaminathan, 1985)

Talvez a vantagem mais importante dos modelos de resposta de itens é que, dado um conjunto de itens de teste que são adequadamente ajustados a um determinado modelo isto é, os parâmetros dos itens são conhecidos e bem especificados, é possível estimar a habilidade de um indivíduo na mesma escala na qual os dados foram ajustados. Dessa forma, sem levar em consideração o número de itens administrados, a estimação de habilidade para cada estudante será teoricamente verdadeira e mais justa. Portanto, os modelos de respostas fornecem uma maneira de comparar o desempenho

de estudantes, mesmo que eles tenham recebido diferentes conjuntos de itens de teste em uma avaliação.

Quando um modelo de resposta é devidamente escolhido para o ajuste dos itens de teste de um determinado domínio de conhecimento, é obtida uma distribuição de habilidade de acordo com os parâmetros de cada item, fornecendo assim informações sobre o comportamento do mesmo dentro do processo de teste. Tal distribuição é chamada de Curva Característica de um Item[1] (CCI) e está presente em todos os modelos de resposta, tendo suas particularidades em cada um deles. Estes modelos serão estudados na próxima subseção.

4.2 Modelos de Resposta de Itens

Os modelos de resposta de itens são utilizados para uma análise detalhada dos itens de um teste observando principalmente seu comportamento em uma avaliação. Geralmente, um modelo de resposta consiste na aplicação e implementação de várias equações matemáticas (estas com objetivos variados), por exemplo: nível de informação de um item, estimação de parâmetros e habilidade, e medida de erros.

Uma das maneiras de classificar os modelos de resposta de itens, baseia-se nos tipos de resposta que podem ser empregadas. Considerando a forma como o aluno responde os testes, objetiva (usando testes objetivos) ou subjetiva (respostas nominais e dissertativas) é possível aplicar diferentes modelos de resposta. Na Tabela 4.2 são mostrados os principais modelos de respostas.

Tabela 4.2: Modelos de resposta de um item, para respostas objetivas e subjetivas (Hambleton and Swaminathan, 1985)

Tipos de Respostas	***Modelo***
Respostas Objetivas	Modelo Linear (*Latent Linear*)
	Modelo da Escala Perfeita (*Perfect Scale*)
	Modelo da Distância Latente (*Latent Distance*)
	Modelo Normal de 1,2 e 3 parâmetros (One,Two-,Tree Parameter Normal)
	Modelo Logístico de 1,2 e 3 parâmetros (One,Two-,Tree Parameter Logistic)
	Modelo Logístico de 4 parâmetros(Four-Parameter Logistic)
Respostas Subjetivas	Resposta Nominal (*Nominal Response*)
	Resposta de Classificação (*Graded Response*)
	Modelo de Crédito Parcial (*Partial Credit Model*)

Todos os modelos descritos na Tabela 4.2 são unidimensionais, ou seja, assumem que os itens do teste foram ajustados para medir uma única habilidade. As principais características que diferenciam os modelos de resposta são suas fórmulas matemáticas, conseqüentemente suas curvas

[1] *Item Curve Caracteristic* - ICC em inglês

características, e a forma de pontuação empregada. Neste trabalho veremos apenas os principais modelos de respostas aplicados aos testes objetivos, especificamente os modelos de um, dois e três parâmetros, pois a aplicação desta teoria neste projeto contempla os testes objetivos para avaliação da proficiência em língua inglesa.

4.2.1 Modelo Normal de Dois Parâmetros

Segundo Hambleton and Swaminathan (1985), Lord (1952, 1953a) propôs o primeiro modelo de resposta de dois parâmetros, no qual a curva característica do item assume sua forma sob a distribuição normal dada por:

$$P_i(\theta) = \int_{-\infty}^{a_i \cdot (\theta - b_i)} \frac{1}{\sqrt{2\pi}} \cdot \exp^{-z^2/2} dz$$

onde $P_i(\theta)$ é a probabilidade que um determinado aluno selecionado aleatoriamente com habilidade θ, responda a um item i corretamente; as variáveis b_i e a_i são parâmetros que caracterizam o item i; e z é o desvio normal da distribuição com média em b_i e desvio padrão $1/a_i$. O parâmetro b_i é comumente referenciado como índice de ***dificuldade*** e representa o nível de dificuldade de um determinado item, que inicialmente é atribuído pelo projetista do teste baseado em seu conhecimento sobre o item, podendo ser alterado à medida que um estudante o responde correta ou incorretamente. O parâmetro a_i é chamado de índice de *discriminação* e representa a capacidade que um item possui de discriminar indivíduos que apresentam nível de habilidade muito próximo.

Geralmente, os valores de b_i variam tipicamente, neste modelo, entre -2.0 e +2.0. Valores de b_i próximos ou iguais a -2.0 correspondem a itens fáceis, e valores de b_i próximos a +2.0 correspondem a itens difíceis. Por sua vez, o item de discriminação a_i, é definido teoricamente no intervalo de $-\infty$ a $+\infty$. Entretanto, itens que possuem o valor de a_i negativos são descartados para os testes de habilidade, pois não possuem boa capacidade de discriminação. Do mesmo modo, valores de a_i muito altos não são usuais, já que resultam em curvas características de item muito acentuadas. Contundo, valores baixos de a_i permitem uma distribuição gradual da curva, em função e ao longo da escala de habilidade do estudante. Assim, valores típicos do índice de discriminalidade situam-se entre 0 e +2.

A Figura 4.2 mostra a curva característica de um item para o modelo normal de dois parâmetros. Note que quando o valor de a é negativo ocorre uma inversão de sentido da curva, indicando que o item possui baixo poder discriminativo. Para os valores de b vemos que seu aumento é gradativo, de acordo como os valores de θ, que significa dizer que quanto mais alto o valor da habilidade maior será a probabilidade de um indivíduo acertar um determinado item.

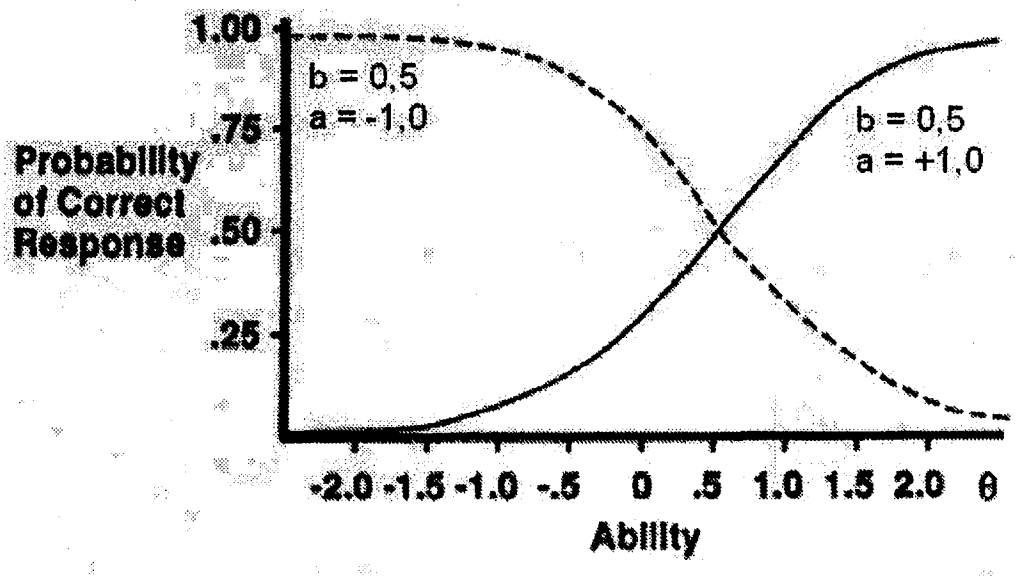

Figura 4.2: Curva Característica de um Item, com valores de *a* positivo e negativo (Baker, 1992)

4.2.2 Modelo Logístico de Dois Parâmetros

A curva característica de um item nos modelos logísticos de dois parâmetros é dada pela fórmula matemática:

$$P_i(\theta) = \frac{\exp^{D \cdot a_i \cdot (\theta - b_i)}}{1 + \exp^{D \cdot a_i \cdot (\theta - b_i)}}$$

com $i = 1, 2, 3, ..., n$.Existe uma maneira alternativa para escrever a função $P_i(\theta)$ acima. Multiplicando o numerador e o denominador da equação pelo fator: $\exp^{-D \cdot a_i \cdot (\theta - b_i)}$ a fórmula torna-se:

$$P_i(\theta) = (1 + \exp^{-D \cdot a_i \cdot (\theta - b_i)})^{-1}$$

e que pode ser escrita como uma melhor alternativa e de leitura mais fácil da seguinte forma:

$$P_i(\theta) = \frac{1}{1 + \exp^{-D \cdot a_i \cdot (\theta - b_i)}}$$

com $i = 1, 2, 3, ..., n$

Os parâmetros a, b e θ possuem essencialmente a mesma interpretação que no modelo normal, e a constante D é o fator de escala. A maior vantagem do modelo logístico de dois parâmetros em relação ao modelo normal, é a facilidade de lidar com a equação, já que esta é mais tratável matematicamente e não envolve a função de integração presente no modelo normal. Além disso, o modelo logístico é computacionalmente mais simples de implementar (Hambleton and Swaminathan, 1985; Revuelta and Ponsoda, 1997).

A Figura 4.3 mostra a forma geral das curvas características, tanto do modelo normal como logístico. Note que suas aparências são similares, diferenciando apenas no ponto de inclinação, pois compartilham os mesmos parâmetros. Os dois modelos têm o ponto de inflexão situado no parâmetro b, visto que este valor tem o mesmo significado nos dois modelos. Entretanto, os mesmos valores numéricos do parâmetro a produzem diferentes pontos de inclinação.

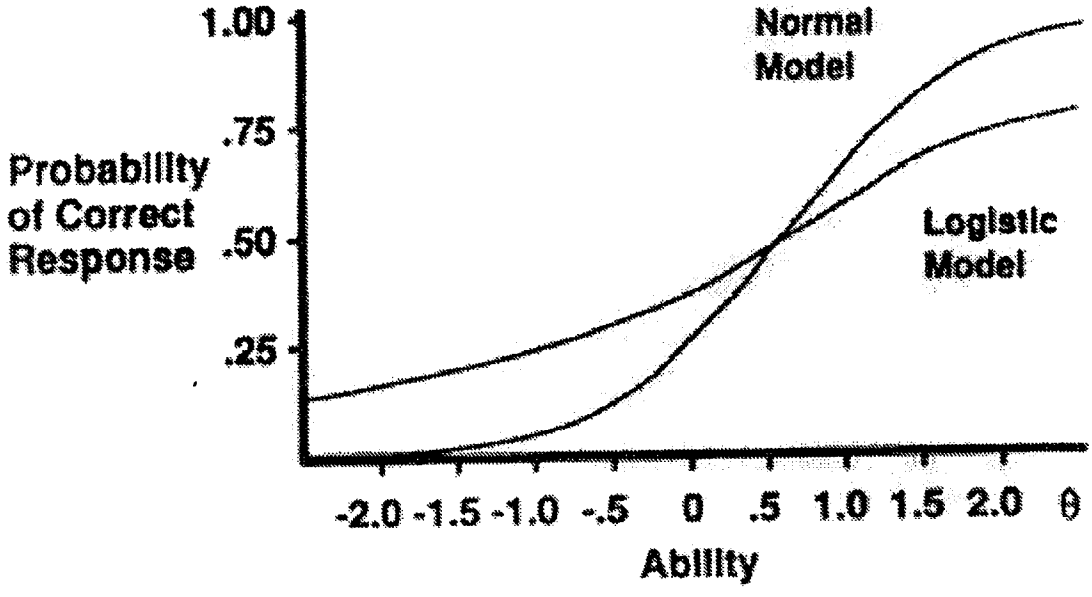

Figura 4.3: Curvas Características de um Item baseado nos modelos Normal e Logístico com b = 0.6 e a = 1.2 (Baker, 1992)

Os modelos de dois parâmetros, tanto o normal como o logístico assumem que os estudantes não podem acertar um determinado item dada a uma baixa habilidade, mas existem possíveis situações em que um indivíduo consegue responder a um item corretamente tendo um baixo nível de habilidade ou competência. O modelo logístico de três parâmetros adiciona em suas equações essa possibilidade, e será estudado na próxima subseção.

4.2.3 Modelo Logístico de Três Parâmetros

O modelo logístico de três parâmetros pode ser obtido a partir do modelo logístico de dois parâmetros pela adição de um terceiro parâmetro, chamado c ou índice de *adivinhação*. Tal índice representa a possibilidade de se ter uma resposta correta dado um baixo nível de habilidade, ou seja, é a mínima probabilidade de acerto.

A fórmula matemática do modelo logístico de três parâmetros é a seguinte:

$$P_i(\theta) = c_i + \frac{(1 - c_i)}{[1 + \exp^{-D \cdot a_i \cdot (\theta - b_i)}]}$$

na qual $P_i(\theta)$ é a probabilidade que um indivíduo com nível de habilidade θ responda um item i corretamente, b_i é o índice de dificuldade, a_i é o índice de discriminação, c_i é o de adivinhação e D é o fator de escala.

O valor do parâmetro c pode ser atribuído a um item por meio do desenvolvimento das opções distratoras pertencentes a um teste objetivo. Conforme visto no Capítulo 2, os distratores são as opções atrativas, mas incorretas, de uma determinada questão. Os estudantes que possuem uma baixa habilidade são atraídos por essas "opções incorretas" (Hambleton and Swaminathan, 1985). Os valores do índice de adivinhação devem ser baixos, já que um valor alto indica uma grande possibilidade de acerto de uma questão dada uma baixa habilidade.

A Figura 4.4 fornece um exemplo de uma curva característica de um item baseado no modelo logístico de três parâmetros. Primeiramente, observamos na parte inferior do gráfico uma linha assintótica a um valor de $P_i(\theta) = 0.15$, que define o limite inferior da curva. Segundo, o limite superior da curva do gráfico é de 1.0. Finalmente, os valores de a e b definem as características da curva de um item (ao longo da escala de habilidade) pela equação logística de três parâmetros, a qual é compreendida pelo limite superior de 1.0 e o valor de c que define o limite inferior, nesse caso $c = 0.15$. Assim, os valores de $P_i(\theta)$ dependerão diretamente dos valores de a, b, c e θ, sendo que o valor de c será o limite inferior da curva (Baker, 1992).

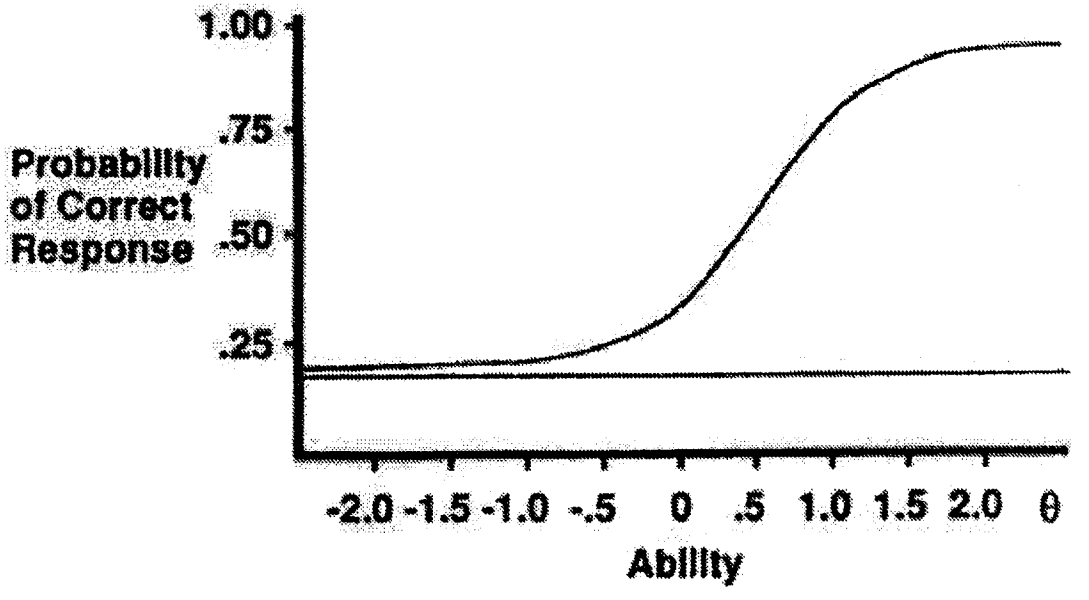

Figura 4.4: Curva Característica de um Item baseado no modelo logístico de três parâmetros, com b=0.6, a=1.2 e c=0.15 (Baker, 1992)

Para obter o modelo logístico de dois parâmetros a partir do modelo de três parâmetros, basta assumir que o valor do índice de adivinhação c é sempre igual a zero ($c = 0$).

4.2.4 Modelo Logístico de Um Parâmetro (*Rasch Model*)

O modelo logístico de um parâmetro, também chamado de ***Rasch Model*** em homenagem a seu criador Georg Rasch, citado por Hambleton and Swaminathan (1985), pode ser visto como um modelo de resposta, no qual a curva característica de um item é baseada em uma função de apenas um parâmetro.

O Rasch Model é um caso especial do modelo logístico de três parâmetros de maneira que todos os itens assumem uma capacidade semelhante de discriminação (mesmo valor do parâmetro a) e mínima possibilidade de adivinhação (baixos ou nulos valores do parâmetro c). Dessa forma, o único parâmetro que o modelo utiliza é o índice de dificuldade, representado pelo valor de b. Assim, este modelo pode ser considerado como um caso especial dos modelos de dois e três parâmetros (Hambleton and Swaminathan, 1985; Baker, 1992), pois possui algumas propriedades específicas que o torna atrativo: primeira, o modelo envolve um único parâmetro facilitando o trabalho de administração dos itens; e segundo, agiliza a tarefa de estimação do parâmetro em relação a outros modelos. A equação do Rasch Model é a seguinte:

$$P_i(\theta) = \frac{\exp^{D \cdot a_i \cdot (\theta - b_i)}}{1 + \exp^{D \cdot a_i \cdot (\theta - b_i)}}$$

A Figura 4.5 mostra a curva característica de um item regida pela equação logística de um parâmetro. O gráfico da Figura 4.5 relata a probabilidade de uma resposta correta de um item dada uma escala de habilidade.

Conforme pode ser visto, a forma da curva não parece com qualquer uma das curvas vistas anteriormente (modelos de dois e três parâmetros), já que a diferença básica desta curva característica é que tanto a escala de habilidade θ quanto o valor de b (índice de dificuldade) variam no intervalo de 0 ao ∞ (infinito) com uma unidade arbitrária de medida, isso porque neste modelo, as chances de sucesso de um aluno com habilidade θ em um item com nível de dificuldade b são dadas pela relação direta entre a habilidade do aluno e dificuldade do item.

Dadas as principais formas e modelos de repostas dos itens, na próxima subseção veremos como os testes adaptativos podem ser implementados usando os métodos da Teoria de Resposta de Itens e seus diversos modelos de resposta.

4.3 Função de Informação de um Item

Os parâmetros que caracterizam um item, além de ditarem a forma das curvas características dos modelos de resposta, também são combinados de maneira a medir o nível de informação que um determinado item fornece ao longo da distribuição da escala de habilidade. Dessa forma, podem descrever quão precisamente um item ajusta-se a um dado nível de habilidade.

O nível de informação de um item representa a quantidade proporcionada, baseado em seus parâmetros, a um determinado ponto da escala de habilidade. A ***curva de informação de um item*** tem seu "pico" representado pelo índice de dificuldade (parâmetro b) e a altura desse "pico"

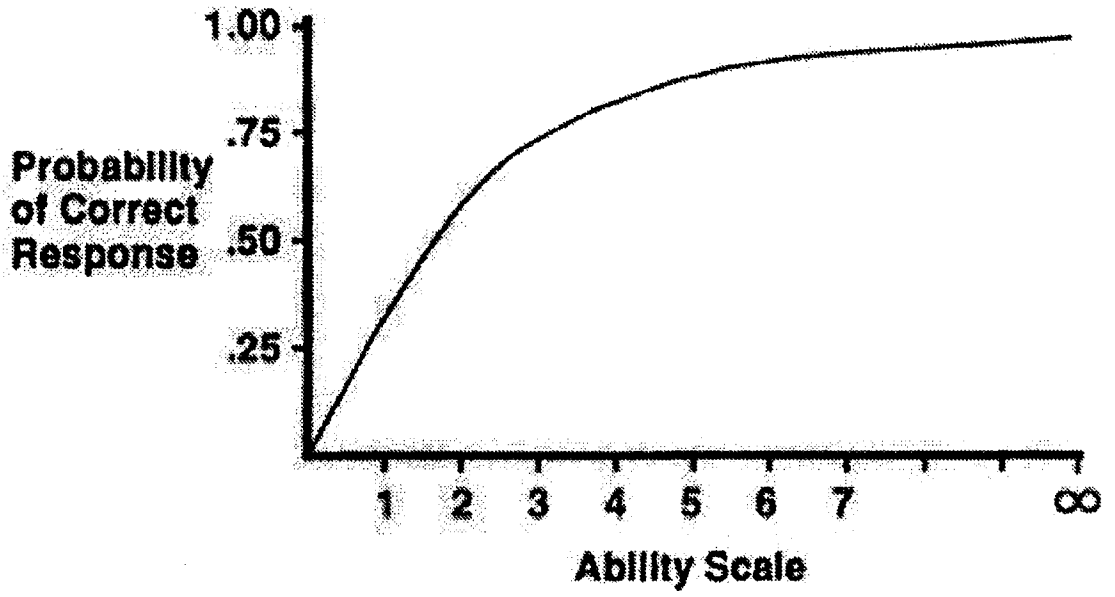

Figura 4.5: Curva Característica de um Item baseado no Rasch Model, com b = 1.822 (Baker, 1992)

representa o grau de precisão do item em um ponto da escala de habilidade, e é regida pelo índice de discriminação (parâmetro a) (Weiss, 1985); em outras palavras, o parâmetro de discriminação a revela a extensão da curva de informação ao longo da escala de habilidade. Assim, um item que possui baixo valor de a terá um baixo pico e sua distribuição será mais achatada, enquanto que um pico alto da curva de informação com achatamento mínimo indica um alto valor do índice de discriminalidade.

Dessa maneira, dado os parâmetros da TRI para um conjunto de itens, o valor do nível de informação pode ser calculado para cada possível valor da habilidade avaliada.

A equação abaixo representa o nível de informação de um item, dado um valor θ da escala de habilidade. A função de informação de um item pode variar de acordo com a quantidade de parâmetros envolvidos no modelo de resposta adotado, e assume a seguinte forma:

$$I_i(\theta) = \frac{(P\prime_i(\theta))^2}{P_i(\theta) \cdot Q_i(\theta)}$$

no qual

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

e $P\prime_i(\theta)$ é a primeira derivada da função $P_i(\theta)$.

A Figura 4.6 mostra 10 curvas de informações de itens hipotéticos. O eixo horizontal representa a habilidade θ sendo medida e o eixo vertical representa o nível de informação de um item dada uma habilidade qualquer.

Valores altos no nível de informação indicam que um item fornece maior quantidade de informação naquele nível de habilidade. O item 9 por exemplo, tem o maior nível de informação, e é o item mais discriminativo entre os 10, seguidos pelos itens 4 e 1. Os itens menos discriminativos são os itens 3 e 7 que possuem a curvas mais achatadas entre os 10 itens. Como mostrado na Figura 4.6 a curva de pico mais alto (item 9) possui menor distância nas bases. Isso significa que o item tem capacidade de diferenciar níveis de habilidade muito próximos.

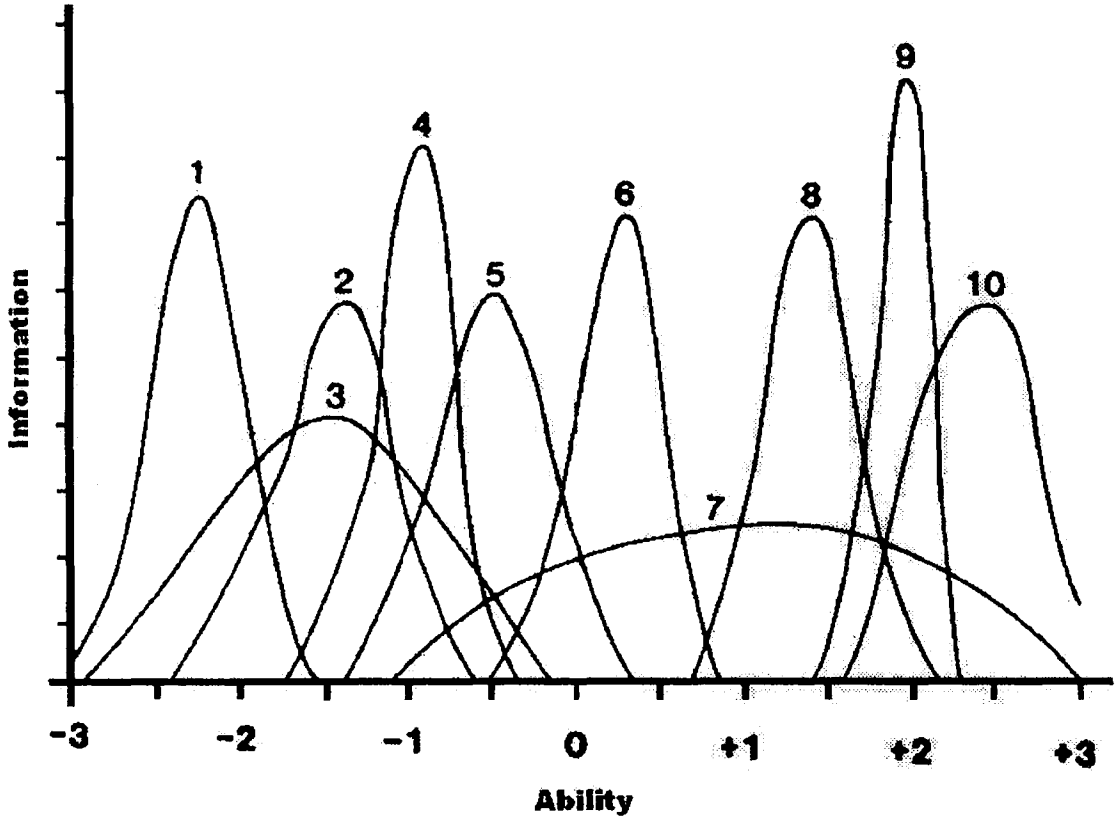

Figura 4.6: Curvas de informação de 10 itens hipotéticos (Weiss, 1985)

Note, ainda na Figura 4.6, que as curvas de informação dos itens diferem sua localização ao longo da escala de habilidade. Por exemplo, os itens 1, 6 e 8 possuem curvas de informação bastante parecidas, tanto na altura do pico quanto no poder de discriminação, mas são localizadas em diferentes regiões. Observe que o pico de informação da curva do item 1 corresponde ao valor de habilidade θ próximo de -2.5, enquanto que o pico da curva do item 6 é próximo a 0.5 e do item 8 a 1.5. Apesar de terem formas parecidas o item 1 diferencia indivíduos com baixa habilidade, e é mais fácil que os outros dois itens (6 e 8), já o item 8 é mais difícil que o item 6, porque em geral o item 8 diferencia indivíduos com maiores níveis de habilidade.

Como os índices de dificuldade e discriminalidade (parâmetros b e a) são independentes em cada item, pode ocorrer que dois ou mais itens tenham o mesmo valor de dificuldade mas discriminalidades distintas. Quando isso acontece, as curvas de informação dos itens terão pico no mesmo ponto da escala de habilidade, mas a altura e o achatamento da curva serão diferenciados. Tal fato pode ser visto entre os itens 2 e 3 na Figura 4.6, pois ambos possuem o mesmo índice de dificuldade -1.5, mas o item 2 é mais discriminativo do que o item 3 pois fornece maior informação.

4.4 Abordagens de Seleção de Itens Baseados na TRI

O advento da Teoria da Resposta de Itens proporcionou o desenvolvimento de várias abordagens de seleção de itens voltadas para os testes adaptativos. Esta subseção faz uma revisão dos procedimentos clássicos e alternativos de seleção de itens destacando suas vantagens e desvantagens.

Os procedimentos de seleção de itens podem ser categorizados em dois grupos: os *pré-estruturados* e os de *ramificação variável*. A maior diferença entre estes dois grupos está na capacidade de adaptação permitida. Os procedimentos pré-estruturados resultam em uma adaptação determinística, com ramificação mais contida durante a execução do teste e os procedimentos com ramificação variável permitem que o processo de adaptação e ajuste dos itens aos alunos seja mais específico e eficiente, quase ilimitado ao longo do teste (Kingsbury and Zara, 1989).

O grupo de procedimento *pré-estruturados* foi criado quando os testes adaptativos começaram a se desenvolver, e os mais conhecidos são: *Teste de Dois Estágios* (Two Stage Test), o *Teste Piramidal* (Pyramidal Test) e os *Testes Adaptativos Estratificados* (Stratified Adaptative Testing), estudados na seção 3.1.2 do Capítulo 3.

As técnicas pré-estruturadas de seleção de itens não são muito adequadas para as avaliações adaptativas, pois elas limitam a capacidade de adaptação que pode ser realizada sobre um indivíduo, além de não usarem todas as informações disponíveis dos estudantes para a realização da seleção dos itens (Kingsbury and Zara, 1989; Olea et al., 1999).

O grupo de procedimento de *ramificação variável* é o de desenvolvimentos mais recente na seleção de itens aplicados aos testes adaptativos, destacando entre eles o **Método da Máxima Informação** e o **Método Bayesiano**.

4.4.1 Método da Máxima Informação

O Método da Máxima Informação consiste na escolha de itens que irão maximizar as informações existentes em cada item quando o estudante responde uma questão. Para selecionar o "melhor" item a informação fornecida por cada questão disponibilizada no banco de itens é calculada, dada a habilidade corrente do indivíduo e os parâmetros (a e/ou b e/ou c) dos itens. Aquele que fornecer mais informação em função da habilidade do indivíduo será selecionado e administrado, assim os cálculos para descobrir qual será o melhor item são sempre realizados depois que o estudante responde uma questão e sua nova habilidade é computada.

As vantagens fornecidas pelo Método da Máxima Informação sobre os métodos pré-estruturados são consideráveis. Primeiro, porque sempre o item mais informativo será administrado, e segundo, devido a esse fato, a eficiência do teste é aprimorada. Além disso, este método permite diferentes ramificações ao longo do teste, devido à busca constante do melhor item a cada habilidade estimada do indivíduo, fornecendo uma medida precisa (Kingsbury and Zara, 1989).

As desvantagens do método de seleção de Máxima Informação estão intimamente ligadas aos procedimentos de busca de itens, principalmente quando existe um grande banco de questões, pois a constante e repetida tarefa de seleção pode consumir tempo de administração desses itens, fazendo com que o indivíduo fique esperando, o que pode ser impraticável em algumas situações de teste.

Especificação do Modelo de Máxima Informação e sua aplicação nos TAI's

A maioria dos testes adaptativos informatizados desenvolvidos utiliza as informações de um item para executar a tarefa de adaptação ao nível do estudante. Assim, a estratégia da máxima informação aplicada aos testes adaptativos indica que o item a ser administrado a qualquer momento do teste é o que fornece maior quantidade de informação, dado o nível corrente de habilidade do aluno (Weiss, 1985).

Para exemplificar esta estratégia de seleção observe as Figuras 4.7 4.8 e 4.9. Como exposto na Figura 4.6 o eixo horizontal representa o nível de habilidade e o eixo vertical representa o nível de informação. Na Figura 4.7, a linha vertical pontilhada ao nível 0 da escala de habilidade representa a estimação inicial da habilidade de um determinado estudante no início do teste. Neste exemplo, com a habilidade no valor 0, três itens fornecem níveis de informação diferentes: os itens 5, 6 e 7. As curvas de informação dos itens 5 e 7 se cruzam no mesmo ponto em intersecção com a linha pontilhada, de maneira que estes itens fornecem a mesma quantidade de informação a este nível. Entretanto o item 6 fornece mais informação em relação aos dois itens anteriores. Desse modo, como nenhum outro item fornece mais informações, o item 6 é administrado ao estudante, que por sua vez o responde e um novo nível de habilidade é calculado.

Quando um aluno responde corretamente um item sua estimação de habilidade é levemente elevada, enquanto que uma resposta incorreta conduz a um decréscimo da estimação de habilidade. Assumindo que o aluno respondeu incorretamente o item 6, o novo nível de habilidade estimado será menor, valor de $\theta = -1.0$ (como representado na Figura 4.8 na linha pontilhada). Para o nível de habilidade com valor -1.0, o item 7 fornece a menor quantidade de informação, sendo que os itens 5 e 2 fornecem, de certa forma, a mesma quantia. O item 3, por sua vez, fornece ligeiramente mais informação que os outros anteriores e o item 4 fornece o maior grau de informação.

Depois que o item 4 foi administrado e respondido, o novo nível de habilidade é calculado, recebendo o valor de -0.5, supondo que o item 4 foi respondido corretamente (como indica a linha pontilhada da Figura 4.9). A este nível de habilidade, os itens 3, 5 e 7 fornecem diferentes níveis de informação, e como o item 5 fornece o maior nível, este é administrado.

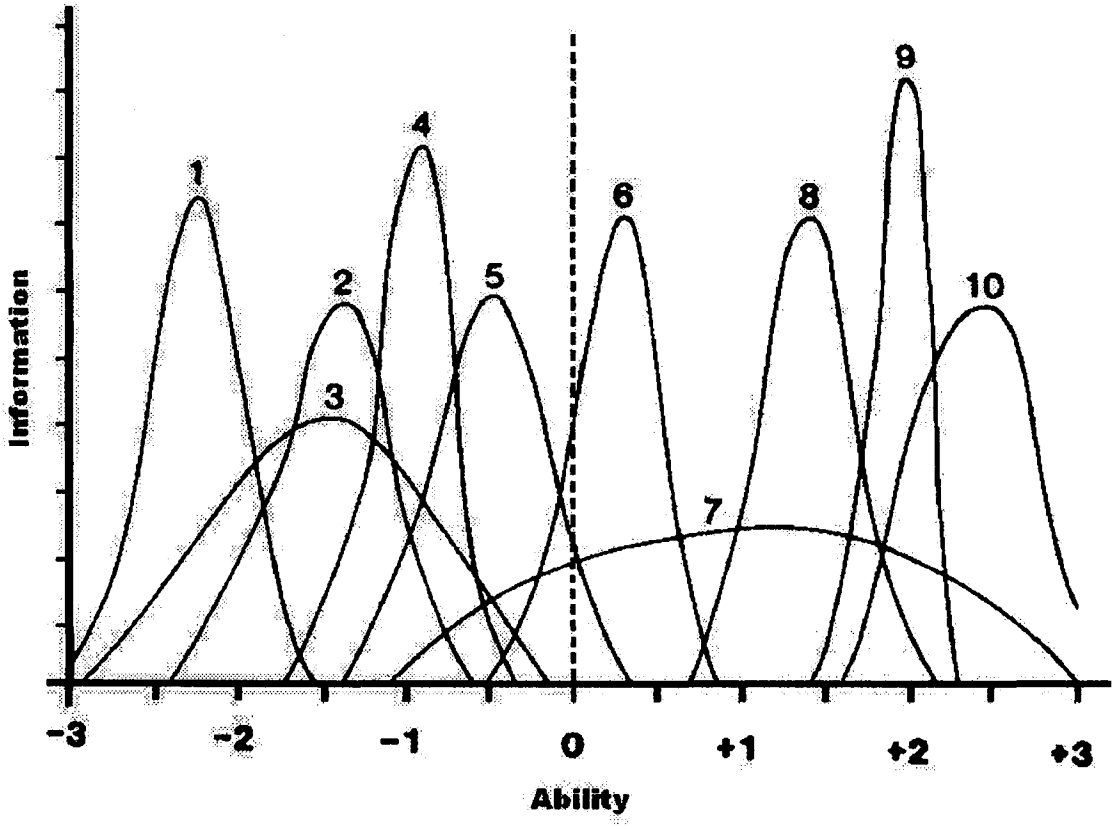

Figura 4.7: Representação das curvas de informação de 10 itens e o nível de habilidade com valor 0 (Weiss, 1985)

Este processo de seleção de itens é mantido até que seja encontrado o critério de parada adotado ou os itens disponíveis se esgotem. Com o uso de um banco de itens reais, diferentes critérios de parada podem ser implementados.

Inerente a esta estratégia de seleção de itens, os métodos de pontuação também podem produzir um erro padrão após a administração de cada item, o qual serve como excelente critério de parada, de maneira que o teste pode ser continuado até que uma determinada medida de erro possa ser alcançada.

4.4.2 Método Bayesiano

O Método Bayesiano (Owen, 1975) de seleção de itens é similar ao Método da Máxima Informação, mas conceitualmente mais complicado e mais custoso de implementar (Kingsbury and Zara, 1989).

Cada indivíduo começa o teste com uma estimação inicial de habilidade e um intervalo de confiança associado a essa estimação. Eles são operacionalizados como a média e variância de uma distribuição normal da habilidade sendo medida. À medida que cada item é respondido, uma

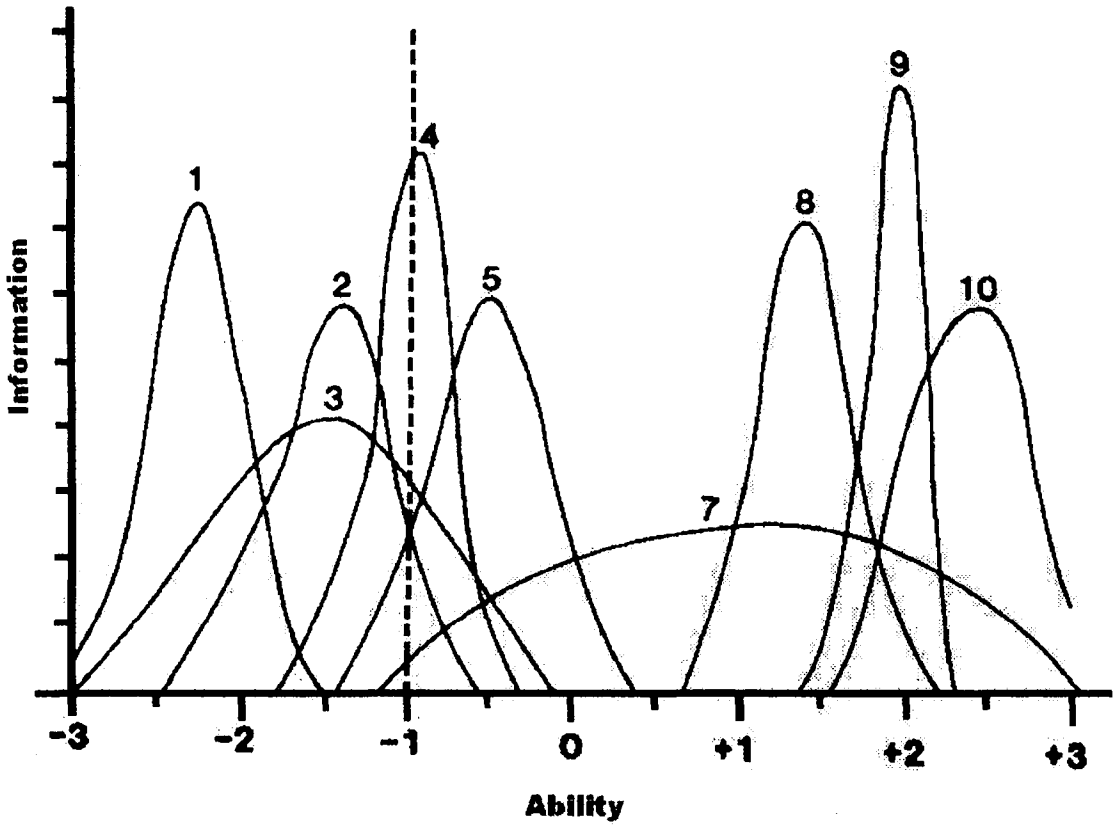

Figura 4.8: Representação das curvas de informação de 9 itens e o nível de habilidade com valor -1 (Weiss, 1985)

nova estimação de habilidade é calculada, usando a resposta do aluno e os valores da distribuição anterior, de maneira que uma nova distribuição de habilidade com um novo intervalo de confiança seja produzido. O Método Bayesiano de seleção escolhe um item que mais reduz esse novo valor da Variância, em outras palavras, tal variância é calculada para cada item disponível do banco, dada a habilidade corrente do indivíduo e os parâmetros dos itens, sendo escolhida a que fornece o menor valor.

As desvantagens associadas ao Método Bayesiano estão também relacionadas aos problemas de busca de itens, principalmente quando existe um grande banco de questões. Outra desvantagem que cabe salientar, diz respeito ao uso do valor da variância sobre a estimação de habilidade para selecionar o "melhor" item, pois em determinadas situações pode ocorrer uma escolha não apropriada, já que a verdadeira estimativa de habilidade não está sendo considerada.

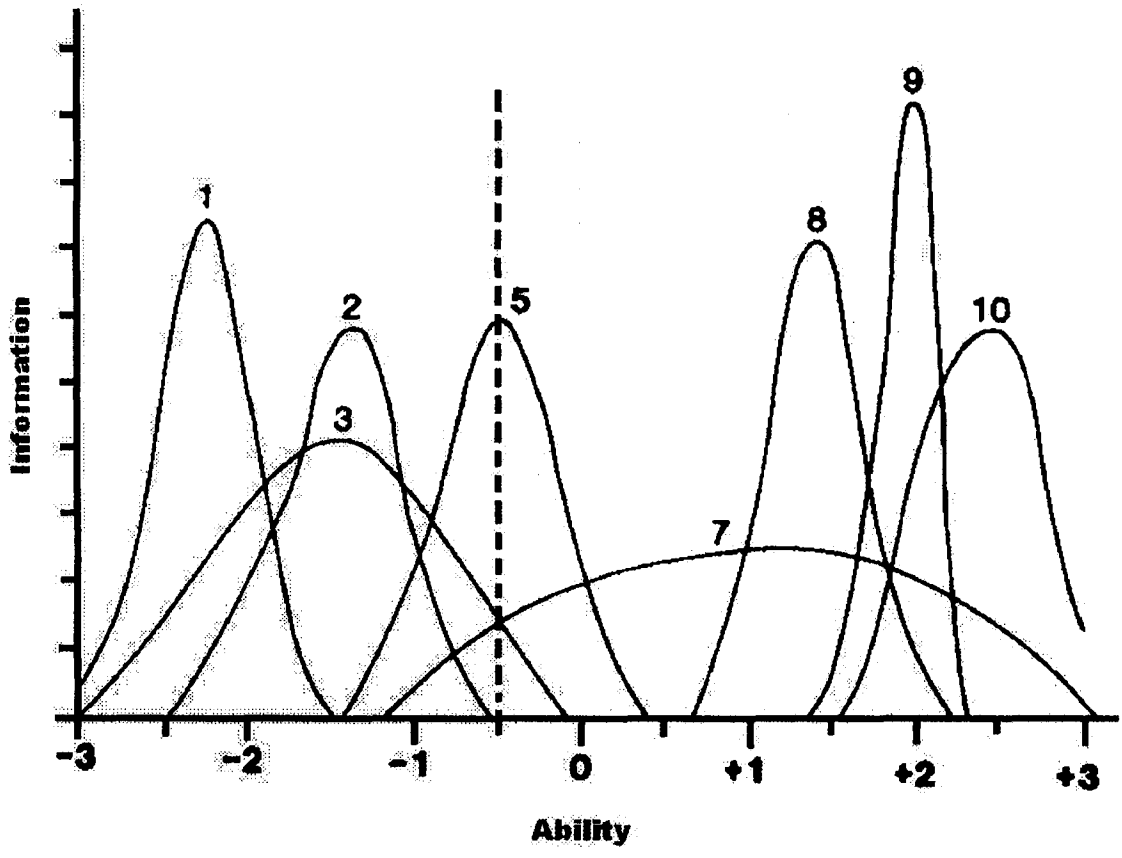

Figura 4.9: Representação das curvas de informação de 8 itens e o nível de habilidade com valor -0,5 (Weiss, 1985)

4.4.3 Técnicas de Busca de Itens

Conforme dito anteriormente, a principal desvantagem relacionada com as abordagens de seleção de itens, tanto do Método da Máxima Informação quanto o Método Bayesiano, é a tarefa de busca de itens, sobretudo quando existe um grande banco de itens. O tempo (*Delay*) existente entre a resposta do indivíduo e a administração do próximo item não pode ser longo. Portanto, Kingsbury and Zara (1989) sugerem três técnicas que podem ser utilizadas para eliminar ou diminuir o intervalo de busca entre uma questão e outra.

A primeira técnica requer que um algoritmo realize duas buscas enquanto um determinado estudante responde uma questão. Uma busca consiste em supor que o aluno acertou a questão corrente, e a outra supõe que ele a errou. Dessa forma, depois da sua resposta, o item apropriado é fornecido, e o *Delay* envolvido nesse processo é mínimo, já que o algoritmo já realizou a busca enquanto o estudante respondia a questão. Entretanto, esta técnica não é recomendada para testes em que o tempo de resposta requerido de cada questão é pequeno, pois pode aumentar o atraso entre a administração dos itens.

A segunda técnica de busca que pode diminuir o *Delay* consiste em limitar o número de itens disponíveis em dado momento do teste, considerando apenas os itens que possuem potencial para serem escolhidos de acordo com a habilidade provisória do estudante. Contudo, essa técnica pode não ser adequada quando o banco de itens é muito extenso, pois encontrar os itens "possíveis" e entre estes o mais adequado pode levar muito tempo.

A terceira técnica requer que todos os itens sejam pré-processados, antes do início do teste, criando uma tabela com os níveis de informação fornecidos, assim, o procedimento seleciona o primeiro item, que não tenha sido administrado, a partir de uma busca na tabela do item que oferece a maior informação à corrente habilidade do indivíduo.

4.5 Estimação de Parâmetros

Como mencionado no Capítulo 3, o processo de estimação de parâmetros consiste no cálculo dos parâmetros pertencentes aos itens do banco, sendo estes referentes a um determinado modelo da Teoria de Resposta de Itens (TRI), ou ainda ao nível de habilidade θ de cada aluno durante um teste. No âmbito dos testes adaptativos informatizados, a estimação de parâmetros acontece em dois momentos distintos: o primeiro momento acontece durante a construção do banco de itens, fazendo parte do processo de calibração dos itens que irão compor o banco, antes da aplicação do teste. O segundo momento acontece durante a execução de um teste adaptativo, e corresponde ao cálculo do valor da nova habilidade de um aluno após o mesmo ter respondido a um item.

De maneira resumida, a tarefa de estimação de parâmetros baseia-se em descobrir o valor de uma determinada variável que mais se aproxima de seu valor real, dado para a execução do cálculo, uma amostra de valores aleatórios regidos por uma função ou modelo de distribuição.

Por exemplo: seja p uma variável qualquer que se deseja estimar, e X um conjunto de amostras aleatórias $(X_1, X_2, X_3, ..., X_n)$ pertencentes a um modelo de distribuição $f(X)=p$; então a tarefa é descobrir se existe alguma estimativa $H(X_1, X_2, X_3, ..., X_n)$ que possa ser empregada para encontrar um valor para a variável p. Dessa forma, o resultado da estimação, usando o modelo definido por $f(X) = p$, é um valor $\hat{p}$ que representa a estimativa de p, a qual mais se aproxima de seu valor real (Meyer, 1973).

Para a estimação dos parâmetros dos itens e do valor da habilidade dos estudantes em um teste adaptativo são utilizados principalmente o método da Estimativa de Máxima Verossimilhança[2] (EMV) e o Método Bayesiano. A seguir veremos a descrição de cada um desses métodos.

4.5.1 Método da Estimativa de Máxima Verossimilhança (EMV)

EMV é um procedimento geral de estimação de parâmetros que, segundo Meyer (1973) e Bilmes (1998), conduz a valores de estimativas razoáveis quando aplicado.

[2] Maximum Likelihood Estimation - MLE em inglês

A fim de explicar o método EMV, definimos uma função de distribuição de probabilidade (fdp), ou modelo de distribuição, da forma $f(X|\theta)$ que depende da amostra X em relação a θ, que é a variável que desejamos estimar.

Seja X um conjunto de valores amostrais $(x_1, x_2, x_3, ..., x_n)$ conhecidos. Definimos a *função de verossimilhança* L como sendo a seguinte função em relação a θ:

$$L(x_1, x_2, x_3, ..., x_n|\theta) = f(x_1|\theta)f(x_2|\theta)f(x_3|\theta)...f(x_n|\theta)$$

ou, em termos de probabilidades podemos generalizar:

$$f(x_1|\theta)f(x_2|\theta)f(x_3|\theta)...f(x_n|\theta) = P(x_1|\theta)P(x_2|\theta)P(x_3|\theta)...P(x_n|\theta)$$

A estimativa da variável θ será o valor que maximizar a função $L(x_1, x_2, x_3, ..., x_n|\theta)$. Em outras palavras, devemos calcular o valor de θ que torna a função L máxima.

Definição: *A estimativa de máxima verossimilhança de θ ($\hat{\theta}$) baseada em uma amostra aleatória $x_1, x_2, x_3, ..., x_n$ é aquele valor de θ que torna $L(x_1, x_2, x_3, ..., x_n|\theta)$ máxima, considerada como função de θ para uma dada amostra $X_1, X_2, X_3, ..., X_n$, sendo L definida pela equação $L(x_1, x_2, x_3, ..., x_n|\theta) = f(x_1|\theta)f(x_2|\theta)f(x_3|\theta)...f(x_n|\theta)$ e, é geralmente referida como estimativa MV (Meyer, 1973).*

Na maioria das tarefas de estimação de parâmetros, a variável θ representa um único valor real isolado. Entretanto, em alguns casos é necessário estimar dois ou mais valores (por exemplo, a estimação dos parâmetros do modelo da TRI), em tais casos a variável θ pode representar um vetor contendo as variáveis que se quer estimar, por exemplo, $\theta = (a, b, c)$.

Dado o fato de que devemos encontrar o valor de θ que maximiza a função L, é conveniente aplicarmos a função *ln(x)*, visto que o logaritmo natural é uma função crescente de x. Portanto, com a aplicação do *ln* sobre a função L, a equação de distribuição fica:

$$lnL(x_1, x_2, x_3, ..., x_n|\theta)$$

e alcança o valor máximo de θ para o mesmo valor que faria a função L. Por isso, em linhas gerais admitimos que θ seja um valor real, ou um vetor de valores reais, e que a função L seja uma função derivável em relação a θ, possibilitando, assim, conseguir a estimativa de máxima verossimilhança pela solução de:

$$\frac{\partial}{\partial\theta} lnL(x_1, x_2, x_3, ..., x_n|\theta) = 0.$$

a qual é conhecida como *equação de verossimilhança.*

Se a variável θ for um vetor de valores, como por exemplo, $\theta = (\alpha, \beta)$, a equação acima deverá ser substituída pelas equações simultâneas sobre cada variável pertencente ao vetor, assim as equações

parciais ficam:

$$\frac{\partial}{\partial \alpha} lnL(x_1, x_2, x_3, ..., x_n|\alpha, \beta) = 0.$$

$$\frac{\partial}{\partial \beta} lnL(x_1, x_2, x_3, ..., x_n|\alpha, \beta) = 0.$$

O Algoritmo EM (Expectation-Maximization)

O algoritmo EM pode ser entendido como um método geral para encontrar as estimativas de máxima verossimilhança dos parâmetros de um determinado modelo de distribuição $f(x|\theta)$ para uma dada amostra de valores aleatórios e, os parâmetros que se deseja estimar (Bilmes, 1998). Como descrito anteriormente, podemos assumir que o conjunto de amostras é representado por X e a variável que se deseja estimar é representada por θ.

A execução do algoritmo EM gera seqüências de estimativas θ^s, em que $s = 1, 2, 3...n$ e representa o número de iterações de cada passo do algoritmo, dado o valor inicial θ^0. Há dois passos a cada iteração: o passo E *(Expectation)* e o passo M *(Maximization)*.

No passo E o valor da ***esperança condicional*** sobre o modelo de distribuição $\boldsymbol{lnf(x|\theta)}$ representa a distribuição condicional da variável que se deseja estimar (θ), sendo fornecido o conjunto de amostras X, assim, considerando a variável θ (que desejamos estimar), a equação que produz o valor da *esperança condicional* a cada iteração s é descrita da seguinte forma:

$$Q(\theta|\theta^s) = E_\theta \left\{ ln \left[\prod_{i=1}^{n} f(x_i|\theta) \right] |X, \theta^s \right\}$$

com $s = 1, 2, 3, ..., n$ e sendo o valor θ^s calculado no passo anterior $(s-1)$.

O passo M (maximização) encontra valores de θ que maximizam a função de distribuição $lnf(x|\theta)$. Cada iteração s tem o objetivo de maximizar a função $Q(\theta|\theta^s)$, de maneira que as novas estimativas $\theta^{(s+1)}$ produzidas no passo M, a cada iteração s, sejam usadas no passo E da iteração $(s+1)$ e, assim sucessivamente, até a convergência do modelo ser obtida (Bilmes, 1998; Woodruff and Hanson, 1997; Hanson, 1996).

No contexto dos testes adaptativos, podemos utilizar a estimativa de verossimilhança em conjunto com o algoritmo EM, tanto para estimar os parâmetros dos itens (**a** e/ou **b** e/ou **c**) como para estimar o valor da habilidade θ do aluno. Dado este fato, é importante descrever as respectivas funções de estimação do algoritmo EM para a tarefa de estimação desses parâmetros.

4.5.1.1 Estimatição dos Parâmetros dos Itens

Esta subseção apresenta o algoritmo EM para calcular as estimativas dos parâmetros dos itens, os quais compõem um banco que será utilizado em um teste adaptativo. Tais itens possuem respostas

dicotômicas, ou seja, a resposta de cada item deve ter valor 1, representando uma resposta correta, ou 0, caso seja incorreta.

Para itens dicotômicos a equação do algoritmo EM pode ser escrita em termos da Função de Resposta de um Item $P_i(\theta_j)$; que conforme visto no início deste capítulo, é a probabilidade do aluno j com nível de habilidade θ responder um item i corretamente. Dado que Δ é o conjunto de parâmetros de um determinado item i, $\Delta_i = (a_i, b_i, c_i)$, a probabilidade de um aluno responder um dado item de forma correta é dada por: $P(\theta_j|\Delta_i)$ e de forma incorreta: $Q(\theta_j|\Delta_i) = 1 - P(\theta_j|\Delta_i)$.

Assim, considerando um vetor de respostas U, sendo que $U_i = 1$ representa uma resposta correta e $U_i = 0$ uma incorreta, o modelo de distribuição dado o nível de habilidade θ e os parâmetros de um item é:

$$f(U_{ij}|\theta_j, \Delta_i) = P(\theta_j|\Delta_i)^{U_{ij}} \cdot Q(\theta_j|\Delta_i)^{1-U_{ij}}$$

$$f(U_{ij}|\theta_j, \Delta_i) = \prod_{i=1}^{I} P(\theta_j|\Delta_i)^{U_{ij}} \cdot Q(\theta_j|\Delta_i)^{1-U_{ij}}$$

e a equação do algoritmo EM para o passo E é definida como:

$$Q(\Delta|\Delta^s) = E\left\{ ln\left[\prod_{i=1}^{J} f(U_{ij}|\theta_j, \Delta_i)\right] |\theta_j, U_{ij}, \Delta^s \right\}$$

com $s = 1, 2, 3, ..., n$ representando o número de iterações.

4.5.1.2 Estimação de Habilidade θ

Existem diferentes formas de estimar a habilidade inicial de um estudante. Entre elas está uma simples que consiste em atribuir um valor aleatório entre -1.0 e 1.0 a cada indivíduo, ou ainda um valor *default* e a partir daí, o ajuste da habilidade se dá a partir das respostas obtidas durante o teste.

Outro método adotado na estimação da habilidade inicial é a Estimativa de Máxima Verossimilhança. Este método estima a habilidade a partir do fornecimento de uma amostra comum de itens a cada estudante, o qual os responde, e baseado na quantidade de respostas certas e erradas dessa amostra o valor de θ é calculado. Suponha que um determinado aluno recebeu uma amostra de cinco itens que devem ser respondidos, sendo que seu vetor de respostas foi o seguinte:

$$U = (u_1, u_2, u_3, u_4, u_5) = (10110)$$

em que um valor de $U = 1$ representa um resposta correta e o valor de $U = 0$ representa uma resposta incorreta. Dado o vetor de respostas, é possível estimar a habilidade inicial de um estudante usando

a seguinte equação:

$$\theta_0 = ln(\frac{r}{(n-r)})$$

em que r é quantidade de respostas corretas e n a quantidade de itens propostos.

No exemplo acima, o estudante acertou 3 questões ($r = 3$) de um total de 5 ($n = 5$) e sua habilidade inicial vale:

$$\theta_0 = ln(\frac{3}{(5-3)}) = 0.4054$$

e com este valor de $\theta = 0.4054$ é possível iniciar o processo de teste (Hambleton and Swaminathan, 1985).

Outra maneira de estimar o valor da habilidade inicial também pode ser empregada de acordo com o conhecimento prévio dos alunos, como utilizado no CBAT-2 (Huang, 1996). Quando o ambiente no qual a avaliação está inserida já está padronizado, com as características mais ou menos padrão, é possível por meio de informações empíricas, tais como histórico escolar, experiências vividas, currículo, etc. traçar um perfil do candidato de maneira que essa análise possa ser quantificada e representar a habilidade inicial para que o teste adaptativo seja iniciado.

O método da Estimativa de Máxima Verossimilhança também pode ser utilizado para estimar a habilidade de um indivíduo enquanto o teste está ocorrendo, ou seja, à medida em que cada estudante responde um dado item, e dependendo de sua resposta é possível estimar sua habilidade para que outro item seja selecionado e o processo de teste continue até o critério de parada ser encontrado.

Da mesma forma a probabilidade de uma resposta correta, dado um nível de habilidade θ é: $P(U_i = 1|\theta)$ e incorreta $P(U_i = 0|\theta)$. Assim, dada toda a seqüência de respostas de um candidato em um teste, o cálculo da probabilidade de respostas corretas e incorretas é dado por:

$$P(U_i|\theta) = P(U_i = 1|\theta)^{U_i} \cdot P(U_i = 0|\theta)^{1-U_i}$$

$$P(U_i|\theta) = P_i^{U_i} \cdot (1 - P_i)^{1-U_i}$$

$$P(U_i|\theta) = P_i^{U_i} \cdot Q_i^{1-U_i}$$

em que $Q_i = 1 - P_i$.

Dessa maneira, se um estudante com nível de habilidade θ responde n itens, o vetor de respostas desse estudante pode ser representado por: $U_i = (U_1, U_2, U_3, ..., U_n)$ e a probabilidade de suas respostas será: $P(U_1, U_2, U_3, ..., U_n|\theta)$. Generalizando, pode ser definido que dada a habilidade θ e as n respostas de um estudante, a função de verossimilhança (***Likelihood Function***) expressa como:

Seja

$$P(U_1, U_2, U_3, ..., U_n|\theta) = P(U_i|\theta), P(U_2|\theta), P(U_3|\theta), \ldots, P(U_n|\theta)$$

$$'' = \prod_{i=1}^{n} P(U_i|\theta)$$

$$'' = \prod_{i=1}^{n} P_i^{U_i} \cdot Q_i^{(1-U_i)}$$

$$L(u_1, u_2, u_3, \ldots, u_n|\theta) = \prod_{i=1}^{n} P_i^{U_i} \cdot Q_i^{(1-U_i)}$$

o que significa dizer que quando $u_i = 1$ o termo Q_i torna-se 0 (zero) e quando $u_i = 0$ o termo P_i assume valor 0 (zero), saindo da fórmula.

Por exemplo, supondo que o vetor de respostas de cinco itens de um determinado aluno seja o seguinte:

$$u = (10110)$$

a função de verossimilhança fica:

$$L(u|\theta) = P_1 \cdot Q_2 \cdot P_3 \cdot P_4 \cdot Q_5$$

Dessa maneira, inserindo a função crescente ln, a função de Estimativa de Máxima Verossimilhança (MLE) pode ser expressa como:

$$ln(L(u|\theta)) = lnP_1 + lnQ_2 + lnP_3 + lnP_4 + lnQ_5$$

que, generalizando, a função MLE assume a forma a seguir:

$$ln(L(u|\theta)) = \sum_{i=1}^{n} [u_i \cdot ln(P_i) + (1 - u_i) \cdot ln(Q_i)]$$

O cálculo da nova estimativa de habilidade θ, depois de um estudante responder um dado item, pode ser dada por meio da seguinte equação: $\theta_{i+1} = \theta_i - h$ na qual o valor h é dado por:

$$h = \frac{[\frac{d}{d\theta} \cdot ln(L(u|\theta))]}{[\frac{d^2}{d\theta^2} \cdot ln(L(u|\theta))]}$$

ou, em outras palavras, o valor h pode ser visto como a divisão entre a primeira derivada e segunda derivada da função de Máxima Verosimilhança.

Para estimar o nível de habilidade θ de um determinado aluno j utilizando o algoritmo EM, basta alterar a equação do passo E, descrita na subseção anterior da seguinte maneira:

$$Q(\theta|\theta^s) = E\left\{ ln\left[\prod_{i=1}^{n} f(U_{ij}|\theta_j, \Delta_i)\right] |\Delta_i, \theta^s \right\}$$

com $s = 1, 2, 3, ..., n$ representando o número de iterações.

4.5.1.3 Método Bayesiano de Estimativa de Habilidade

Quando qualquer informação prévia sobre a distribuição de habilidade dos estudantes é disponível antes do início do teste, o Método Bayesiano de estimação de habilidade pode fornecer resultados significativos (Hambleton and Swaminathan, 1985). Dado que a habilidade de um aluno é representada por θ_a, onde a representa cada aluno variando de *1, 2, ... N*, é possível considerar, baseado nessas informações prévias, que o valor da habilidade seja um valor aleatório dentro de um intervalo que deve ser especificado. Por exemplo, os projetistas do teste acreditam que uma pequena porção de estudantes tenha habilidades fora de um intervalo, então a maioria dos estudantes tem níveis de habilidade em torno de uma média. Assim, pode ser indicado que a habilidade obedece a uma distribuição normal dada por:

$$\theta_a \sim N(\mu, \phi)$$

sendo que $N(\mu, \phi)$ denota uma distribuição normal com média μ e variância ϕ ,de maneira que seus valores precisam ser especificados. Segundo Owen (1975), no contexto dos testes adaptativos esses valores podem ser $\mu = 0$ e $\phi = 1$.

Assume-se que a distribuição de habilidade de θ_a pode ser a função logística $f(\theta)$:

$$f(\theta) = \frac{\exp(\theta)}{[1 + \exp(\theta)]^2}$$

O teorema de Bayes é o principal componente do Método Bayesiano de estimação de parâmetros e relata as probabilidades condicionais:

$$P(B|A) = P(A|B) \cdot P(B)/P(A)$$

No contexto da estimação de habilidade, A pode ser considerada como θ_a e B como um conjunto de respostas observadas em n itens e é denominado como u. Então a equação das probabilidades condicionais pode ser transformada em:

$$P(\theta_a|u) = P(u|\theta_a)P(\theta_a)/P(u)$$

4.5.2 O Programa XCALIBRE

A presente subseção analisa as principais características do programa XCALIBRE Versão 1.0 para Windows que é utilizado na tarefa de calibração dos itens durante processo de construção do banco. Esse programa, além de possuir uma versão de demonstração grátis, implementa os modelos de 2 e 3 parâmetros da TRI, o que justifica sua escolha para o processo de calibração dos itens. A descrição detalhada das operações de estimação deste programa se dá pelo fato de que há a intenção de que no futuro próximo mais itens sejam estimados usando o XCALIBRE, com a intenção de incrementar o banco de itens.

O processo de adaptação ocorrido durante a execução de um teste adaptativo ocorre graças à capacidade que os modelos da Teoria de Resposta de Itens (TRI) possuem de se adequarem a cada nível de habilidade do estudante, dadas as variáveis psicométricas (a, b e c) que cada item possui. Entretanto, a tarefa de especificar os parâmetros dos itens que permitem tal adaptação é custosa, pois é preciso aplicar funções estatísticas complexas, pertencentes à TRI.

A maneira mais comum de estimar os parâmetros dos itens de um banco é organizar os itens em diferentes grupos (como se fosse um teste à parte) e aplicar esses grupos de itens a um determinado conjunto de alunos. Com as respostas obtidas dessa aplicação é preciso elaborar amostras contendo informações dos itens e suas respectivas respostas fornecidas. A tarefa de estimação dos parâmetros consiste em analisar tais amostras por meio de cálculos estatísticos de maneira a determinar os valores dos parâmetros de cada item administrado.

O programa XCALIBRE engloba a análise de testes e de itens, fazendo parte do pacote **Micro-CAT Testing System** disponibilizado pela empresa *Assessment System Corporation*. O XCALIBRE utiliza os modelos de dois e três parâmetros da TRI para estimar as variáveis de Discriminação (**a**), Dificuldade (**b**) e Adivinhação (**c**) associadas a cada item que compõe um banco, usando para isso as amostras de item-resposta.

O processo de estimação dos parâmetros dos itens é complexo por envolver cálculos matemáticos avançados e deve ser tratado por programas especialmente projetados para esta tarefa. Segundo o manual (XcalibreManual, 1997) as técnicas de estimação de parâmetros têm experimentado significativos progressos nos últimos anos, e provavelmente a Estimativa de Máxima Verossimilhança (EMV) é a opção mais avançada. O XCALIBRE implementa o algoritmo EM (*Expectation-Maximization*) em conjunto com a EMV para estimar os parâmetros.

4.5.2.1 Formatação dos Dados de Entrada

Os dados de entrada (resposta dos itens) do programa XCALIBRE devem estar contidos em um arquivo no formato ASCII (somente-texto), que deve ser elaborado seguindo o formato padrão do programa em qualquer editor de texto. Todos os dados a serem incluídos na análise devem estar contidos em um único arquivo, sendo que o programa permite a inclusão de mais de 750 itens (colunas) e o número de alunos (linhas) ilimitado para realizar a análise. Entretanto, se o conjunto de dados for muito grande o processo de estimação pode levar muito tempo para ser realizado,

devido aos cálculos complexos que devem ser efetuados (XcalibreManual, 1997). Caso exista uma amostra de dados extremamente grande, é necessário criar "pequenas amostras" armazenadas em arquivos menores para que a análise seja realizada.

A Figura 4.10 Mostra um exemplo de um arquivo de entrada do XCALIBRE.

```
 10 0 N  4
1435342435  RESPOSTAS
5555555555  NRO. DE ALTERNATIVAS
YYYYYYYYYY  ITENS A INCLUIR
A001 1543542143  ALUNO 1
A002 1435342445  ALUNO 2
A003 1435342235  ALUNO 3
A004 3514342453  ALUNO 4
A005 5135424431  ALUNO 5
...
...
```

Figura 4.10: Um exemplo do formato do arquivo de entrada dos dados do programa XCALIBRE

O arquivo de entrada de dados consiste de cinco componentes, apresentados na ordem mostrada abaixo:

1. Uma linha de controle que descreve os dados (primeira linha);
2. Uma linha contendo a resposta correta dos itens (segunda linha);
3. Uma linha com o número de alternativas de cada item (terceira linha);
4. Uma linha que especifica quais itens entrarão na análise (quarta linha);
5. Os dados propriamente ditos que serão base para a análise (restante das linhas).

os outros dados são comentários e podem ser incluídos.

A primeira linha do arquivo de dados, linha de controle, contém informações que especificam os parâmetros que ditam a análise a ser realizada. Os três primeiros caracteres devem conter o número de itens que serão incluídos na análise e devem ser justificados à direita. No exemplo da Figura 4.10 o número de itens é dez. O quarto caractere deve ser um espaço em branco que serve para separar do quinto caractere que representa um código alfanumérico utilizado para respostas que foram omitidas (não respondidas) pelo aluno. No exemplo esse caractere é descrito pela letra O. O sexto caractere também deve ser um espaço em branco e o sétimo representa um código alfanumérico utilizado para os itens que não foram alcançados pelo aluno durante um teste (letra N). Finalmente, o oitavo caractere é mais um espaço em branco e os caracteres de nove a dez representam o número de caracteres que identificam um determinado aluno no conjunto de dados. No exemplo esse número é quatro e define, por exemplo, o aluno A001.

A Tabela 4.3 mostra a descrição de cada caractere da linha de controle.

Tabela 4.3: Descrição dos caracteres da linha de controle do arquivo de entrada de dados

Coluna	Representação dos Dados
1-3	Número de itens que serão incluídos na análise
4	Branco
5	Código alfanumérico que representa a omissão de respostas
6	Branco
7	Código alfanumérico que representa um item não alcançado pelo aluno
8	Branco
9-10	Número de caracteres que identifica um determinado aluno

A segunda linha contém a resposta correta (chave) de cada item incluído na análise. Cada resposta é representada por um caractere, e todas as respostas devem estar contidas em somente uma linha. Assim, seguindo o exemplo da Figura 4.10, a resposta correta do item 1 é "1", do item 2 é "4", do 3 é "3" e assim sucessivamente até o décimo item cuja resposta é "5". Cada caractere que identifica uma resposta representa a alternativa correta de um determinado item, e dessa forma pode ser representada pelos caracteres de "1" a "9" ou pelas letras de "A" a "I". Por conveniência, é definido que uma resposta codificada como "1" é igual a "A" e "a", assim como "2" = "B" = "b", até "9" = "I" = "i". Não existe equivalência para o caractere 0, e todas as respostas codificadas de maneira diferente desses caracteres serão consideradas como incorretas.

A terceira linha deve especificar o número de alternativas de resposta que cada item possui. Este número deve ser igual ao número de alternativas fornecidas para o item. No exemplo fornecido, o número de alternativas é cinco, ou seja, as opções de resposta de um item durante a execução de um teste variam de "1" a "5" ou de "A" a "E". Caso o modelo da TRI selecionado seja o de 3 parâmetros, o programa XCALIBRE usa o número de alternativas do item para calcular o parâmetro de Adivinhação (c) no início do processo de estimação.

A quarta linha do arquivo de entrada contém o código de inclusão de cada item na análise. O código de inclusão pode ser o caractere "Y" (yes) ou "N" (no) e indica se um determinado item deve ou não ser considerado no processo de estimação. No exemplo da Figura 4.10 todos os itens foram incluídos na análise.

Os dados de análise propriamente ditos utilizados para a estimação dos parâmetros são representados a partir da quinta linha em diante, até o final do arquivo. O conjunto de respostas aos itens de cada aluno deve ser organizado em uma única linha, de maneira a representar apenas um aluno, começando com os caracteres que identificam um aluno, por exemplo, A001 respeitando para isso a quantidade de caracteres definida na linha de controle (primeira linha). Qualquer caractere que representa uma resposta do aluno, ou o código de omissão, ou código de item não alcançado pode estar presente no conjunto de dados que representa as respostas dos alunos.

4.5.2.2 Descrição das Interfaces do Programa XCALIBRE

Em linhas gerais, o programa XCALIBRE é de simples instalação e uso, fornecendo interface gráfica amigável e bem organizada, facilitando as atividades do usuário. Quando o XCALIBRE inicia sua execução a primeira tela apresentada contém um menu com cinco opções. O menu ***File*** abriga os comandos de imprimir (*Print*) e sair (*Exit*); o menu ***Edit***, permite o acesso ao editor de textos ASCII que acompanha o XCALIBRE para a edição dos arquivos de entrada e visualização dos resultados da análise. Sob o menu ***Configure***, com a ação do comando *Go* é apresentada a tela de opções de configuração do processo de estimação, também sendo possível pela ação associada ao botão representado por um *lápis* logo abaixo das opções de menu. O XCALIBRE inicia o processo de análise dos dados por meio da opção *Go* sob menu ***Analyze***, o que também é possível pelo botão representado por um *pequeno computador*. Finalmente, o menu ***Help*** libera o acesso aos arquivos de ajuda do programa. A ação de sair do programa (*Exit*) também é executada pelo botão representado pela figura de uma *porta*. A Figura 4.11 apresenta a primeira tela do programa.

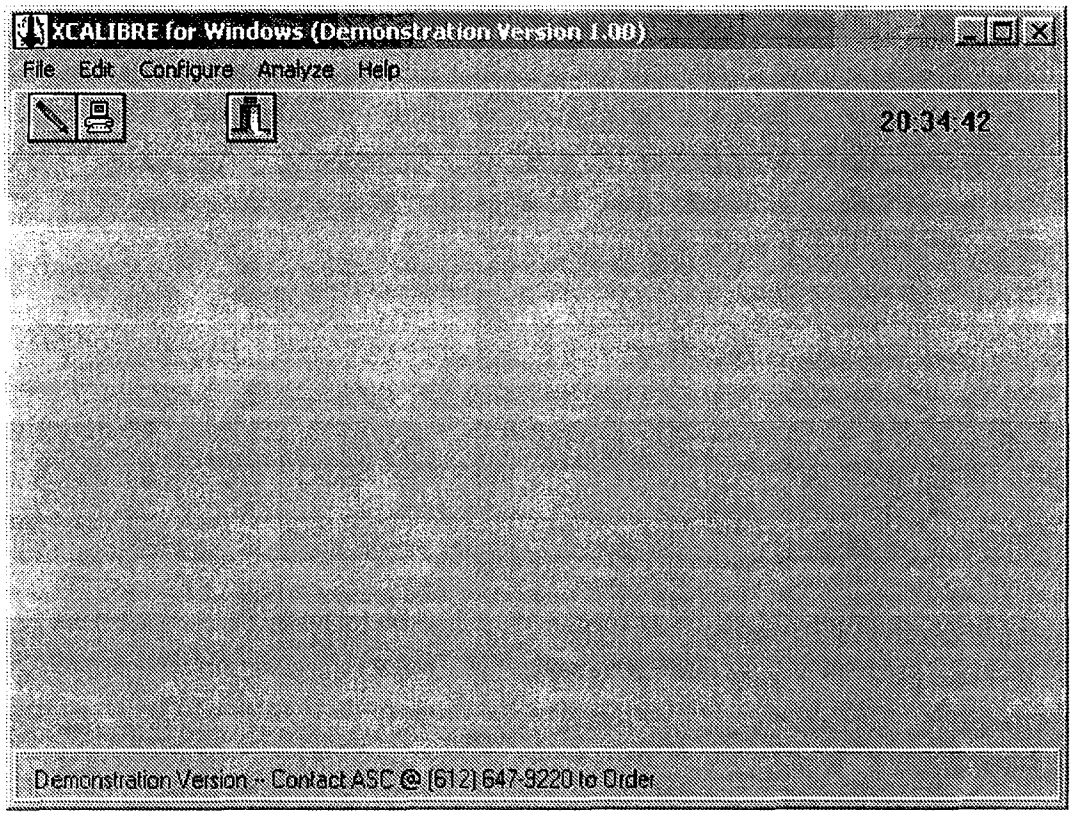

Figura 4.11: Primeira tela do programa XCALIBRE

Quando o usuário aciona o comando *Go* do menu *Analyze*, (ou clica no botão *lápis*) uma tela contendo duas pastas (*folders*) são apresentadas. A primeira pasta, chamada de *Files*, oferece as opções de seleção dos arquivos que serão utilizados na análise. São fornecidos dois campos do tipo

Edit onde o usuário pode editar o nome do arquivo de entrada (que contém as amostras) e outro para a edição do arquivo de saída (que conterá os resultados da estimação). Além dos campos *Edit*, ainda há três conjuntos de *Radio Buttons* que permitem ao usuário criar arquivos adicionais que contenham: a pontuação dos alunos, os resultados das estatísticas e a identificação de cada item pertencente à análise. Caso o usuário selecione a opção YES em qualquer um dos *Radio Buttons*, um botão representado por um *Folder* aparece automaticamente, permitindo a seleção e criação do arquivo. A Figura 4.12 mostra a tela que representa a pasta *Files*.

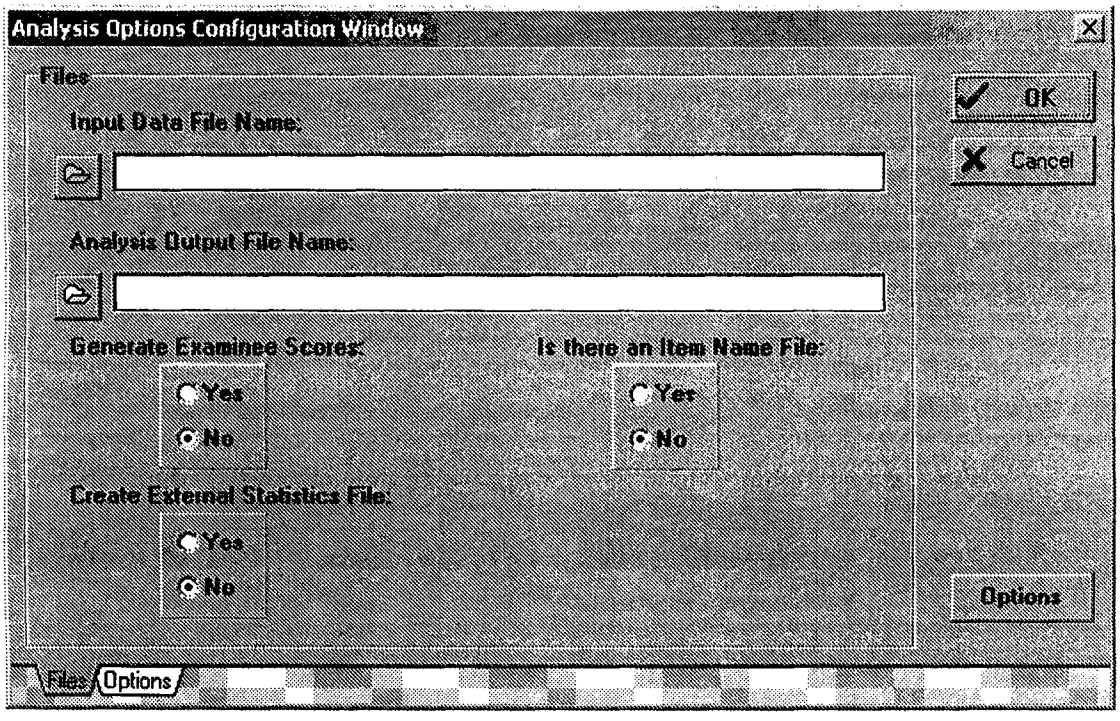

Figura 4.12: Janela da pasta *Files*

A segunda pasta, chamada *Options*, permite ao usuário especificar certas opções que regem o processo de estimação dos parâmetros e conseqüentemente a execução do programa. Dentre as opções oferecidas estão:

IRT Model: permite a seleção do modelo da TRI (2 ou 3 parâmetros) que serve de base para o processo de estimação dos parâmetros;

Priors Distributions: esta opção permite a especificação do tipo de distribuição (média e desvio padrão) que será imposta ao processo de estimação dos parâmetros durante o processo de análise. Há três tipos de distribuição pré-estabelecidos: *Default*, *Common* e *Separate*. Caso o usuário escolha o tipo *Separate* é necessário a inclusão de um arquivo com as especificações;

Maximum Loops: esta opção permite que o usuário defina o número máximo de iterações (ciclos EM) que o programa irá executar durante a estimação. De acordo com o manual a maioria

dos cálculos das estimações convergem em menos de 15 ciclos. Entretanto, essa opção aceita até 100 ciclos;

Floating Priors: permite que a média definida na opção *Priors Distributions* seja atualizada a cada ciclo EM, de maneira que o valor da média navegue, ou seja possivelmente trocado durante a fase de estimação;

Prior Distributions Moments: esta opção mostra ao usuário os valores da média e desvio padrão de acordo com a distribuição selecionada (*Common* ou *Default*), permitindo ainda que os mesmos sejam alterados. Caso a opção tenha sido *Separate* os respectivos valores não são mostrados.

A Figura 4.13 mostra a tela com a pasta *Options*.

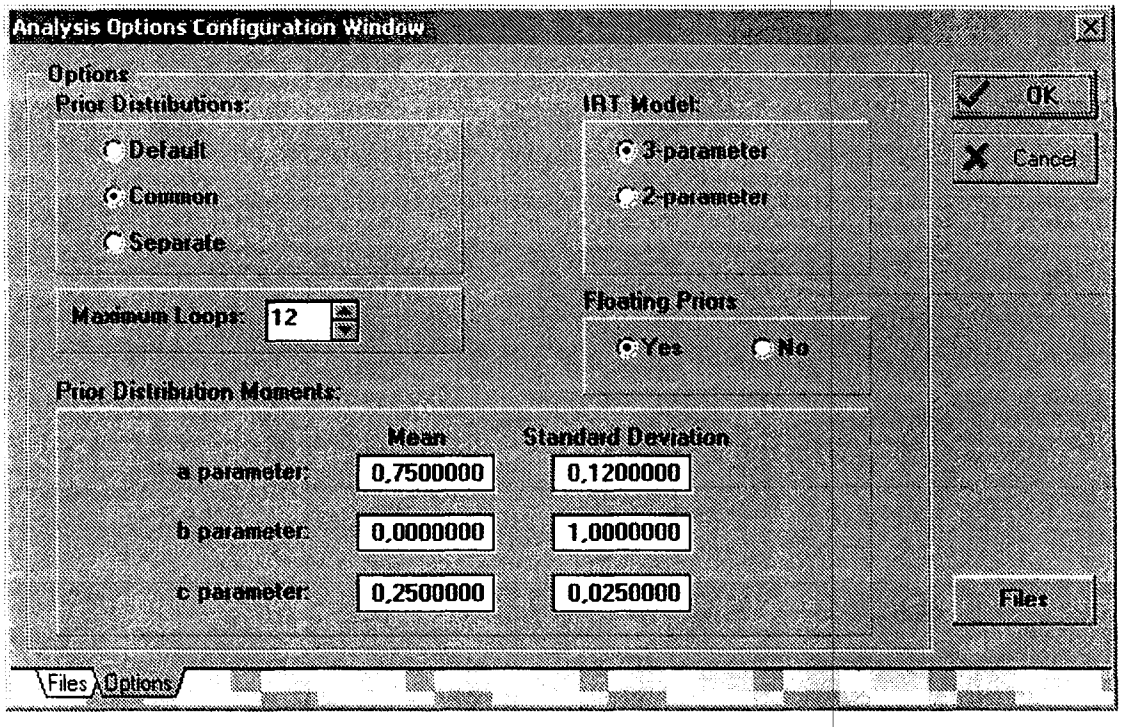

Figura 4.13: Janela da pasta *Options*

Uma vez que foram selecionados os arquivos de entrada e saída, e as opções de análise tenham sido definidas, é preciso iniciar o processo de estimação dos parâmetros propriamente dito. Este processo é iniciado com o comando *Go* do menu ***Analyze*** ou através do clique no botão representado pelo pequeno computador (Figura 4.11) da primeira tela. O início do processo de estimação é marcado com a abertura de uma pequena janela contendo uma barra de porcentagem que indica o andamento da fase de estimação, ficando explícito o progresso de cada ciclo executado.

Quando a tarefa de estimação está completa, é apresentada uma janela contendo três botões. O primeiro botão tem a função de abrir o editor do XCALIBRE com arquivo de resultados (especificado

na pasta *Files*) e apresentá-lo ao usuário. O arquivo de resultados nada mais é do que um relatório completo, contendo todas as informações da estimação realizada, bem como os resultados (parâmetros) da estimação. O segundo botão, quando acionado, imprime o arquivo de resultados na impressora padrão instalada. O terceiro botão sai do XCALIBRE.

4.5.2.3 Interpretação dos Resultados da Calibração

O relatório produzido pelo programa XCALIBRE contendo todos os resultados da estimação dos parâmetros é armazenado em um arquivo com extensão ".OUT" (ponto out). As primeiras duas páginas do relatório (arquivo) contém as informações fornecidas pelo usuário na fase de especificação, incluindo os nomes dos arquivos de entrada e saída e as opções de análise selecionadas. Estas informações podem ser muitos úteis no futuro quando outra análise for executada sobre os mesmos dados.

Seguindo as informações do relatório são fornecidos detalhes específicos sobre a amostra dos dados e o progresso passo-a-passo da análise realizada. O progresso da estimação dos parâmetros é uma lista de valores que registra a alteração máxima ocorrida entre um ciclo e outro das iterações do algoritmo EM. Esse valor é calculado pela diferença das estimações parciais de cada ciclo determinadas por: a_n, b_n e c_n (passo n) e a_{n-1}, b_{n-1} e c_{n-1} (passo $n-1$). Assim, o cálculo da alteração máxima (AM) no passo n é dado por:

$$AM_n = |(a_n - a_{n-1})| + |(b_n - b_{n-1})| + |(c_n - c_{n-1})|$$

A média e o desvio padrão geral de cada parâmetro obtidos durante o processo de estimação também são apresentados no relatório. Estes valores são determinados empiricamente pela estimação final dos parâmetros.

As últimas informações contidas no relatório final são os próprios valores dos parâmetros estimados (a,b e/ou c) e dois gráficos representando, respectivamente, a função de informação e a função de resposta referentes ao teste.

A Figura 4.14 apresenta um exemplo dos valores finais da estimação dos parâmetros e as respectivas variáveis estatísticas que representam suas características.

Conforme visto da Figura 4.14 os valores finais da estimação dos parâmetros são agrupados em onze colunas, cada uma delas representando uma determinada característica de cada item que teve seus parâmetros estimados. A seguir segue uma descrição de cada uma delas.

Item: esta coluna mostra posição seqüencial de cada item disposto no arquivo de entrada;

Lnk: a coluna *Lnk* denota se um item é de ligação. Este tipo de item já tem os valores de seus parâmetros fixos especificados em um arquivo com o nome FIXITEMS.DAT para propostas de análise de atualização dos parâmetros;

Flg: a terceira coluna representa vários *flags* referentes ao item. Estes *flags* são:

```
                XCALIBRE (tm) for Windows95/NT -- Version 1.10
  Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
         Marginal Maximum-Likelihood IRT Parameter Estimation Program

XCALIBRE Analysis from Data File: D:\XCALIBRE\W95-110c\Sample1.dat
Date: Jul 14, 1997                                              Time:  3:56 PM

ITEM PARAMETER ESTIMATES W/STANDARD ERRORS
                          a              b              c
Item Lnk Flg    a      error     b      error     c      error   Resid   Item name
---- --- --- ------  -------  ------  -------  ------  -------  -------  ----------

  1            0.67    0.120   -1.09    0.137    0.25     ***     0.49
  2            0.98    0.129   -0.43    0.096    0.23     ***     0.54
  3            1.11    0.119   -2.14    0.150    0.25     ***     0.86
  4            0.75    0.103   -2.05    0.161    0.25     ***     0.24
  5       P    0.71    0.106   -3.00    0.239    0.25     ***     1.06
  6            0.84    0.133   -0.46    0.109    0.25     ***     0.37
  7            0.55    0.097   -2.84    0.230    0.25     ***     1.28
  8            0.59    0.108   -1.69    0.164    0.25     ***     0.92
  9            0.84    0.109   -1.41    0.125    0.25     ***     0.39
 10            0.86    0.144   -0.14    0.10E    0.25     ***     0.45
  .             .       .       .        .        .        .       .
  .             .       .       .        .        .        .       .
  .             .       .       .        .        .        .       .
```

Figura 4.14: Exemplo de valores da estimação final dos parâmetros

- **P** – denota um item potencialmente problemático. Itens problemáticos possuem os valores dos parâmetros:
  - $a < 0,30$
  - $b > 2,95$
  - $b < -2,95$
  - $c > 0,40$
- **K** – indica um possível erro de resposta, significando que uma alternativa incorreta tem alta correlação com a pontuação total;
- **R** – indica que o valor residual padrão do modelo de ajuste excedeu o valor de 2,0.

a,b,c: estas três colunas contém os valores das estimativas finais dos parâmetros a,b e c correspondentes ao modelo da TRI (2 ou 3 parâmetros) selecionado. Se o modelo de 2 parâmetros foi selecionado o valor do parâmetro c será 0 (zero);

Errors: são três colunas que indicam o erro padrão associado a cada estimativa dos parâmetros;

Resid: esta coluna contém o valor residual padrão que indica a proporção que um determinado item se ajusta ao modelo da TRI selecionado. Valores próximos de 0 (zero) significam um bom ajuste ao modelo;

Item: esta coluna lista o nome (descritor) do item, caso tenha sido fornecido um arquivo com os nomes dos mesmos no momento da entrada de dados. Esta descrição pode conter até 80 caracteres.

Na última parte do relatório são traçados dois gráficos, representando a função de informação e a função de resposta do teste. Tais gráficos são elaborados de acordo com o somatório das funções de informação e de resposta que cada item pertencente ao conjunto de amostra fornece dado seus parâmetros estimados. Entretanto, os gráficos destas funções não são muito úteis, já que a avaliação do processo de estimação realizada sobre o relatório final diz respeito a cada item individualmente, e não ao teste como um todo. Além disso, o esboço das funções não é apresentado por uma linha contínua e os gráficos são mal desenhados, caracterizando uma desvantagem do XCALIBRE e não fornecendo detalhes suficientes para uma análise confiável. A Figura 4.15 mostra o gráfico da função de informação de um teste. Apesar disto, o programa fornece dois valores importantes para a análise dos itens:

Expected Information: representa a quantidade média de informação que um teste formado com os itens calibrados forneceria para uma amostra de alunos com distribuição normal de habilidade;

Average Information: representa a quantidade média de informação que um teste formado com os itens calibrados forneceria para uma amostra de alunos retangularmente distribuídos.

Dado o fato que o processo de seleção dos itens se dá a partir da quantidade de informação fornecida por um item durante um teste dado o nível de habilidade do aluno, tais valores são de relativa importância, já que com eles é possível conhecer a quantidade média de informação fornecida pelos itens conforme a distribuição de habilidade dos alunos.

4.6 Procedimentos de Pontuação

Vistos os principais métodos de estimação de habilidade, é necessário estabelecer critérios que estabelecem a verdadeira pontuação obtida por um estudante em um teste. Os procedimentos de pontuação, então, têm a responsabilidade de converter a habilidade estimada durante todo teste em pontos convencionais, que permitem classificar o estudante após a avaliação.

As diferenças entre a escala de distribuição de habilidade e a verdadeira pontuação do estudante em um determinado teste, dependem, entre outras coisas, do nível de dificuldade que cada teste administrado representou, de acordo com a habilidade desenvolvida pelo indivíduo. Dessa forma, não é possível realizar uma conversão direta, sendo necessário considerar o nível de dificuldade de itens fornecidos ao estudante e as variações de habilidade (quando acontece um acerto ou erro) ocorridas durante um teste. Por exemplo, se alguém aplicasse dois testes medindo a mesma habilidade para um mesmo grupo de estudantes, não sendo necessariamente paralelos, duas diferentes habilidades, e consequentemente, suas pontuações seriam distintas, pois os itens disponibilizados,

XCALIBRE Analysis from Data File: D:\XCALIBRE\W95-110c\Sample1.dat
Date: Jul 14, 1997 Time: 3:56 PM

Test characteristics:	K-R 21 Reliability	Expected Information	Average Information
	0.790	3.769	3.312

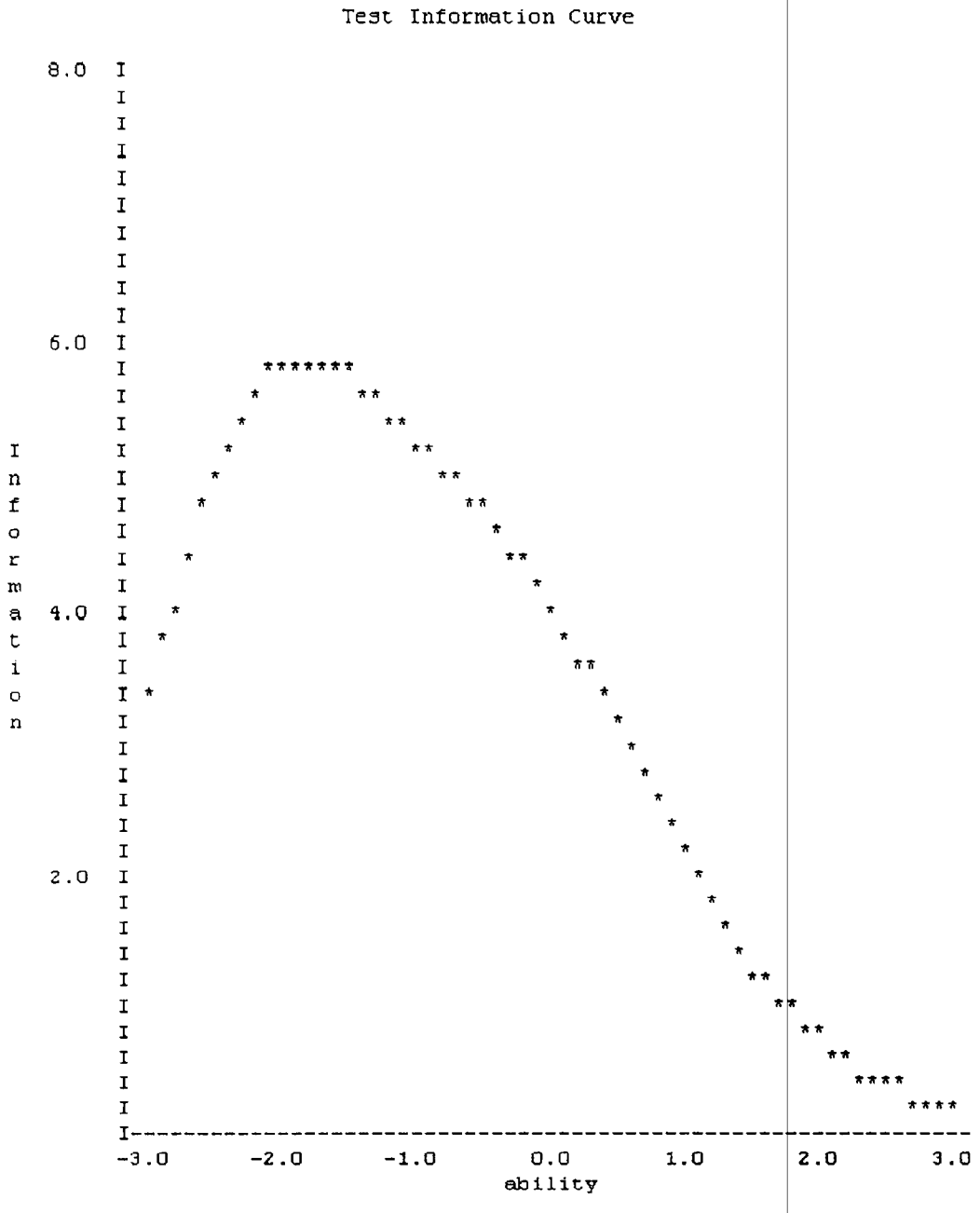

Figura 4.15: Gráfico da função de informação de um teste

os níveis de dificuldade e as habilidades estimadas durante o teste foram diferentes. Ou seja, é preciso considerar os itens fornecidos, bem como seus parâmetros, e as habilidades calculadas, para encontrar de maneira justa a pontuação final do aluno.

A pontuação verdadeira possui uma interessante relação com a habilidade θ de um indivíduo. Segundo Hambleton and Swaminathan, (1985), a pontuação observada (obtida durante um teste) de um estudante é definida como r, baseada no somatório dos n itens apresentados em um teste. Então:

$$r = \sum_{i=1}^{n} U_i$$

onde U_i representa a resposta de um item i e pode ser 1 ou 0 (zero).

Em linhas gerais, a medida de pontuação verdadeira é dada pelo somatório dos itens que foram respondidos corretamente durante um teste, dividido pelo total de itens, e pode ser representada pela equação:

$$\pi = \frac{1}{n} \cdot \sum_{i=1}^{n} U_i$$

Entretanto, esse método não considera, tanto os parâmetros dos itens disponibilizados no teste, quanto as habilidades individuais de cada indivíduo, já que apenas as questões certas serão somadas. Então, é necessário que a equação acima considere as características dos itens fornecidos para um estudante com nível de habilidade θ. Assim, a equação torna-se:

$$\pi|\theta = \frac{1}{n} \cdot \sum_{i=1}^{n} (U_i|\theta)$$

e visto que Ui é uma variável que assume valores 1 ou 0, segue:

$$(U_i|\theta) = (U_i = 1)P_i(U_i = 1|\theta) + (U_i = 0)P_i(U_i = 0|\theta)$$

$$(U_i|\theta) = 1 \cdot P_i(U_i = 1|\theta) + 0 \cdot P_i(U_i = 0|\theta)$$

$$(U_i|\theta) = P_i(U_i = 1|\theta)$$

$$(U_i|\theta) = P_i(\theta)$$

que é a Função de Resposta de um Item dado o modelo de resposta adotado. Então, a função de conversão do nível de habilidade de um estudante ao final de uma avaliação para a sua pontuação

verdadeira é definida da seguinte forma:

$$\pi|\theta = \pi = \frac{1}{n} \cdot \sum_{i=1}^{n} P_i(\theta)$$

Segundo Hambleton and Swaminathan, (1985), esta função é conhecida como ***Função Característica de um Teste*** (*Test Characteristic Function*) e sua curva no gráfico bidimensional é referida como ***Curva Característica de um Teste*** (*Test Characteristic Curve*) e é utilizada no teste adaptativo CBAT-2 (Huang, 1996) e no sistema SIETTE (Rios et al., 1998) para o cálculo da pontuação verdadeira.

Vistas as principais características, aplicações e funcionalidades da Teoria de Resposta de Itens (TRI) aplicada aos testes adaptativos, o capítulo seguinte observa os principais aspectos envolvendo os programas que empregam os princípios dos testes adaptativos. Várias características são estudadas, entre elas a construção dos testes, a administração e a análise dos resultados.

CAPÍTULO 5

Programas para a construção e administração dos testes informatizados

O objetivo do presente capítulo é descrever os tipos de programas disponíveis, bem como as principais características envolvidas, destinados à construção e administração de testes com o uso do computador.

Na atualidade, a quantidade e variedade de programas existentes, e a rapidez com que surgem novas versões e novos programas é enorme. Segundo Olea et al. (1999), os primeiros programas destinados à avaliação informatizada foram elaborados por professores e educadores em universidades e instituições escolares, e eram implementados para uso em grandes computadores existentes na época. A generalização do uso dos computadores pessoais provocou a criação de novos programas de avaliação, e no final dos anos oitenta e início dos anos noventa a quantidade desses programas já era considerável. No entanto, os programas desenvolvidos nessa época já estão obsoletos, devido principalmente às capacidades gráficas, ao advento da Internet e à capacidade de agregar funções multimídia pelos atuais computadores. Assim, tais programas vêm sendo substituídos por outras versões que englobam novos avanços tecnológicos.

Contudo, o crescimento exponencial da Internet tem proporcionado o aparecimento de diversas empresas provedoras de testes informatizados e entre eles os testes adaptativos. Tais empresas têm como principal função o desenvolvimento de programas e/ou sistemas que contemplam todas as atividades inerentes a uma avaliação automática. No campo dos testes adaptativos, como principais companhias provedoras deste tipo de teste, podemos citar: a ***CAT Software System*** (*www.catinc.com*) que fornece o pacote *CATGlobal*, que contém vários programas, desde a construção dos testes (*CATBuilder*) até a administração adaptativa (*CATAdministrator*); a ***Assessement System Corporation*** (*www.assess.com*) que fornece o *FastTest Professional* um dos mais tradicionais pacotes de programas que implementa a os testes adaptativos, a ***Ericae.Net*** (*www.ericae.net*), e a ***Prometric Testing Center*** (*www.prometric.com*) é responsável pelo desenvolvimento do

Prometric que implementa os testes adaptativos na IEEE para a avaliação dos profissionais de computação.

A maior desvantagem desses sistemas é que as versões disponibilizadas atualmente são apenas para teste, (chamadas versões *Demo*), na grande maioria deles com duração de apenas um mês, tendo, que a partir daí, pagar direitos de uso às empresas proprietárias, ou seja, existe uma carência grande de programas *Freeware*. Um dos únicos programas que foge a essa regra é o ***AdTest***, desenvolvido pela Universidade Autônoma de Madri, porém é muito simples, com versões apenas para o sistema operacional DOS, com interface texto, não fornecendo flexibilidade para escolha antecipada dos métodos para seleção de itens a estimação de proficiência de um teste adaptativo.

5.1 Características e Tipos de Programas

Os programas especializados em construção e administração de testes oferecem procedimentos padronizados para tornar possível as atividades destes processos. Todos estes programas podem ser compostos por módulos que realizam atividades relacionadas e que atuam entre si de forma coordenada. Geralmente, essas atividades se agrupam em módulos funcionais responsáveis pelas tarefas de **Construção, Administração** e **Análise**.

Construção - compreende a elaboração e a organização dos itens no banco e a distribuição destes em um determinado teste mediante diferentes procedimentos de seleção. Mesmo que alguns pacotes de programas apresentam esta atividade inserida em seu contexto como sendo uma função, também existem alguns programas, como o *CATBuilder*, cujo procedimentos de geração de itens e construção de testes são específicos.

Administração - se referem aos programas que realizam a apresentação dos itens em um teste, bem como a coleta das respostas e os cálculos das pontuações. Neste sentido, os programas diferem quanto aos tipos de testes (tradicionais ou adaptativos) e os modos pelos quais a administração pode ocorrer. Normalmente, os programas podem fazer combinações desses aspectos e trazerem em um só pacote as diversas aplicações de testes informatizados, entretanto, os de mais destaque, como o *AdTest*, priorizam apenas um tipo de abordagem de teste. Dessa forma, podemos centrar os programas que realizam somente os testes tradicionais informatizados ou testes adaptativos.

Análise - são responsáveis pelo cálculo dos índices estatísticos e informações sobre o teste, bem como a elaboração de relatórios a respeito dos resultados da avaliação como um todo. Geralmente, aparecem incluídos em um pacote de programas como sendo apenas uma função específica.

Neste trabalho será feita a análise superficial de apenas dois programas: o ***AdTest***, um programa de caráter experimental que implementa a administração e análise dos testes adaptativos, elaborado pela Universidade Autônoma de Madri; e o *FastTest Professional* da empresa Assessement System

Corporation um pacote de programas que contém os três módulos funcionais acima citados. Também, na última parte deste capítulo, será realizada uma comparação dos resultados de teste um hipotético, dadas as abordagens empregadas pelos dois programas acima citados. Para uma análise mais detalhada de vários programas e sistemas voltados à avaliação informatizada verificar Olea et al., (1999).

5.2 Programas Analisados

5.2.1 O AdTest - Administração e Análise

O *AdTest* (Ponsoda et al., 1994) é um programa que emprega a estratégia dos testes adaptativos, e foi elaborado por pesquisadores da Universidade Autônoma de Madrid. O AdTest foi criado para atender fins docentes e de pesquisa servindo apenas para avaliar e comprovar os diferentes modelos de seleção de itens e estimação de parâmetros. Suas versões disponíveis não são comerciais e podem apenas ser obtidas em contato como os autores

O AdTest visto de uma maneira técnica é um programa baseado no sistema operacional DOS que utiliza um banco de itens de múltipla escolha em formato texto, calibrados com o modelo logístico de três parâmetros. O algoritmo adaptativo empregado escolhe o primeiro item aleatoriamente, sendo este de dificuldade média; a habilidade inicial é estimada como um valor aleatório entre -1.0 e 1.0 (que representa a maioria dos valores iniciais de habilidade dos estudantes), o modelo de seleção de itens é o de máxima informação, a estimação da habilidade segue o método da máxima verosimilhança (MLE), e o critério de parada é dado pela administração de um determinado número de itens. Segundo Olea et al. (1999), algumas versões deste programa incorporam outras opções de estimação de habilidade e a seleção de itens, como métodos bayesianos e randômicos.

O programa analisado nesta subseção contém um banco de trinta e quatro itens de múltipla escolha com seis opções cada um, sendo apenas uma delas correta. As questões são sobre o domínio de História Geral. A tela inicial do programa solicita um "nick" de três letras iniciais do nome do indivíduo que serve para identificá-lo durante o processo de teste (Figura 5.1). Os itens são fornecidos de acordo com a estratégia adaptativa empregada, e o aluno tem um tempo (parâmetro do sistema) para responder a questão sob pena de perder a chance de respondê-la. Se este parâmetro for zero o tempo não é considerado. A Figura 5.2 mostra a tela em que é apresentada uma questão ao estudante. Note que no canto inferior direito há um contador de tempo, e ao seu lado a opção de resposta do aluno.

A versão analisada deste programa fornece cinco itens a cada estudante, sendo que no final é disponibilizado um relatório do resultado do teste, contendo a habilidade alcançada (*score*), os itens respondidos correta e incorretamente, os parâmetros (a, b e c) de cada item administrado, o valor do desvio padrão associado a cada estimação de habilidade e um gráfico que mostra a habilidade atual alcançada numa distribuição normal de habilidade. As Figuras 5.3, 5.4 e 5.5 mostram as telas do relatório de um teste hipotético.

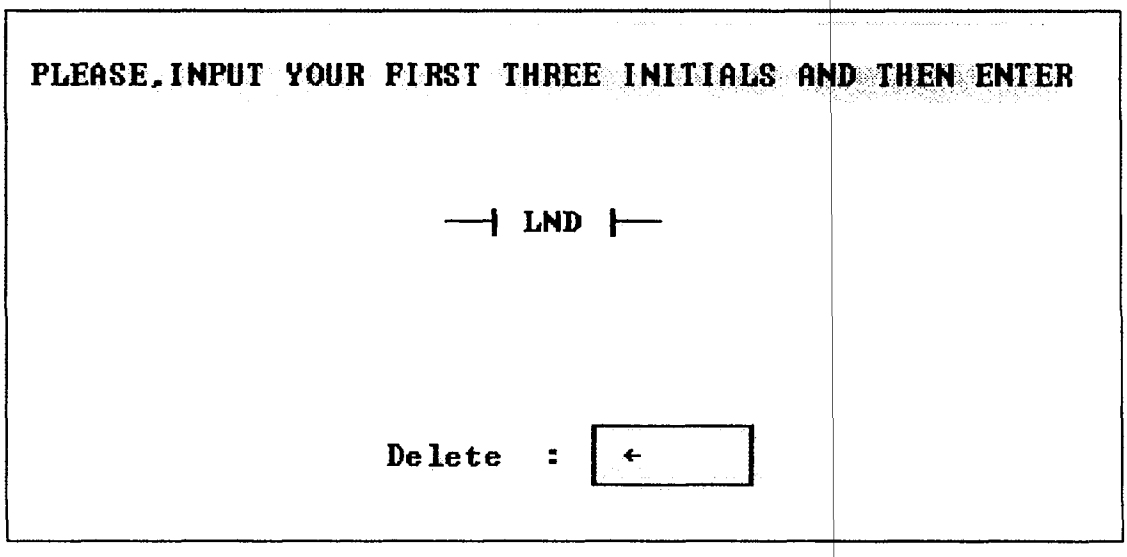

Figura 5.1: Tela de pedido de identificação com as três iniciais

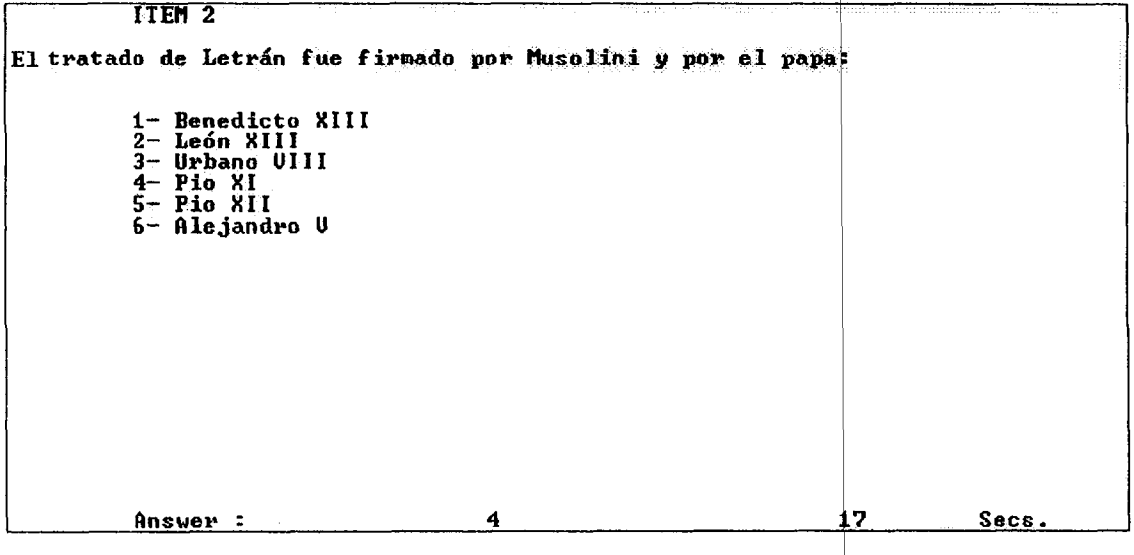

Figura 5.2: Tela que mostra uma questão a um estudante, com a opção 4 e o tempo restante

A Figura 5.3 mostra a primeira tela do relatório dispondo os resultados do teste. Contém a identificação do candidato, a pontuação (*Score*) alcançada, a medida de erro padrão do teste (*Standard error*), os itens administrados e o tempo gasto em todo teste. Além destas informações, do lado direito existe um *menu* de opções em que o aluno, dependendo da sua escolha, pode repetir outro teste, obter mais informações sobre o teste corrente, finalizar, ou salvar os resultados.

Caso o aluno selecione a opção **Information** um relatório mais detalhado é fornecido, conforme ilustrado nas Figuras 5.4 e 5.5.

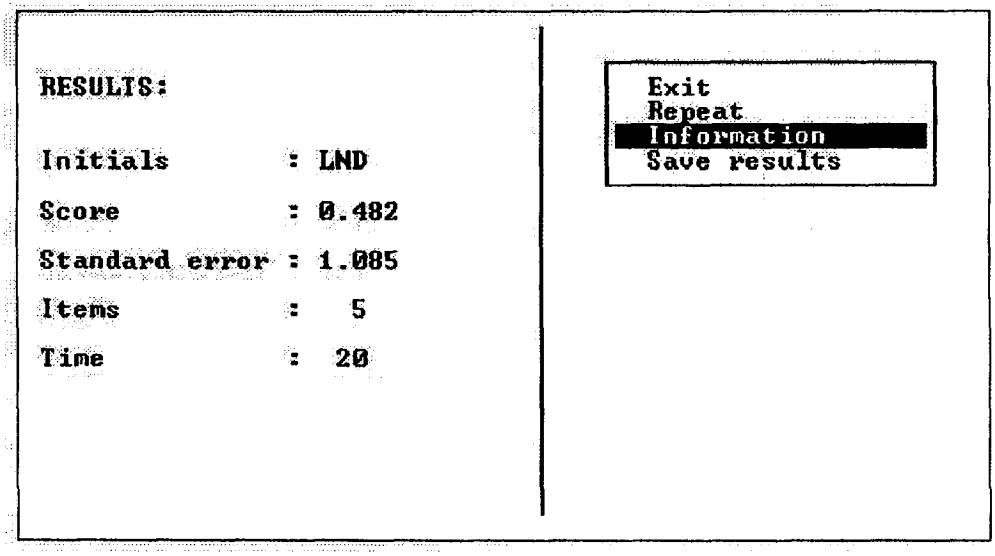

Figura 5.3: Tela que apresenta o resultado ao final de um teste

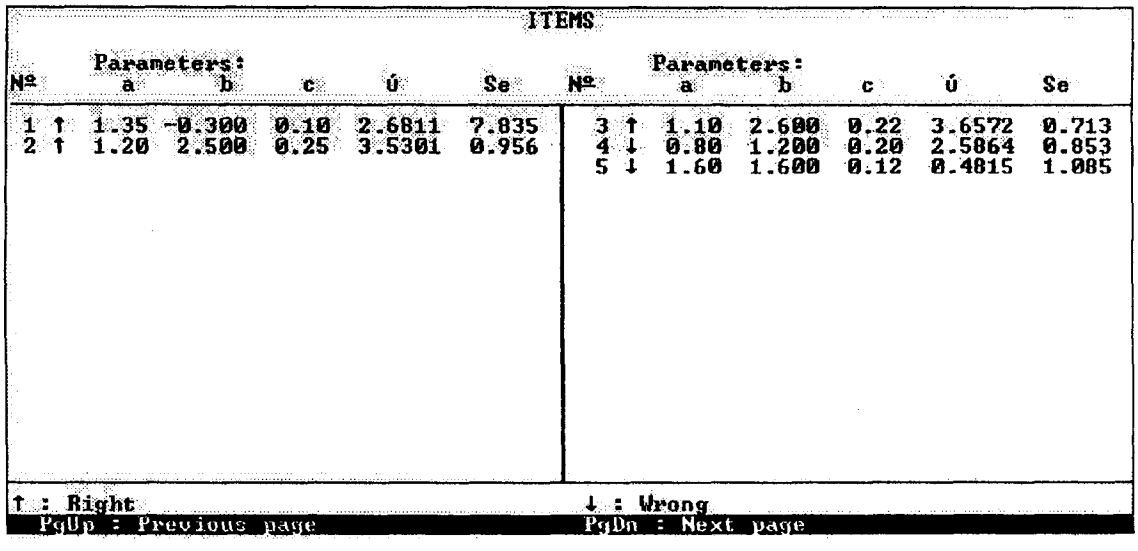

Figura 5.4: Tela que apresenta o resultado detalhado do teste, contendo informações de todos itens administrados.

A Figura 5.4 mostra o resultado detalhado do teste, contendo os valores dos parâmetros a, b e c de cada item administrado, as questões certas e erradas, o *Score* (representado por Ú) alcançado em cada item após a administração de cada um deles e o erro padrão associado (representado por Se). Note que no teste acima o aluno acertou o primeiro, o segundo e o terceiro itens, errando o quarto e o quinto. A habilidade estimada depois do primeiro item ter sido administrado foi de 2.6811 e contando que o segundo item foi respondido corretamente, a habilidade associada ao segundo item foi de 3.5301 e de 3.6572 após o aluno ter acertado o terceiro item. Observe que a

partir do quarto item, o qual o aluno respondeu incorretamente, sua habilidade caiu para 2.5864 tendo diminuído mais ainda com o erro da quinta questão, sendo o valor da habilidade de 0.4815, com valor do erro padrão de 1.085. Como o critério de parada adotado nesta versão foi o número de itens administrados, o processo de teste pára neste ponto e o último nível de habilidade torna-se a pontuação alcançada pelo estudante.

A Figura 5.5 mostra a distribuição normal da habilidade alcançada pelo aluno ao final do teste, com valor de $Z = 0.482$ (pontuação) e também o valor que representa a porcentagem de questões respondidas corretamente, nesse caso de $C = 69.20\%$.

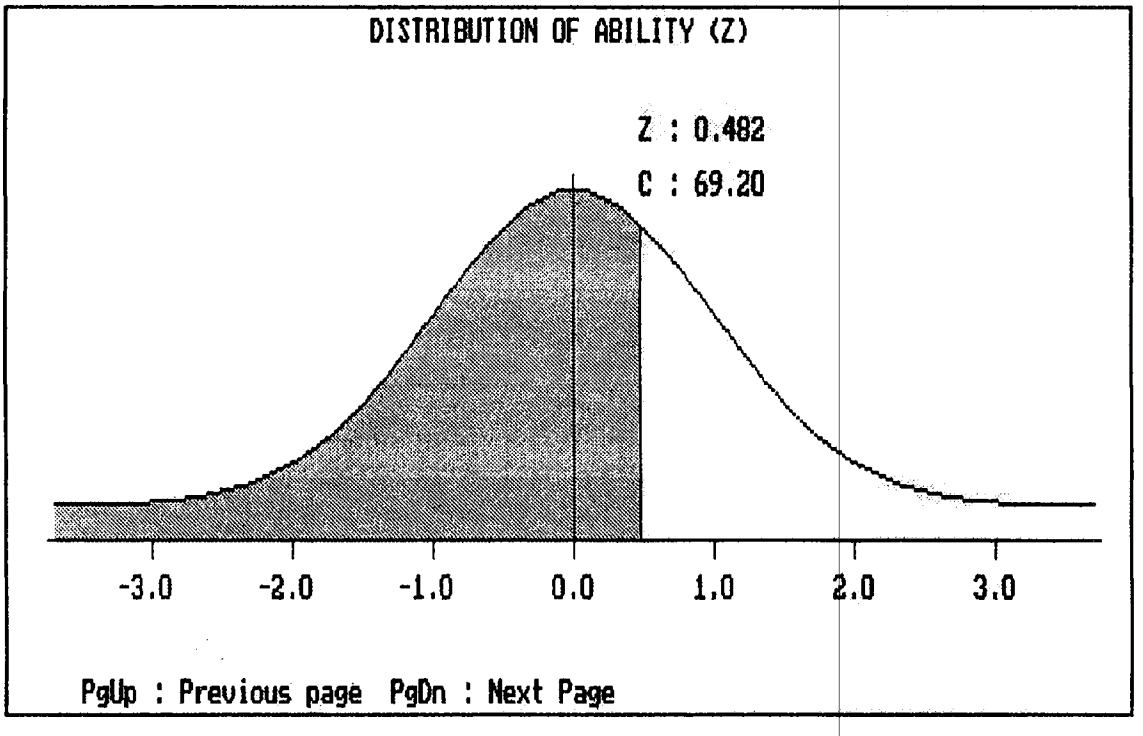

Figura 5.5: Distribuição normal da habilidade alcançada..

5.2.2 O FastTest Professional - Construção, Administração e Análise

O *FastTest Professional*[1] é um pacote de programas que realizam todas as tarefas envolvidas na criação, implementação e administração de um teste informatizado, bem como o desenvolvimento, organização e calibração de um banco de itens. Criado pela ***Assessement System Corporation*** (*www.assess.com*), que é responsável pelo desenvolvimento de vários programas comerciais de avaliação automática, o *FastTest Professional 1.5* (versão analisada nesta subseção) é a versão mais recente da família "FastTest" e traz consigo, além de todos os módulos funcionais acima citados, a opção de administração de testes adaptativos.

[1] Também pode ser referenciado como *FastTest Pro*.

Uma das principais vantagens do FastTest Pro é a possibilidade de incorporar no conteúdo dos testes as potencialidades que os computadores atuais fornecem, como recursos multimída e administração em rede (através da distribuição de licenças). Outra vantagem, que ainda cabe destacar, é a capacidade de criar várias abordagens de avaliações informatizadas, variando desde a implementação de testes convencionais até a aplicação dos modelos mais recentes dos testes adaptativos baseados na TRI. Além disso, pode-se usar uma poderosa ferramenta para a criação de bancos de itens, que por sua vez podem ser explorados em diversas avaliações de diferentes maneiras.

A versão aqui analisada é de demonstração (*Demo*) válida por apenas trinta dias. A Figura 5.6 mostra a tela inicial do programa contendo cinco botões, sendo que quatro deles representam os módulos funcionais que podem ser chamados e executados, e um de finalização (saída do programa). Conforme pode ser visto, a primeira opção (botão **Item Banker**) chama o programa responsável pela criação do Banco de Itens. A segunda opção representa a chamada do construtor do teste (botão **Test Builder**) cuja função é produzir testes com um conjunto de determinados itens pertencentes a um banco qualquer (que tenha sido criado anteriormente com a ferramenta **Item Banker**). Os itens que compõem um teste podem ser todos ou apenas parte dos itens de um banco. O terceiro botão (**Test Configuration**) chama o módulo de configuração de um teste, cuja principal função é decidir qual abordagem (convencional ou adaptativa) será utilizada, bem como quais os métodos e modelos serão seguidos, como os critérios de seleção e de estimação de parâmetros. A quarta opção (**Administrator**) aplica um teste criado pelo programa **Test Configuration**, e tem como principais objetivos coletar as respostas dos estudantes, calcular os resultados, selecionar itens a serem fornecidos e divulgar a pontuação obtida ao final da administração.

A título de experimento foi implementado o mesmo banco de itens do programa **AdTest** (analisado na subseção anterior) contendo itens do domínio de História Geral. Ao clicar na opção Item Banker o programa chama o módulo de criação de um banco de itens mostrado na Figura 5.7. Note que a tela é divida em duas partes. A parte esquerda mostra o banco corrente, contendo (se for o caso) todas as categorias e subdivisões que um banco pode possuir. Tais categorias, poderiam representar, por exemplo no caso da História Geral, o conjunto de itens que representam a História Antiga e Moderna, ou também testes que avaliam diferentes habilidades. A parte direita representa todos itens (ou questões) que fazem parte do banco ou de uma de suas categorias.

A Figura 5.8 representa a tela de edição e criação de um item (ITEM2). Observe que existe um conjunto de paletas (*Identifier, Item Text, etc*) que podem ser selecionadas, tendo cada uma delas uma tarefa específica, variando desde a identificação do item, a pontuação obtida, definição de opções de resposta corretas e incorretas, até a atribuição de parâmetros e outras. No exemplo da Figura 6.8, um determinado item está sob o modelo da TRI de três parâmetros, os quais são especificados.

Depois de criar o banco de itens e definir os tipos e características dos mesmos, o administrador do teste pode então criar o teste propriamente dito podendo administrá-lo em qualquer ocasião. A tarefa de construção, criação e edição de um teste é realizada pelo programa **Test Builder** e consiste na seleção do conjunto de itens de um determinado banco que farão parte do teste. A

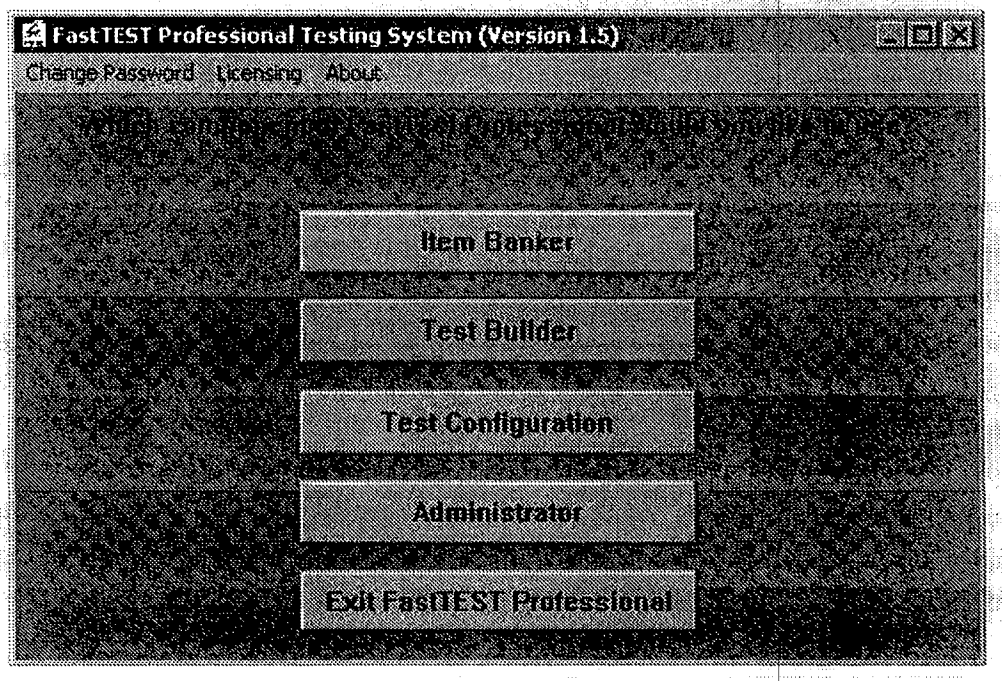

Figura 5.6: Tela inicial do programa FastTest Professional.

Figura 5.9 mostra a tela de criação de um teste, chamado de TSTHIST.TST (sobre o banco de itens de História Geral) em que todos os itens do banco farão parte da avaliação. Note que o programa mostra o banco e os itens selecionados para o teste da seguinte forma: BANKHIST/ITEM1 que descreve que o ITEM1 do banco BANKHIST está inserido no teste criado.

Uma vez realizada a criação de um teste é preciso decidir qual abordagem de administração será utilizada. A configuração do teste é realizada pelo módulo **Test Configurator** e consiste na determinação do modelo de seleção de itens que será adotado (sequencial, aleatório ou adaptativo); se adaptativo, qual dos métodos de estimação de parâmetros será empregado (MLE ou Bayesiano); a quantidade de itens a administrar; o tempo necessário para realização do teste e todas decisões que dizem respeito à maneira de administrar um determinado teste.

A Figura 5.10 mostra a tela de configuração de um teste. Note que o módulo de configuração possui várias paletas, nas quais são possíveis determinar o tipo de teste, a forma de administração, a pontuação e a forma de apresentação dos resultados.

Após a criação do banco, construção e configuração do teste, bem como a decisão da forma de aplicação do mesmo, chega a parte mais importante de todo esse processo: a administração do teste em uma avaliação. O módulo do ***FastTest Pro*** responsável por essa tarefa é o **Administrator**. Sua função consiste em ministrar o teste de acordo com todas as decisões tomadas anteriormente. Em outras palavras, o **Administrator** dirige o indivíduo durante o teste, coleta suas respostas, calcula sua pontuação e apresenta os resultados no final do teste.

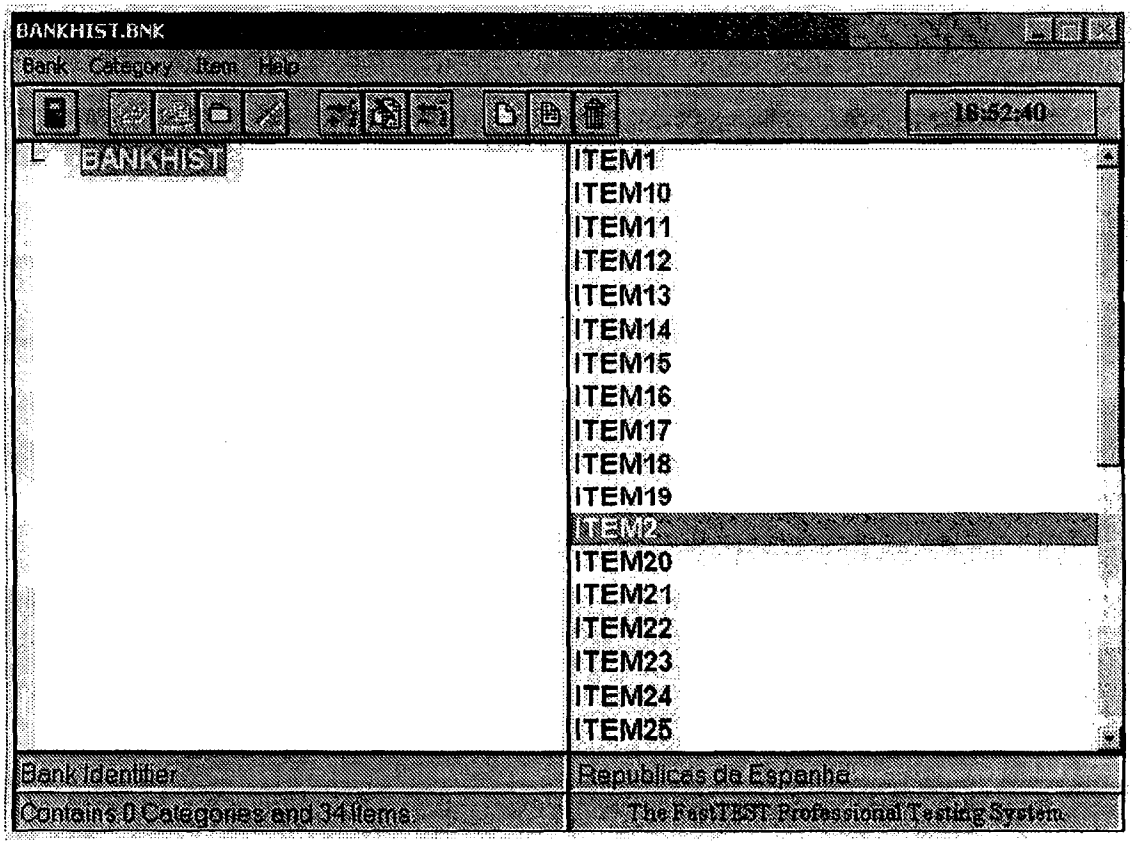

Figura 5.7: Tela de apresentação e operação de banco de itens

Devido ao exposto acima, parece fácil concluir que o **FastTest Pro** é um pacote de programas bastante completo, possuindo todas as ferramentas necessárias para a realização de uma avaliação automática. No entanto, uma desvantagem deste programa é não ser de livre disponibilização (*Freeware*) pois seu custo de manutenção e distribuição de licenças de uso é alto.

Na subseção seguinte é realizada uma comparação entre os dois programas descritos. A comparação foi realizada baseada na execução de cada um deles sob a ótica da abordagem adaptativa.

5.2.3 Comparação entre os Programas AdTest e FastTest Professional sob uma Abordagem Adaptativa

A Tabela 5.1 mostra um resumo das principais características dos programas descritos neste capítulo.

A análise dos programas realizada nesse trabalho teve como propósito apenas a observação da execução dos mesmos, considerando principalmente o comportamento do método de seleção de itens adotado, as variações dos valores da estimação de habilidade sob dois procedimentos (MLE e Bayesiano) e a medida do erro padrão.

A execução dos programas foi realizada sobre um banco de itens contendo trinta e quatro questões do domínio da História Geral, sendo ajustado ao modelo logístico de três parâmetros (a, b, e c) (veja Apêndice A). Para a seleção de itens utilizamos o método da Máxima Informação. Para

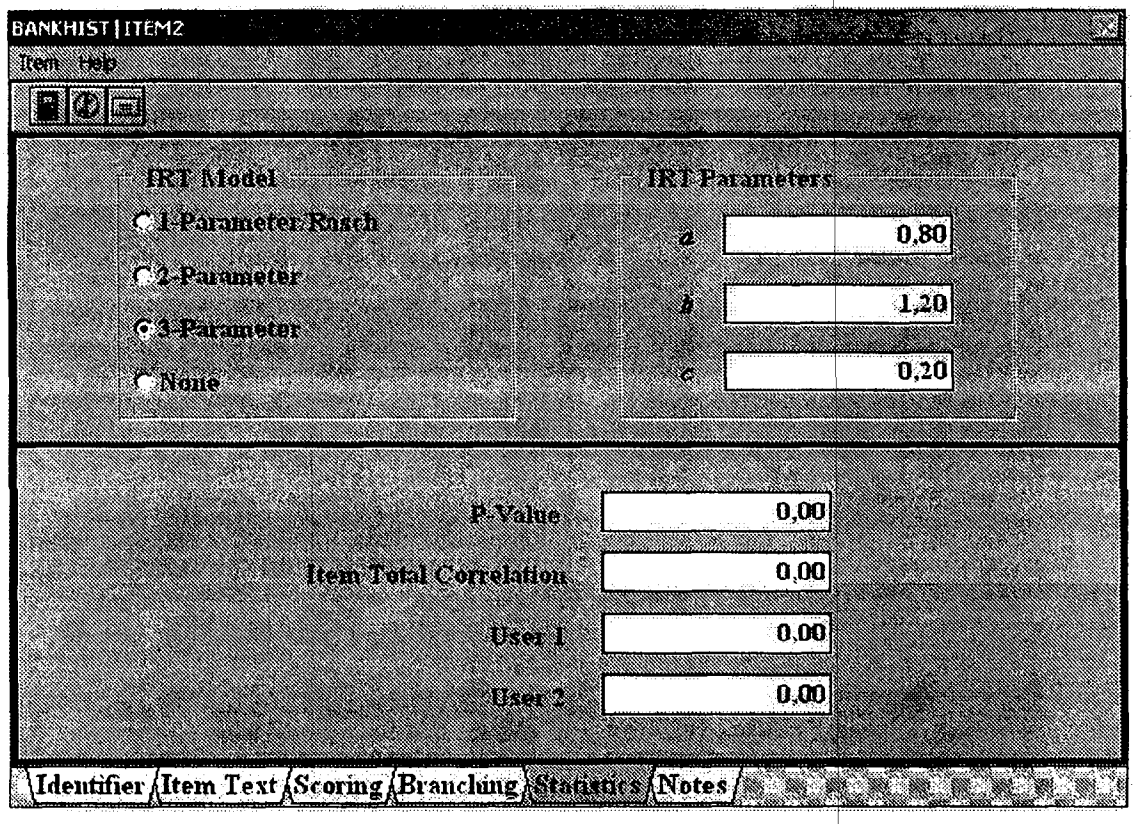

Figura 5.8: Tela que representa as tarefas de edição e criação de um item

Tabela 5.1: Comparação das principais características dos programas FasTest Pro e AdTest

Carac./Soft.	FASTTEST PRO	ADTEST
Abordagem de Teste	Testes Adaptativos	Testes Adaptativos
	Testes Discursivos (Essay)	
Método de Seleção de Itens	Máxima Informação	Máxima Informação
	Aleatório	
	Máxima Informação	
Estimação de Habilidade	Est. de Máxima Verosimilhança (MLE)	Est. de Máxima Verosimilhança
	Bayesiano	
Critério de Parada	Variável - pode ser definida pelo administrador do teste	Fixo (número de itens)
Formatação dos Itens	Múltipla Escolha	Múltipla Escolha
Módulo de Criação do Banco de Itens	Sim	Não
Módulo de Construção e Configuração dos testes	Sim	Não
Administração do Teste	Sim	Sim
Sistema Operacional	Windows	DOS
Disponibilização	Shareware (tempo limitado)	Freeware (livre)

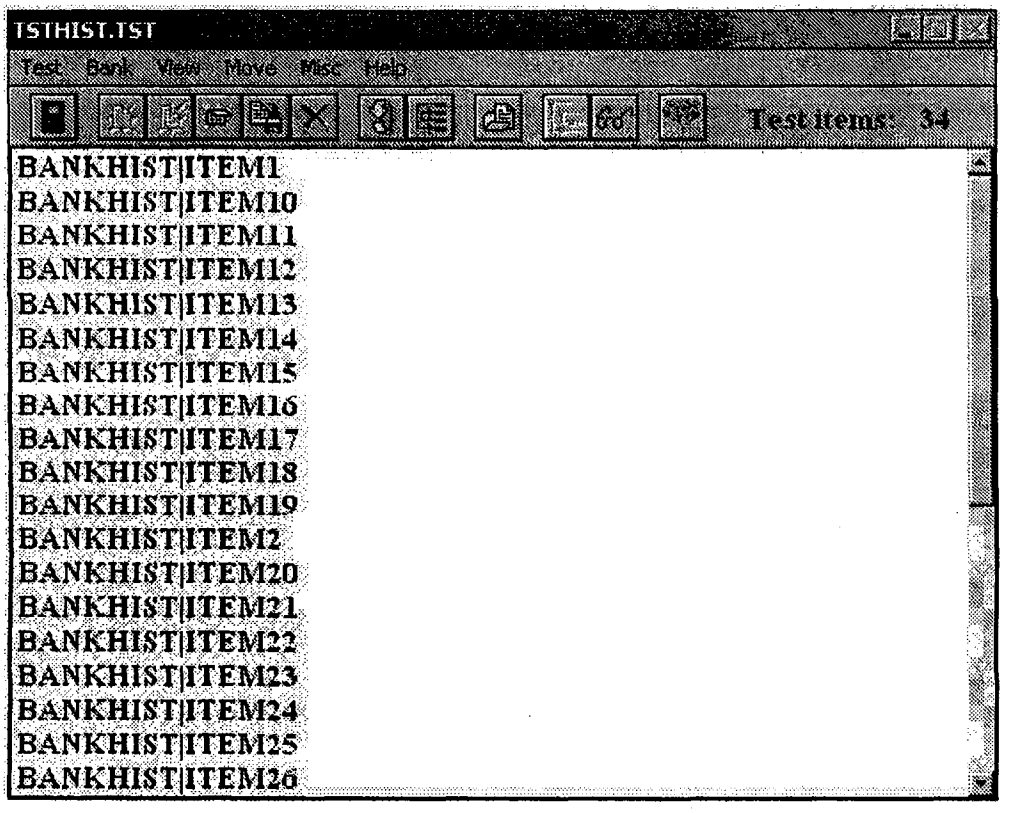

Figura 5.9: Tela que representa a tarefa de criação e edição de um teste

a estimação da habilidade utilizamos o método da Estimativa de Máxima Verosimilhança (MLE) e Bayesiano no programa **FastTest Pro**, e o MLE no **AdTest**. Como critério de parada foi adotada a administração fixa de cinco itens, isso porque o banco de itens é pequeno para uma amostragem maior. A habilidade inicial nos dois programas foi definida como um valor aleatório entre o intervalo -1.0 e +1.0. A Tabela 5.2 mostra os dados obtidos da análise.

A análise foi realizada por meio da simulação de doze situações diferentes de testes, considerando a quantidade de acertos e erros que podem acontecer, bem como sua ordem de acontecimento. Em outras palavras, foram consideradas todas as situações de teste em que um aluno, sempre respondia todos os itens correta ou incorretamente; ou por uma seqüência de respostas certas e erradas e vice-versa; e ora correta e incorreta alternadamente. Observe na Tabela 5.2 que existem duas colunas mestre que representam os programas analisados (FastTest Pro e AdTest), nas quais são especificados (pela subdivisão) os métodos de estimação de habilidade empregados por cada um deles. Cada conjunto de duas linhas (em pares de cima pra baixo) representa uma situação de teste, contendo a seqüência dos itens administrados, a combinação de respostas corretas e incorretas adotada, a medida de habilidade (representada por S), e a medida do erro padrão associado a cada estimativa(representado por SE). Uma resposta correta é representada pelo valor 1 e incorreta pelo

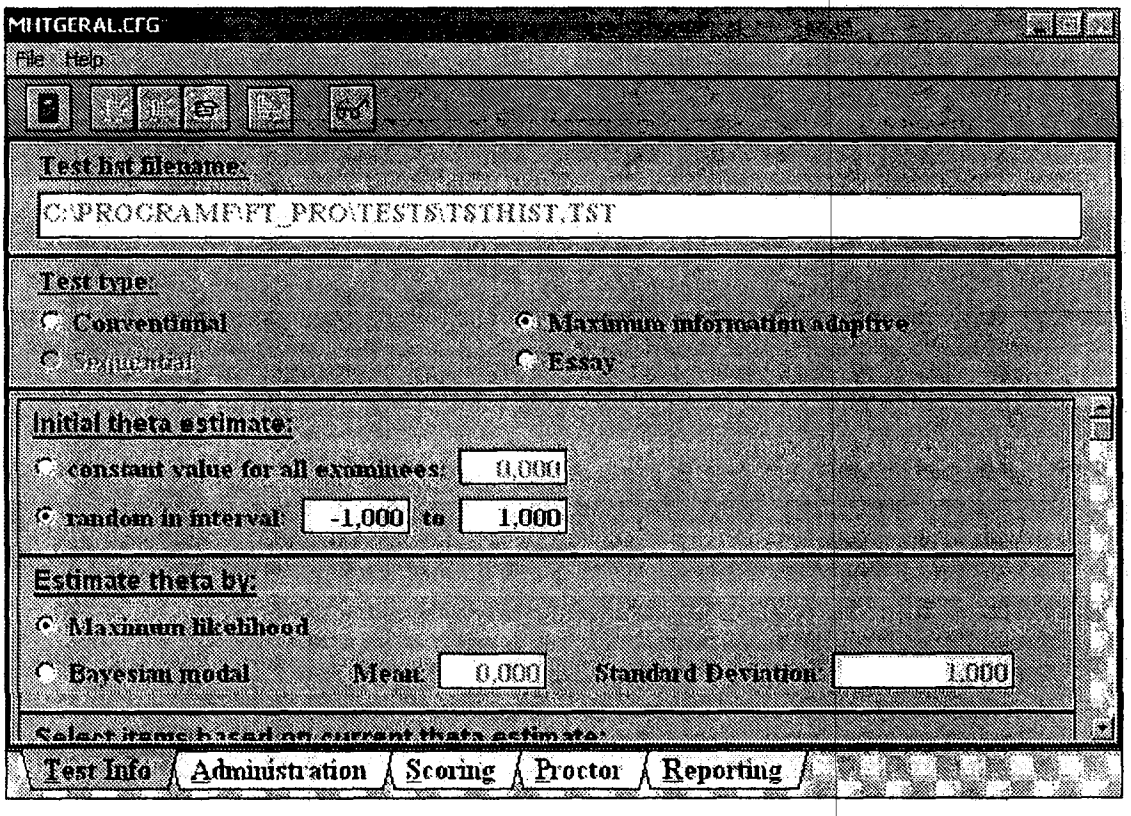

Figura 5.10: Tela de configuração de um determinado teste

valor 0. Por exemplo, a seqüência de respostas 11100 significa que as três primeiras questões foram respondidas corretamente e as duas últimas incorretamente.

Baseado na análise dos programas descrita na Tabela 5.2, tanto no método MLE como no Bayesiano, os itens 20, 21 e 13 foram os mais selecionados no início do processo de teste, sendo que o item 12 foi selecionado apenas uma vez. Isso se deve ao fato de que tais itens oferecem maior nível de informação quando a estimativa de habilidade está entre -1.0 e +1.0 (habilidade inicial). Outro dado que cabe salientar é que quando a primeira questão é respondida corretamente o item mais selecionado, sob o método MLE, foi o 34, tanto nos programas FastTest Pro e AdTest. De outra forma, depois de errar a primeira questão o item que mais apareceu foi o 16 (também nos dois programas) sob o método MLE. Por sua vez, o método Bayesiano selecionou mais vezes o item 21 quando houve um acerto na primeira questão, e variando entre os itens 7 e 20 quando houve um erro inicial. Percebe-se com esses dados que os métodos MLE e Bayesiano selecionam o primeiro item mais ou menos de forma semelhante, alterando o processo de seleção (que pode ser visto como adaptação) já a partir do segundo item administrado.

Outro dado importante que cabe ressaltar, é que a habilidade S estimada com os métodos MLE tanto do FastTest Pro quanto no AdTest, foi semelhante apenas nos casos "extremos" em que o indivíduo acertou ou errou todos os itens. Em outros casos, em que não houve muitos acertos ou

Tabela 5.2: Dados da execução dos programas FastTest Pro e AdTest usando os métodos MLE e Bayesiano sobre um banco de itens de História Geral

	FASTTESTPRO										ADTEST				
	MLE					BAYES					MLE				
Itens	20	34	18	3	27	13	21	25	27	12	20	34	18	3	2
11111	S = 4.0* / SE = 1.0*					S = 1.841 / SE = 0.496					S = 3.701 / SE = 0.646				
Itens	20	34	18	3	27	20	21	12	27	3	20	34	18	3	2
11110	S = 1.780 / SE = 1.0*					S = 1.625 / SE = -0.509					S = 2.802 / SE = 0.766				
Itens	21	34	18	3	27	13	21	25	27	12	13	34	18	3	27
11100	S = 1.147 / SE = 0.843					S = 1.22 / SE = 0.492					S = 0.689 / SE = 1.758				
Itens	13	34	18	27	21	20	21	12	25	6	20	34	18	27	21
11000	S = - 0.308 / SE = 0.829					S = 0.2 / SE = 0.499					S = 0.002 / SE = 0.793				
Itens	20	34	12	21	22	21	25	20	22	13	20	34	12	21	22
10000	S = - 0.707 / SE = 0.652					S = - 1.106 / SE = 0.524					S = - 0.707 / SE = 0.784				
Itens	20	16	1	28	19	21	20	22	13	31	21	16	19	1	26
00000	S = - 4.0* / SE = 1.0*					S = - 1.424 / SE = 0.496					S = - 3.713 / SE = 0.719				
Itens	20	16	1	28	19	13	7	31	18	1	21	16	19	1	28
00001	S = - 3.685 / SE = 1.0*					S = - 1.589 / SE = 0.454					S = - 3.570 / SE = 0.964				
Itens	13	16	1	28	31	13	7	31	28	4	21	16	19	1	28
00011	S = - 1.927 / SE = 0.619					S = - 1.225 / SE = 0.418					S = - 2.003 / SE = 0.788				
Itens	20	16	13	20	5	21	20	22	5	33	21	16	19	1	28
00111	S = - 1.283 / SE = 0.606					S = - 0.055 / SE = 0.431					S = - 1.519 / SE = 0.767				
Itens	20	16	13	20	5	21	20	5	33	30	20	16	3	7	22
01111	S = 0.239 / SE = 0.577					S = 0.43 / SE = 0.478					S = - 0.261 / SE = 0.542				
Itens	13	34	21	27	25	21	25	20	12	6	20	34	12	27	25
10101	S = 1.143 / SE = 0.616					S = 0.556 / SE = 0.471					S = 1.343 / SE 0.605				
Itens	13	16	28	1	31	12	20	21	5	33	20	16	13	28	31
01010	S = - 2.127 / SE = 0.711					S = - 0.116 / SE = 0.441					S = - 1.641 / 0.614				

erros, os valores de S foram diferentes, sendo que os valores do erro padrão no AdTest foram menores em oito das doze situações propostas. Ainda, a estimação de habilidade S no método Bayesiano teve um comportamento uniforme (sem valores muito baixos ou muito altos) em todas as situações de teste, sendo que o erro padrão associado foi sempre menor do que no MLE. Isso reafirma o fato que de a distribuição de habilidade obedece uma distribuição normal em torno de uma média.

O fato mais interessante que pode ser retirado dessa análise é que a ordem das respostas (correta e incorreta) interessa. Observe as linhas em que um indivíduo acertou duas e errou três questões (11000), e a linha onde errou três e acertou duas (00011) nessa ordem. Note que o número total de acertos e erros foi o mesmo, 2 e 3 em cada situação. Entretanto, os itens fornecidos e a habilidade final estimada foram totalmente diferentes (-0.308 e -1.927 respectivamente). Tal fato mostra que os testes adaptativos ajustam os itens ao nível de competência do indivíduo, considerando seus parâmetros e a resposta dada a cada um deles durante o teste.

A análise descrita na Tabela 5.2 possui conteúdo informativo, servindo apenas para mostrar o comportamento geral dos procedimentos de seleção e estimação de habilidade que podem ser empregados nos testes adaptativos, não tendo qualquer objetivo de destacar a importância de um método sobre outro, precisando para isso, uma análise mais detalhada.

CAPÍTULO 6

O Desenvolvimento de um Teste Adaptativo Sensível ao Conteúdo

Os objetivos definidos para este projeto de mestrado foram o estudo da teoria e dos formalismos dos testes adaptativos e, posteriormente, com base nesse estudo, o projeto e implementação de uma aplicação que pudesse disponibilizar a avaliação adaptativa diagnóstica no cenário do Exame de Proficiência em Inglês (EPI) para admissão no programa de mestrado do ICMC. Entretanto, as pesquisas e estudos realizados no primeiro ano de projeto explicitaram uma condição primária para a realização de uma avaliação adaptativa: a necessidade da existência de um banco de itens robusto que pudesse dar suporte à aplicação adaptativa. Assim, concluiu-se que o banco de itens é o elemento central de um sistema adaptativo e que os esforços devem se concentrar principalmente na criação e calibração dos itens que compõem um banco.

Portanto, diante de tal cenário, o segundo ano de trabalho foi dedicado exclusivamente à criação e calibração de um banco de itens sensível ao conteúdo populado com questões referentes ao Exame de Proficiência em Inglês do ICMC. Entre outras tarefas que se fizeram necessárias para a criação do banco, destaca-se o desenvolvimento de um *Sistema Gerenciador do Banco de Itens* (**SisBI**) que tem como objetivo principal facilitar a manutenção da base de itens, oferecendo tarefas específicas como a inclusão, exclusão, alteração e visualização de itens e a análise dos mesmos sob a luz da TRI.

Além da criação do banco de itens, cuja manutenção foi facilitada pela implementação do **SisBI**, também foi desenvolvida a modelagem de um sistema de avaliação adaptativa[1]. Construída em UML (*Unified Modeling Language*) (Craig, 1998), tal modelagem tem o propósito de auxiliar uma posterior implementação do sistema, de maneira a guiar o programador na tarefa de codificar as funções inerentes à aplicação dos processos do teste adaptativo.

[1]Também pode ser chamado de TAEPI - Testes Adaptativos para o Exame de Proficiência em Inglês.

Este capítulo apresenta os resultados obtidos neste trabalho, destacando principalmente as atividades empregadas na criação do banco de itens, na implementação do **SisBI** e o desenvolvimento da modelagem do protótipo do sistema, justificando as decisões do projeto.

6.1 Criação e Calibração do Banco de Itens

6.1.1 Organização do Conteúdo do Banco de Itens

O conteúdo do banco de itens para o teste adaptativo foi organizado de acordo com a distribuição dos módulos da Tabela 1.1. Dessa forma, consideramos que cada módulo do exame corresponde a uma "habilidade" que será avaliada durante um teste. Além disso, cada módulo é divido em partes, que aqui podem ser entendidas como subconteúdo do conteúdo de um módulo. Por sua vez as questões (itens) foram distribuídas entre estas partes de maneira que tal estratégia garanta o *balanceamento de conteúdo*. Para garantir a sensibilidade de conteúdo do banco, também foram previstas as existências de TESTLETS (subseção 2.7.1) cujos grupos de questões foram organizados em algumas das partes presentes nos módulos.

As questões e/ou *Testlets* foram divididos em seis partes diferentes distribuídas entre os módulos. São elas: GAP e PURPOSE (pertencentes ao módulo 1), INTRODUCTION e ABSTRACT (módulo 2), SETTING e REVIEW (módulo 4)[2]. O módulo 3 (compreensão de texto) não contém partes e suas questões são agrupadas em *Testlets*, pois todos se referem a um único texto. Abaixo segue uma descrição de cada parte e as questões a que se referem:

- GAP: abriga questões que avaliam o conhecimento sobre os marcadores de discurso utilizados para indicar as lacunas encontradas na Revisão da Literatura;
- PURPOSE: representa as questões que avaliam o conhecimento sobre os tempos verbais utilizadas na indicação do propósito de um trabalho;
- INTRODUCTION: armazena *Testlets* (grupos de questões) avaliando a estruturação da seção de Introdução de um artigo;
- ABSTRACT: armazena *Testlets* avaliando a estruturação de resumos (Abstracts) de um artigo;
- SETTING: parte que contém questões referentes às estratégias de escrita relativas ao componente *contexto* (setting) de uma seção de introdução;
- REVIEW: reúne questões referentes às estratégias de escrita relativas ao componente de *revisão bibliográfica* (review) de uma seção de introdução.

[2] Atualmente, somente dois componentes da estrutura de uma Introdução são utilizados para tratar das convenções da língua inglesa para textos científicos. Da mesma forma, somente dois componentes também da Introdução são tratados nas estratégias do módulo 4.

A organização do conteúdo do banco de itens de acordo com cada módulo e suas respectivas partes é apresentada na Figura 6.1. Observe que cada módulo representa uma divisão do conteúdo, e suas respectivas partes podem ser vistas como especializações deste conteúdo tornando-o flexível no momento de uma avaliação.

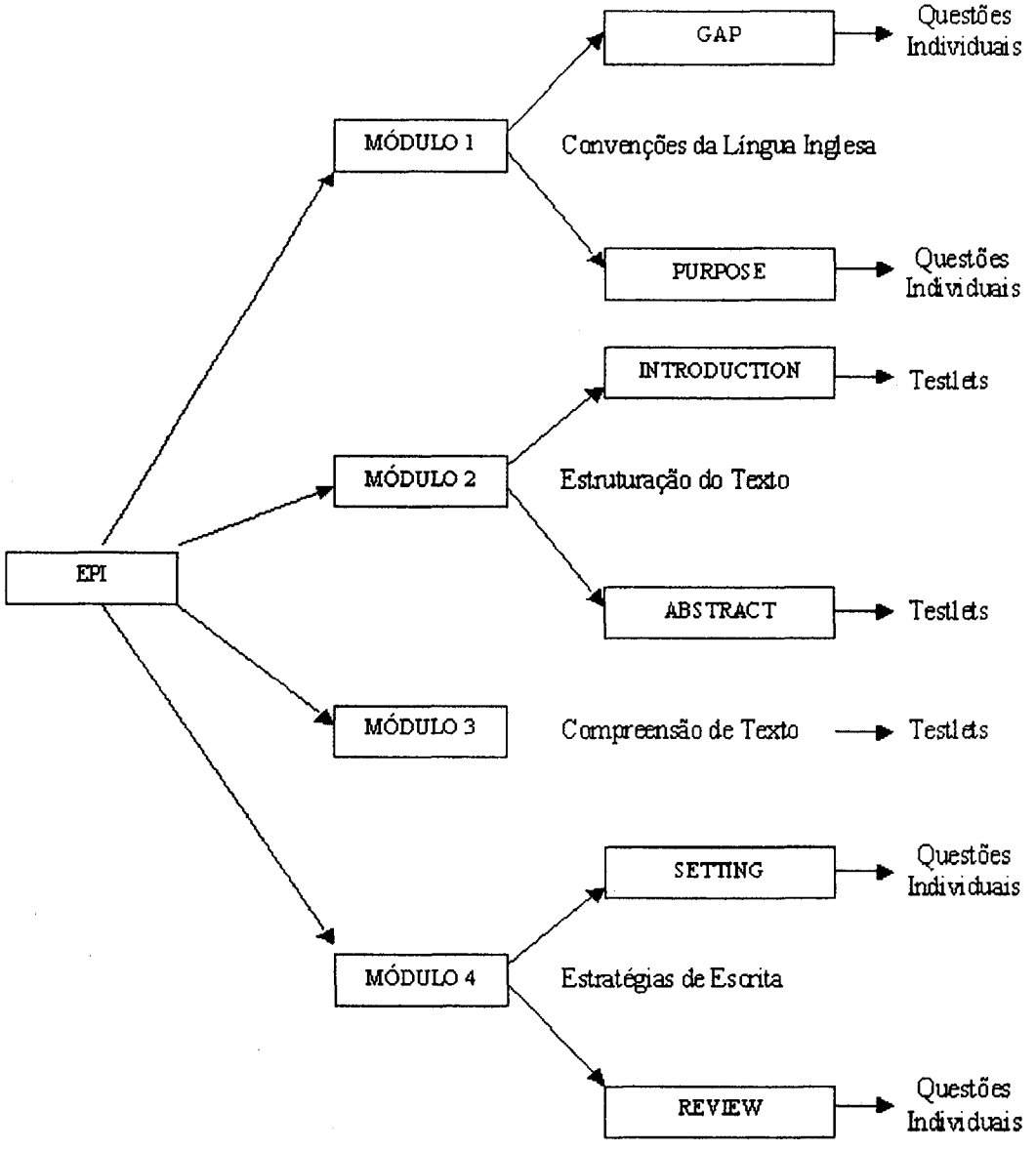

Figura 6.1: Organização do conteúdo (módulos e partes) do banco de itens para o Exame Diagnóstico Adaptativo

6.1.2 Coleta dos Dados de Análise

Conforme dito no Capítulo 1, a partir do primeiro semestre de 2001 o Exame de Proficiência em Inglês passou a ser realizado de forma computadorizada (por meio do sistema CAPTEAP). Durante todo o ano de 2001 foram aplicados vários exames aos alunos da pós-graduação do ICMC, sendo que alguns deles tinham caráter de treinamento, com intuito de familiarizar os alunos com a interface e conteúdo dos exames do CAPTEAP, e outros de caráter definitivo e formal, cujos resultados eram considerados oficiais pelo programa de pós-graduação deste instituto.

Portanto, como a intenção deste mestrado era construir e calibrar um banco de itens relacionados ao EPI, os resultados da aplicação destes exames, considerando tanto os de treinamento quanto os oficiais, foram utilizados, pois para o processo de calibração dos itens era preciso um número de ocorrências (alunos que fizeram um exame) considerável.

Ao todo foram coletados oito exames (provas), cada um com um número diferente de alunos que os realizaram. Cada prova pertence a uma dada subárea de pesquisa que é contemplada pela pós-graduação deste Instituto, sendo estas: Computação, Matemática Computacional e Estatística. A divisão das provas de acordo com a subárea de concentração do Instituto ocorre porque as questões de algumas partes incluídas no exame, por exemplo INTRODUCTION e ABSTRACT (módulo 2), dependem diretamente de um texto que é fornecido ao aluno no momento de avaliação, sendo este pertencente à área em que o aluno está inserido. A Tabela 6.1 mostra a identificação das provas coletadas e a área a qual pertencem.

Tabela 6.1: Identificação das provas e área pertencente

Identificação da Prova	Área
Código 27	Computação
Código 29	Computação
Código 30	Estatística
Código 32	Matemática Computacional
Código 38	Computação
Código 39	Estatística
Código 40	Matemática Computacional
Código 41	Computação

A prova de Código 27 foi elaborada com o intuito de coletar dados massivos de respostas de alunos (para favorecer o processo de estimação) por meio de um pedido de *auxílio a experimento* para a comunidade de alunos de pós-graduação do ICMC. A Tabela 6.2 mostra a carta de solicitação enviada aos alunos.

6.1.3 Organização das Provas do EPI

Dado o fato de que o conteúdo do banco de itens é dividido em módulos (obedecendo à hierarquia apresentada na Figura 6.1), a primeira tarefa desempenhada tendo a posse das provas, foi identificar

Tabela 6.2: Carta de solicitação de colaboração enviada aos alunos do ICMC

Solicitação de Colaboração com um Experimento Científico

No programa de mestrado do Instituto de Ciências Matemáticas e de Computação (ICMC-USP), os estudantes ingressantes são avaliados com relação à proficiência em inglês por meio de um exame informatizado que além de avaliar a compreensão de texto, avalia a habilidade de reconhecer o gênero de textos científicos em inglês, com a estrutura e convenções da língua que lhes são características.
No mesmo Instituto há um trabalho de mestrado que estuda os Testes Adaptativos Informatizados (TAI's) - CAT Computer Adaptive Testing, em inglês - e tem por objetivo implementar um teste de diagnóstico, usando a abordagem adaptativa, sob o domínio de conhecimento do teste acima descrito. Entretanto, para a implementação deste teste, é preciso elaborar e calibrar um banco de questões referente ao conteúdo do exame. A calibração do banco é realizada por meio de uma análise estatística das questões (pertencentes ao banco) na qual são considerados, principalmente, suas respostas e o número de pessoas que as responderam.
Contudo, para atingirmos um bom resultado é preciso que um considerável número de pessoas (voluntários) tenha a disposição para realizar um exame/teste, para que tenhamos uma boa "massa" de informações e possamos realizar uma análise confiável.
Este e-mail tem o objetivo de solicitar a participação voluntária de vocês no sentido de colaborar com a calibração de um banco de questões, permitindo a coleta de informações empíricas sobre o exame. A colaboração consiste:1) no acesso ao site: http://nilc.icmc.sc.usp.br/staff/capteap/pgprovateste.php e na resolução do exame de proficiência em inglês lá alocado.2) no envio de e-mail para leandroh@icmc.sc.usp.br dizendo a duração do exame realizado por você. Assim, esperando sua valorosa colaboração, nos despedimos agradecendo desde já a atenção e apoio dispensados.

Atenciosamente,

Leandro Henrique Mendonça de Oliveira - ICMC-USP

Profa. Sandra Maria Aluísio - ICMC-USP

quais módulos e partes eram cobertos em cada uma delas. Assim, para facilitar a identificação dos módulos e suas respectivas partes foi definida a seguinte notação[3] de acordo com a Tabela 6.3.

A tarefa de identificação do conteúdo das provas produziu a Tabela 6.4, que mostra os módulos e partes cobertas por cada prova. Observe que os mesmos estão dispostos na mesma ordem que uma prova do EPI utiliza, e que de certa forma estão uniformemente distribuídos cobrindo todos os módulos existentes. Repare na Tabela 6.4 que o conteúdo relacionado aos módulos M2I, M2A e M3 está presente na grande maioria das provas. Os módulos M1G, M1P, M4S e M4R estão alternados em cada uma delas. Por último, veja que as provas de códigos 29, 30 e 32 contêm todos os módulos e partes existentes.

[3]Daqui para frente toda referência a um dado módulo e parte será realizada por esta identificação.

Tabela 6.3: Identificação de Módulos e Partes

Módulo e Parte	Identificação
Módulo 1 parte GAP	M1G
Módulo 1 parte PURPOSE	M1P
Módulo 2 parte INTRODUCTION	M2I
Módulo 2 parte ABSTRACT	M2A
Módulo 3	M3
Módulo 4 parte SETTING	M4S
Módulo 4 parte REVIEW	M4R

Tabela 6.4: Distribuição dos módulos e partes em cada prova

Código 27 - Computação							**Código 38 - Computação**				
M2I	M4S	M3	M1G	M1P			M2I	M2A	M4S	M3	M1G
Código 29 - Computação							**Código 39 - Estatística**				
M2I	M2A	M4S	M4R	M3	M1G	M1P	M2I	M2A	M4S	M3	M1G
Código 30 - Estatística							**Código 40 - Matemática**				
M2I	M2A	M4S	M4R	M3	M1G	M1P	M2I	M2A	M4S	M3	M1G
Código 32 - Matemática							**Código 41 - Computação**				
M2I	M2A	M4S	M4R	M3	M1G	M1P	M2I	M2A	M4R	M3	M1P

6.1.4 Análise e Seleção das Questões Válidas

Terminada a atividade de identificação dos módulos e partes contidas nas provas, foi necessário realizar uma análise dos itens (questões) com a finalidade de identificar questões semelhantes entre todas as provas e separá-las em grupos relacionados com cada parte. Tal atividade permitiu a contabilização das questões, a formação dos conjuntos de estimação (formado por todas as questões pertencentes a uma determinada parte) e o conhecimento dos itens que fariam parte do processo de estimação.

A tarefa de identificação de itens semelhantes foi realizada através da observação cuidadosa de todas as questões das provas, sendo verificada parte por parte e questão por questão individualmente, tendo como saída um relatório final contendo todas as questões semelhantes. O relatório final dessa análise é mostrado na Tabela 6.5.

Na primeira linha da Tabela 6.5 encontra-se a identificação de cada prova. A quantidade de questões de cada prova é apresentada ao final de cada coluna que representa uma dada prova, por exemplo, o total de questões contidas na prova de *Código 27* foi 32, na prova *Código 29* foi de 26 e assim sucessivamente. O total geral de questões considerando todas as provas foi de 190. Observando a tabela é possível encontrar várias linhas contendo códigos do tipo **20=21.29** os quais indicam que a questão de número **20** da prova em questão (código na primeira linha) é igual à questão de número **21** da prova de *Código 29*. Um código do tipo **28=19** indica que a questão **28**

Tabela 6.5: Relatório das questões semelhantes entre as provas e número total de questões válidas

Cód. 27	Cód. 29	Cód. 30	Cód. 32	Cód. 38	Cód. 39	Cód. 40	Cód. 41
01	01	01	01	01	01=02.30	01=01.32	01=01.29
02	02	02	02	02	02=03.30	02=02.32	02=02.29
03	03	03	03	03	03=04.30	03=04.32	03=03.29
04	04	04	04	04	04=05.30	04=05.32	04=04.29
05	05	05	05	05	05	05	05
06	06	06	06	06	06	06	06
07	07	07	07	07	07	07	07
08	08	08	08	08	08	08	08
09	09	09	09	09=09.27	09=09.38	09=09.39	09-Perdida
10	10	10	10	10=12.32	10=10.38	10=10.39	10-Perdida
11	11	11	11	11=13.32	11=11.38	11=11.39	11-Perdida
12=09	12	12=12.29	12	12	12=12.38	12=12.39	12=12.40
13	13	13=13.29	13	13	13=13.28	13=13.39	13
14	14	14=14.29	14=12.30	14=15.29	14=14.38	14=14.39	14
15	15	15	15=13.30	15=16.29	15=15.38	15=15.39	15
16	16	16	16=14.30	16=17.29	16=16.38	16=16.39	16=16.40
17	17	17	17=15.30	17	17=17.38	17=17.39	17=17.40
18	18	18	18=16.30	18=20.29	18=18.38	18=18.39	18=25.29
19	19	19	19=17.30	19=21.29	19=19.38	19=19.39	19=26.29
20	20	20=21.29	20	20=22.29	20=20.38	20=20.39	20
21	21	21=22.29	21				
22	22	22	22				
23=22	23	23	23				
24	24	24	24				
25	25	25	25				
26=17	26		26				
27			27				
28=19							
29							
30							
31							
32							
28 q.	26 q.	20 q.	21 q.	11 q.	4 q.	4 q.	8 q.

é semelhante à questão **19** da mesma prova. O número de questões semelhantes em todas as provas foi de *68*.

A última linha da Tabela 6.5 indica a quantidade de questões válidas de cada prova, ou seja, o total de questões autênticas que cada prova possui. Assim, o total de itens válidos que deveriam participar do processo de estimação de parâmetros foi de 122. As questões 09, 10 e 11 da prova de *Código 41* foram anuladas.

6.1.5 Elaboração dos Conjuntos de Dados para Estimação

Encerrada a fase de identificação das questões semelhantes contidas em todas as provas, bem como a contabilização das questões válidas, foi necessário elaborar vários conjuntos de questões que posteriormente iriam servir para facilitar o processo de calibração.

A aplicação de uma avaliação multidimensional (medindo várias habilidades) exige a construção de um banco de itens sensível ao conteúdo, obedecendo a regra de que cada conteúdo inserido no banco representa uma habilidade a ser medida. Entretanto, o processo de estimação dos paramentos de um banco de itens que implementa o balanceamento de conteúdo é realizado separadamente, sobre

cada conteúdo que será incluído no banco. Assim, para o EPI, as questões referentes a cada módulo e parte foram e futuramente serão estimadas de forma isolada, como se cada módulo fosse um banco de itens diferente.

Em razão da sensibilidade de conteúdo do banco de itens, proporcionada pela existência dos vários módulos que caracterizam a possibilidade de uma avaliação multidimensional, bem com a necessidade de estimação separada dos conteúdos do banco, houve a necessidade de uma organização das questões das provas em determinados grupos, que foram aqui chamados de *Conjuntos de Análise*. Além disso, outro fato que justificou a criação dos *conjuntos de análise* foi que as questões semelhantes deviam ser agrupadas em um único conjunto, aproveitando assim as ocorrências de resposta (respostas dos alunos) de todas as provas a fim de melhorar o resultado da calibração. Assim, um *conjunto de análise* é um grupo de questões referentes a um mesmo módulo e parte contendo todas as ocorrências de respostas fornecidas pelos alunos, sendo estas pertencentes a uma ou mais provas. Por exemplo: tomamos o módulo e parte identificada pela sigla **M2I** conforme descrição da Tabela 6.3. Observe na Tabela 6.5 que as quatro primeiras questões da prova **29** são iguais às quatro primeiras questões da prova **41**, e que de acordo com a Tabela 6.4 representam o módulo 2 da parte de INTRODUCTION. Assim, um conjunto de análise poderia ser formado com estas questões, já que elas são iguais e representam o mesmo módulo e parte.

A elaboração dos conjuntos de análise foi realizada identificando as questões iguais em todas as provas em relação a cada módulo pertencente ao banco. Assim, foram definidos **31** conjuntos de análise, alguns deles contendo um único grupo de questões referentes a uma única prova, outros possuindo grupos de questões de até quatro provas diferentes. A Tabela 6.6 mostra todos os conjuntos de análise definidos. Por exemplo: observe que o módulo e parte definida pela sigla **M4S** contêm quatro conjuntos de análise. O primeiro conjunto contém as questões de 8 a 13 da prova **27**, identificada pelo código ***C27-08-13***. O segundo conjunto contém três questões (de $9-11$) da prova **29** representado pelo código ***C29-09-11*** e o terceiro pelas questões 10 e 11 da prova **30**. O último conjunto, o maior deles, contém as questões 11 a 13 da prova **32** (***código C32-11-13***) e os itens 10 e 11 das provas **38, 39 e 41 (códigos: C38-10-11,C39-10-11 e C40-10-11)** respectivamente. Vale lembrar que todas as questões desse último conjunto são iguais e dizem respeito ao mesmo conteúdo do módulo 4.

Generalizando, cada conjunto de análise pode ser entendido como uma reunião de itens e suas respostas, que fará parte do processo de estimação de parâmetros realizada automaticamente[4], ou seja, um conjunto análise nada mais é do que os dados que farão parte do arquivo de entrada do programa que realiza a estimação.

Dessa forma, podemos entender os conjuntos de análise como sendo os "grupos" de itens que serão estimados, respeitando o balanceamento de conteúdo definido na estrutura do banco de itens.

[4]No nosso caso, o processo de estimação foi realizado pelo programa XCALIBRE.

Tabela 6.6: Relação dos conjuntos de análise distribuídos por módulo

Módulos e Partes	Conjuntos de Análise						
M2I	C27-01-07	C29-01-04 C41-01-04	C30-01-05 C39-01-04	C32-01-06 C40-01-04	C38-01-04		
M2A	C29-05-08	C30-06-09	C32-07-10	C38-05-08	C39-05-08	C40-05-08	C41-05-08
M4S	C27-08-13	C29-09-11	C30-10-11	C32-11-13 C38-10-11 C39-10-11 C40-10-11			
M3	C27-14-16	C29-15-18 C38-14-16 C39-14-16 C40-14-16	C30-15-17 C32-17-19	C38-12-13 C38-17 C39-12-13 C39-17 C40-12-13 C40-17 C41-12-13 C41-17	C41-13-15		
M1G	C27-17-22 C27-24-25 C27-27	C29-19-24 C30-20-21 C38-18-20 C39-18-20 C40-18-20	C30-18-19 C30-22-23	C32-20-25			
M1P	C27-29-32	C29-25-26 C41-18-19	C30-24-25	C32-26-27	C41-20		
M4R	C29-12-14 C30-12-14 C32-14-16						

6.1.6 O Processo de Calibração

A elaboração dos conjuntos de análise, descrita na subseção anterior é a fase que antecede a atividade de estimação de parâmetros da TRI e conseqüentemente a calibração dos itens do banco. A estimação dos parâmetros consiste na submissão dos itens a um método de estimação regido por um modelo matemático, geralmente executado por um programa de computador.

Para a estimação dos parâmetros das questões do EPI foi aplicado o método da Estimativa de Máxima Verossimilhança (EMV - subseção 4.5.1) sob o modelo logístico de dois e três parâmetros da TRI ($P_i(\theta)$ - subseção 4.2.3) e pelo emprego do Algoritmo EM (seção 4.5), todos implementados pelo programa XCALIBRE.

Dessa forma, utilizando o programa XCALIBRE, todos os conjuntos de análise descritos na Tabela 6.6 foram formatados segundo o modelo de arquivo de entrada, descrito da Figura 4.10, e submetidos ao processo de estimação. Cada conjunto de análise foi sujeito à estimação dos modelos de 2 e 3 parâmetros (2P e 3P)[5], já que o programa XCALIBRE permite essa opção. Essa submissão aos modelos de 2P e 3P foi estabelecida porque, após a estimação, foi realizada uma análise dos resultados deste processo (por meio dos relatórios finais do XCALIBRE) com o intuito de selecionar qual modelo melhor se ajustava ao comportamento empírico das questões do EPI.

Entre os 31 conjuntos de análise elaborados e submetidos ao processo de estimação, apenas 3 (conjuntos **C39-05-08, C40-05-08 e C41-20**) não foram estimados por nenhum modelo (2P e 3P) por conterem poucas ocorrências (número de respostas para estimação insuficiente). Dos 28

[5] A partir deste ponto, o modelo de 2 parâmetros será referido por 2P e o de 3 parâmetros por 3P.

conjuntos restantes, 13 foram estimados pelo modelo 2P e 15 pelo modelos de 2P e 3P. Assim, a análise dos relatórios finais com intuito de selecionar qual modelo melhor se ajustava aos itens, foi realizada apenas em 15 conjuntos, ou seja, apenas naqueles em que foram estimados para o modelo de 2P e 3P. A Tabela 6.7 mostra o resultado da estimação dos conjuntos de análise, contendo o número de questões que cada conjunto possui e o(s) modelo(s) pelo qual foi estimado.

Tabela 6.7: Resultados da estimação dos conjuntos de análise

Módulos e Partes	Conjuntos de Análise						
M2I	C27-01-07	C29-01-04 C41-01-04	C30-01-05 C39-01-04	C32-01-06 C40-01-04	C38-01-04		
	2P–3P	2P	2P	2P–3P	2P–3P		
	7q.	4q.	5q.	6q.	4q.		
M2A	C29-05-08	C30-06-09	C32-07-10	C38-05-08	C39-05-08	C40-05-08	C41-05-08
	2P	2P	2P–3P	2P–3P	P. Ocorr.	P. Ocorr.	2P–3P
	4q.	4q.	4q.	4q.	4q.	4q.	4q.
M4S	C27-08-13	C29-09-11	C30-10-11	C32-11-13 C38-10-11 C39-10-11 C40-10-11			
	2P–3P	2P–3P	2P–3P	2P			
	5q.	3q.	2q.	3q.			
M3	C27-14-16	C29-15-18 C38-14-16 C39-14-16 C40-14-16	C30-15-17 C32-17-19	C38-12-13 C38-17 C39-12-13 C39-17 C40-12-13 C40-17 C41-12-13 C41-17	C41-13-15		
	2P–3P	2P–3P	2P	2P	2P		
	3q.	4q.	3q.	3q.	4q.		
M1G	C27-17-22 C27-24-25 C27-27	C29-19-24 C30-20-21 C38-18-20 C39-18-20 C40-18-20	C30-18-19 C30-22-23	C32-20-25			
	2P–3P	2P–3P	2P	2P–3P			
	9q.	6q.	4q.	6q.			
M1P	C27-29-32	C29-25-26 C41-18-19	C30-24-25	C32-26-27	C41-20		
	2P–3P	2P	2P	2P	P. Ocorr.		
	4q.	2q.	2q	2q.	1q.		
M4R	C29-12-14 C30-12-14 C32-14-16						
	2P						
	3q.						

Das 122 questões que foram distribuídas nos conjuntos de análise e sujeitas à estimação, 113 foram estimadas, sendo que 42 para o modelo de 2 parâmetros e 71 para o modelo de 2P e 3P. As 9 questões restantes pertenciam aos conjuntos não estimados citados anteriormente. Mais uma vez, vale ressaltar que a análise e seleção do melhor modelo de resposta foram realizadas nas 71 questões estimadas para 2P e 3P.

6.1.7 Resultados da Estimação dos Parâmetros

A análise e seleção do modelo de ajuste dos 71 itens estimados para o modelo de 2 e 3 parâmetros foram realizadas baseadas no relatório final (Figura 4.14 e 4.15) que o programa XCALIBRE produz ao final do processo de estimação.

Como mostrado na Figura 4.14, o relatório final traz vários índices que auxiliam na tarefa de escolha do melhor modelo. A decisão de ajustar um determinado item a um modelo foi efetuada com base na comparação dos valores dos seguintes índices: o erro padrão associado a cada estimativa dos parâmetros e o valor residual padrão de cada item. Um determinado item, estimado para os modelos de 2P e 3P, é escolhido quando os valores dos erros padrões e os valores residuais de um dos módulos forem menores ou próximos de zero.

As Figuras 6.2a e 6.2b mostram um exemplo no qual 7 itens foram estimados para os modelos de 2P e 3P respectivamente. Observe, por exemplo, o Item 2. Os valores dos erros padrões são $0,223$ e $0,247$ para o parâmetro A e $0,171$ e $0,168$ para o parâmetro B, respectivamente para os modelos de 2P e 3P. Os valores residuais padrões são, respectivamente, $0,90$ e $0,37$. Para esse item especificamente, repare que o erro padrão do parâmetro A é menor no modelo de 2P do que no modelo de 3P. Em contrapartida, o valor do erro padrão do parâmetro B e o valor residual padrão são maiores para o modelo 2P do que o 3P. Para este item especificamente, o modelo que melhor se ajusta é o de 3 parâmetros. Observe outro exemplo do Item 3. Os valores dos erros padrões são $0,189$ e $0,348$ para o parâmetro A e $0,245$ e $0,348$ para o parâmetro B, respectivamente para os modelos 2P e 3P. Os valores residuais são $0,39$ e $0,41$, respectivamente. No caso específico desse item, tanto os valores dos erros padrões quanto o valor residual são menores para o modelo de 2P, para qual o item é ajustado.

Uma pergunta que pode surgir neste ponto é: *Por que alguns conjuntos de análise foram estimados para os modelos 2P e 3P e outros só para o modelo 2P?* A explicação para tal ocorrência pode ser entendida se pensarmos no comportamento de cada questão inserida no contexto da avaliação como um todo. Como o processo de estimação dos parâmetros se dá a partir da análise das respostas fornecidas às questões em uma situação de teste, o que pode ser dito é que o volume massivo destes dados carrega um comportamento empírico das questões frente às suas aplicações em uma avaliação. Este comportamento é oriundo de diversos fatores, destacando principalmente entre eles o ambiente de aplicação da avaliação, o nível de comprometimento dos alunos com a avaliação e qual o objetivo da avaliação em si para os alunos. Tais fatores são diretamente refletidos nas respostas dos alunos, que será a base para a estimação, cujo resultado reflete o comportamento dos itens. Pode acontecer que no futuro, uma nova estimação dos parâmetros venha a ser totalmente diferente, visto que com o tempo as situações de aplicação de uma avaliação se alterem e sejam refletidas nos parâmetros estimados.

Matematicamente, tal ocorrência pode ser explicada pelo método de estimação. Os resultados matemáticos de uma tarefa de estimação podem ser dois: a ***convergência*** ou a ***divergência*** do modelo. Quando ao final da estimação temos como resultado as variáveis desejadas estimadas, dizemos que

ITEM PARAMETER ESTIMATES W/STANDARD ERRORS

Item	Lnk	Flg	a	a error	b	b error	c	c error	Resid	Item name
1			0.63	0.168	2.57	0.343	0.00	N/A	1.37	
2			0.80	0.223	-0.33	0.171	0.00	N/A	0.90	
3			1.02	0.189	2.02	0.245	0.00	N/A	0.39	
4			1.01	0.210	0.19	0.143	0.00	N/A	0.36	
5		K	0.60	0.166	2.45	0.334	0.00	N/A	1.07	
6			1.05	0.191	2.00	0.242	0.00	N/A	0.25	
7			1.07	0.220	2.54	0.314	0.00	N/A	0.49	

(a)

ITEM PARAMETER ESTIMATES W/STANDARD ERRORS

Item	Lnk	Flg	a	a error	b	b error	c	c error	Resid	Item name
1		PK	1.06	0.618	3.00	0.587	0.10	***	0.67	
2			0.93	0.247	0.00	0.168	0.08	***	0.37	
3			1.08	0.348	2.29	0.348	0.06	***	0.41	
4			0.96	0.259	0.47	0.172	0.08	***	0.24	
5		PK	1.05	0.621	3.00	0.593	0.10	***	0.65	
6			1.07	0.336	2.24	0.339	0.06	***	0.42	
7			1.11	0.402	2.57	0.393	0.05	0.096	0.85	

(b)

Figura 6.2: Exemplo de valores dos parâmetros estimados para os modelos 2P e 3P

o modelo convergiu e encontrou as estimativas, caso contrário dizemos que o modelo divergiu, ou seja, as variáveis não foram estimadas. Portanto, para todos os casos em que um determinado item não teve seus parâmetros estimados, dizemos que o modelo não convergiu.

Tal fato aconteceu em alguns itens do conjunto de análise. Observe na Figura 6.3 que o item 3 do conjunto de 4 itens possui uma cláusula (*deleted*) informando a sua exclusão do conjunto. Esta cláusula é atribuída pelo programa XCALIBRE durante a fase de estimação e significa dizer que o modelo de estimação não convergiu para aquele item especificamente e por isso ele foi excluído.

O resultado final da comparação dos resultados da estimação dos itens é apresentado na Tabela 6.8. A tabela mostra todas as questões que participaram do processo de estimação (total de 122) distribuídas por módulo e parte, respectivamente. Cada coluna mestra identifica um módulo e parte, que é subdividida em 3 colunas filhas. A primeira coluna (filha) contém o código de identificação da questão. Uma questão identificada pelo código **19.30** representa a questão de número **19** da prova **Código 30**. A segunda coluna contém a quantidade de alunos que responderam a questão (dados de estimação), ou o caractere **X**, significando que a quantidade de alunos foi pequena. A terceira mostra o modelo da TRI (2P ou 3P) que melhor se ajustou à questão, ou os caracteres **NE** e **D**. O caractere **NE** informa que a questão não foi estimada, e o caractere **D** indica que a

```
ITEM PARAMETER ESTIMATES W/STANDARD ERRORS
                          a              b              c
Item Lnk Flg    a      error     b      error     c     error  Resid   Item name
---- --- --- -----    -----   -----    -----   -----   -----  -----   ---------

  1           0.66    0.504    1.39    0.599    0.00     N/A   1.13
  2           0.88    0.457   -1.68    0.585    0.00     N/A   0.41
  3  -- Deleted --
  4           0.98    0.520    1.29    0.447    0.00     N/A   0.27
```

Figura 6.3: Exclusão de um item durante a fase de estimação

questão foi excluída do conjunto de análise (cláusula *deleted*) durante o processo de estimação. As três últimas linhas da coluna mestra, contêm os intervalos dos valores estimados dos parâmetros **A** e **B**, respectivamente, e o número total de questões válidas em cada módulo. Por exemplo, a coluna que identifica módulo e parte **M1P** possui os intervalos do parâmetro A variando de $0,58$ a $1,02$ ($0,58 \leq a \leq 1,02$) e o de B de $-0,87$ a 3 ($-0,87 \leq b \leq 3$) e o total de 10 questões.

Observe que, apesar de 15 conjuntos de análise terem sido estimados para os modelos 2P e 3P, apenas 3 itens (destes conjuntos) ajustaram-se ao modelo 3P. O restante dos itens se ajustou ao modelo 2P, baseado na análise dos relatórios finais produzidos pelo XCALIBRE. Assim, dos 122 itens que participaram do processo de estimação, 17 foram excluídos, ou por terem poucas ocorrências ou por terem sido excluídas durante a estimação; 3 itens foram estimados para o modelo de 3 parâmetros e 100 itens para o modelo de 2 parâmetros. Dessa maneira, temos um total de 103 questões calibradas, o que resulta no aproveitamento de $84,43\%$ ou $15,57\%$ de perda de itens.

Finalizadas as tarefas de organização das provas, divisão do conteúdo em módulos e partes, identificação dos itens semelhantes, elaboração dos conjuntos de análise e calibração dos itens, temos como resultado parcial deste trabalho de mestrado um banco de itens referentes ao Exame de Proficiência em Inglês do ICMC, sensível ao conteúdo e calibrado para o modelo da Teoria de Resposta de Itens de 2 parâmetros. Nesse caso específico, como a maioria dos itens se ajustou ao modelo 2P, todos os itens foram armazenados no banco sob este modelo.

6.1.8 Inclusão de Novas Questões no Banco de Itens

A inclusão de novos itens no banco deve ser realizada de maneira cuidadosa. Para este cenário específico do EPI, é preciso primeiro identificar as questões semelhantes entre as provas que irão fornecer os itens. Em segundo lugar, após ter identificado as questões semelhantes, é necessário formar os conjuntos de análise que serão estimados. Definidos os conjuntos de análise é preciso organizar os arquivos de entrada para o processo de estimação automatizado, para só então submeter tais arquivos (itens) à estimação sob os modelos 2P e 3P. De posse dos resultados da estimação (relatórios) chega o momento de analisar as estimativas, item por item, com o intuito de selecionar aqueles que mais se ajustam a um determinado modelo. Ao fim da análise, teremos os parâmetros

dos itens estimados, que finalmente devem ser armazenados no banco. A seção 6.3 apresenta uma estimação do prazo necessário para a implantação do teste adaptativo no cenário do ICMC.

6.2 Sistema de Gerenciamento do Banco de Itens (SisBI)

A criação do banco de itens, descrita na subseção anterior, seria pífia se não possuíssemos uma ferramenta para a manipulação do mesmo. Com o intuito de aumentar a organização do banco de itens, e conseqüentemente do seu conteúdo (itens), foi proposto o **Sistema de Gerenciamento do Banco de Itens (SisBI)**[6], cujo principal objetivo foi facilitar a manutenção do banco, auxiliando o professor na tarefa de gerenciar os itens de forma simples e otimizada. Dessa forma, o **SisBI** apresenta uma interface, baseada no ambiente Web, de manipulação de questões do EPI voltadas para a avaliação adaptativa. Seus usuários, por sua vez, devem possuir conhecimentos técnicos suficientes das características (parâmetros da TRI, formatos, tipos e etc.) desses itens de maneira a aproveitar todo o potencial desta ferramenta. Uma vantagem que merece ser destacada, é que pelo fato do sistema depender somente de um *Browser* para sua execução, o **SisBI** se torna uma aplicação multiplataforma. Pelo fato de o **SisBI** ser tipicamente um sistema acadêmico, voltado especialmente para professores, seu desenvolvimento e construção seu deu com o uso de ferramentas de domínio público favorecendo assim sua posterior atualização e crescimento conforme seja necessário. A modelagem do sistema foi desenvolvida usando a **UML** (*Unified Modeling Language*), uma linguagem de modelagem de sistemas orientados a objetos que fornece recursos gráficos para visualização e especificação de um sistema (Scott and Fowler, 1998; Cabral and Araújo, 2002). Para a implementação do sistema foi utilizada a linguagem de programação **PHP** (*Personal Home Page* – www.php.net) em conjunto com o banco de dados **MySQL** (www.mysql.com) sobre o *servidor internet* **Apache** (www.apache.org) .

6.2.1 Os Diagramas de Modelagem da UML

A modelagem de um sistema é muito útil por fornecer uma compreensão única do mesmo para todas as pessoas envolvidas no projeto. A UML é uma linguagem de definição e especificação de sistemas orientados a objetos. Segundo Anquetil (2000) a UML é o resultado da unificação de três métodos de modelagem de sistemas de orientação a objetos: o OMT (*Object Modeling Technology*) de Rambanch, o método de Booch e os "casos de uso" de Jacobson. Essa unificação proporcionou bons resultados na modelagem de sistemas, o que provocou sua adoção pela maioria das empresas e fornecedores de *software* (Righetto, 2001). Por meio de diagramas que a UML implementa é possível representar os sistemas sob diversas perspectivas de visualização, documentando assim todas as fases do seu desenvolvimento, desde a modelagem até a implementação (Righetto, 2001). A interface gráfica que a UML utiliza é bastante amigável, de maneira a facilitar o entendimento dos diversos aspectos do sistema. A modelagem do **SisBI** com a UML utilizou somente diagramas referentes à fase de concepção do sistema, a qual contém a definição das classes, as funcionalidades oferecidas,

[6]Disponível no endereço: http://www.nilc.icmc.sc.usp.br/tai/

os usuários e as atividades desempenhadas durante sua execução. Seguindo estas definições, foram criadas todas as classes do sistema e elaborados os seguintes diagramas: Diagrama de Classe, Diagramas de Casos de Uso e Diagrama de Atividades. A seguir veremos a definição e características de cada um deles.

Classe: uma classe é a especificação de um conceito ou situação do mundo real. Essa especificação é sustentada pelos seus atributos e operações. Cada operação pode ser entendida como uma função que manipula os atributos da classe (Anquetil, 2000);

Diagrama de Classes: representa a estrutura estática do sistema. Mostra as classes e suas relações no mundo real, representando o aspecto estrutural do sistema;

Diagrama de Caso de Uso: define as **interações** entre os usuários e o sistema. Mostra os possíveis cenários e quais as atividades que os atores (usuários) podem desempenhar. É muito importante pois é a base de desenvolvimento do sistema;

Diagrama de Atividades: pode ser entendido como diagrama de fluxo de dados (Anquetil, 2000). Mostra como as funções do sistema realizam uma operação permitindo uma análise do **comportamento** do sistema.

6.2.2 Descrição das Interfaces e Funcionalidades do SisBI

O **SisBI** é formado por um único ambiente composto de funcionalidades voltadas para a manutenção de um banco de itens. O único usuário que interage com o sistema é o professor que desempenha as tarefas de inclusão, exclusão, alteração e visualização de questões, bem como a edição de textos e partes (divisão de conteúdo) que podem conter em um exame. Outra funcionalidade desempenhada pelo sistema é a montagem dos gráficos relacionados com a TRI de cada item pertencente ao banco. Tal funcionalidade permite que o professor verifique o comportamento do item, dado os gráficos das funções de resposta e de informação.

Adicionalmente, no futuro pode ser criado o usuário *administrador* possuindo a função de cadastrar novos professores (usuários) que podem também acessar o **SisBI**. A Figura 6.4 mostra o diagrama de caso de uso do **SisBI** com todas as funções que o professor pode desempenhar.

O diagrama de caso de uso mostrado na Figura 6.4 contém todas as funções de gerenciamento do banco de itens. Tais funções são descritas a seguir.

1. **Efetuar Login**: permite o acesso do professor ao sistema. É efetuado por meio do fornecimento do *Nome do Usuário (UserName)* e *Senha* que por sua vez, passam por um mecanismo de validação para garantir a entrada no sistema. Veja Figura B.1 no Apêndice B a tela de entrada do **SisBI**.

2. **Incluir Questão**: permite que o professor realize a inclusão de uma questão que já passou pelo processo de calibração de parâmetros. O professor, depois de acessar o sistema, pode clicar no botão **Incluir** e fornecer os seguintes dados: o módulo e parte aos quais a questão

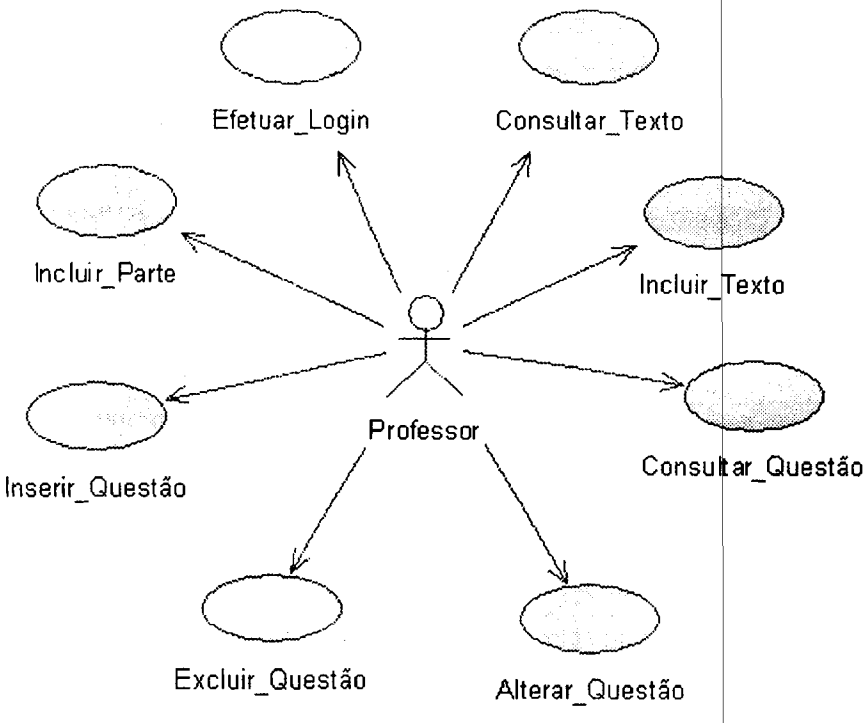

Figura 6.4: Diagrama de Caso de Uso do professor

pertence, o enunciado da questão, o identificador do texto o qual a questão se refere, as alternativas de resposta, a resposta correta, os parâmetros da TRI e a data de criação. Caso o professor não saiba qual identificador do texto (muito comum) associado à questão, existe a opção **Consultar Textos** (vista posteriormente nesta subseção) em que uma nova janela é aberta exibindo todos os textos cadastrados permitindo a sua seleção. Ainda, caso não exista o cadastro do texto, o professor pode incluir um novo texto (botão **Novo Texto**) e voltar à tela de inclusão de questão normalmente. Veja Figura B.2 no Apêndice B o formulário de inclusão de uma nova questão.

3. **Excluir Questão**: possibilita que uma determinada questão, selecionada pelo professor, seja excluída do banco de itens. O professor depois de clicar no botão **Excluir** seleciona o módulo e a parte que a questão pertence. Selecionadas essas opções, são exibidas todas as questões pertencentes ao módulo e parte escolhidos, de maneira que o usuário possa navegar entre elas e escolher qual delas será excluída. A exclusão definitiva acontece quando o usuário aciona o botão **Excluir** ao final da exibição de cada questão. Veja Figura B.3 no Apêndice B a janela de exclusão de questões.

4. **Alterar Questão**: deixa o professor realizar alterações numa determinada questão. Mais uma vez o professor deve selecionar o módulo e parte que a questão está inserida, e a partir dos dados exibidos no formulário o professor insere os novos valores e clica em **Alterar**, o

que efetivamente atualiza a questão no banco. Veja Figura B.4 no Apêndice B a janela de alteração de questões.

5. **Consultar Questão**: possibilita que um professor visualize as questões previamente cadastradas no banco, fornecendo o módulo e parte desejados. São mostradas todas informações das questões, como enunciado, texto a qual se refere, alternativas de resposta, parâmetros da TRI etc. Adicionalmente, o professor pode verificar os gráficos das funções de resposta e de informação (baseado no modelo da TRI) que cada questão fornece, clicando no botão **Exibir Gráficos**. É permitido também que o professor verifique os detalhes do texto (botão **Detalhes do Texto**) referente à questão, como autor, título, URL e conteúdo completo. Veja Figura B.5 no Apêndice B a tela de consulta de questões e a Figura B.6 um exemplo dos gráficos da TRI de uma determinada questão.

6. **Incluir Parte**: permite que o professor inclua uma nova parte sob um determinado módulo. Essa função é muito útil, pois torna o banco flexível à criação de novos conteúdos, e conseqüentemente a atribuição de novas questões a esta nova parte e o crescimento uniforme do banco. O formulário de inclusão de uma nova parte é fornecido quando o professor aciona o botão **Incluir Parte**, sendo necessário o fornecimento de dados como a o módulo a qual pertence, o nome da parte e suas instruções. Veja Figura B.7 no Apêndice B um exemplo do formulário de inclusão de uma parte.

7. **Incluir Texto**: permite a inclusão de um novo texto no banco de maneira que este possa ser associado a diversas questões que farão parte de uma avaliação. Acontece especificamente quando o professor está incluindo uma questão quando aciona o botão **Novo Texto**. Ao clicar neste botão uma nova janela se abre com um formulário, no qual o professor deve fornecer o título, o autor, a URL (caminho da internet) caso houver, e o conteúdo do texto. Veja Figura B.8 no Apêndice B um exemplo de inclusão de um novo texto.

8. **Consultar Textos**: operação que possibilita a consulta de todos os textos cadastrados no banco. Acontece quando o professor deseja associar um texto a uma determinada questão que está sendo inserida clicando no botão **Consultar Textos**.

Neste ponto é preciso discorrer sobre uma padronização realizada na concepção deste banco de itens. As questões do EPI variam muito quanto à forma, dependendo do conteúdo (módulo e parte) ao qual pertence. Assim, temos questões que se referem a um texto propriamente dito, sendo este uma Introdução de um artigo (**Módulo 2 parte Introduction**) ou um *Abstract* (**Módulo 2 parte Abstract**), ou ainda, questões que se referem a uma única oração (**Módulo 1 parte GAP**), sendo esta parte de um texto geralmente extenso. Convencionou-se que quando nos referimos a um "texto" no contexto deste banco de itens, não quer dizer que este seja um artigo propriamente dito, com todo seu conteúdo, mas sim, uma forma geral de referência a um texto (seja este uma oração, citação, *Introdução* ou *Abstract*) que está associada a uma questão quando esta é selecionada numa avaliação.

No Apêndice A é apresentado um exemplo de uma prova do EPI. Para uma maior compreensão e entendimento do **SisBI**, o Apêndice B mostra todas as telas e janelas que representam a sua interface, bem como a sua modelagem, contendo a descrição das classes e os diagramas: de classes, de caso de uso, e de atividades desenvolvidos sob a notação UML.

6.3 Discussão dos Resultados

Ao longo desta dissertação foi possível conhecer e estudar as teorias que regem a avaliação informatizada, e em destaque os testes adaptativos. Dentre as características que delineiam a avaliação adaptativa, os métodos de seleção de itens, os modelos de estimação de habilidade do estudante, a calibração do banco de itens e os critérios de parada da avaliação merecem atenção especial.

Neste contexto, há dois principais obstáculos que dificultam a implantação dos testes adaptativos: 1) a necessidade de um estudo empírico de larga escala para o processo de calibração dos itens e 2) a criação de testes multidimensionais, com balanceamento de conteúdo, de forma a medir diferentes habilidades em um mesmo domínio de conhecimento (Huang, 1996). Entretanto, em grandes centros de educação especializada, como é o caso do ETS (*Educational Testing Service*), que gerencia o exame TOEFL, tais obstáculos podem ser superados, visto que nestes existe uma larga experiência vinda de anos de trabalho dedicados às avaliações e grandes volumes de dados para realizar a calibração dos itens. Assim, de maneira geral, o processo de implantação da avaliação adaptativa no ambiente de pequenas instituições, seja empresarial ou acadêmica, torna-se difícil e demorado.

Para tentar superar estes obstáculos e implantar os testes adaptativos em pequenas instituições, diversas abordagens de seleção de itens e de estimação de parâmetros e habilidade foram desenvolvidas (veja seção 6.5). Conforme pôde ser visto na Figura 4.1 no Capítulo 4, a tarefa de estimação dos parâmetros dos itens no processo de calibração do banco de itens deve seguir quatro passos, dentre eles a análise de um conjunto de amostras de respostas aplicados anteriormente. Portanto, pode-se dizer que a metodologia de calibração de itens descrita na Figura 4.1 é a mais adequada, visto que seus resultados são fiéis ao comportamento dos itens em uma avaliação, pois utiliza-se dados reais.

Deste modo, o processo de calibração do banco de itens desenvolvido neste trabalho de mestrado seguiu esta metodologia, mesmo que para isso fosse conseguido um pequeno número de itens calibrados durante o desenvolvimento deste projeto. Portanto, a adoção de tal metodologia é justificada pelo fato de que a tendência do banco de itens é crescer com o tempo, dado que no cenário do ICMC novos EPI's acontecem a cada semestre, de maneira que as questões pertencentes a estes exames sofrerão normalmente o processo de calibração e serão inseridas no banco.

No entanto, tendo em vista que o banco de itens desenvolvido para o EPI implementa o balanceamento de conteúdo, seu crescimento linear depende diretamente do crescimento de cada módulo e parte (subdivisão de conteúdo) incluídos no banco. Além disso, em testes adaptativos sensíveis ao conteúdo (caso do EPI do ICMC–USP), cada módulo é avaliado como sendo um teste

adaptativo à parte, ou seja, é necessário que cada módulo de conteúdo cresça de acordo com a sua importância no contexto geral da avaliação.

Dessa maneira, considerando a regra dita por Olea et al. (1999) de que o número total de itens de um banco deve ser dez vezes maior do que o número de itens que um teste possui, e atentando para o fato de que cada módulo representa uma avaliação, a Tabela 6.9 mostra a perspectiva do tempo necessário para que o banco de itens atual esteja adequado para o uso em uma avaliação adaptativa no cenário do ICMC. A primeira coluna representa os módulos e suas respectivas partes. Tomado como base os dados advindos do exame de setembro de 2001, observe que a segunda coluna representa a quantidade de questões correspondentes a cada módulo e parte das provas do EPI. Assim, selecione como exemplo a linha que representa o módulo **M1G**. Dado que o número de questões da prova é 3, a quantidade mínima necessária é de 30 questões ($3 \cdot 10 = 30$), e ainda, visto que a quantidade de questões atualmente pertencente a esse módulo é de 22 (Tabela 6.9) ficam restando para completar as 30 questões (necessárias), o total de 8 ($30 - 22$). Assim, já que a cada semestre teremos no mínimo 3 provas, podemos contar que a cada semestre somaremos 9 questões adicionais ($3 \cdot 3 = 9$) e, dessa maneira, calcular a perspectiva de tempo para conseguirmos o mínimo de questões necessário. O cálculo da perspectiva de tempo se dá por meio da divisão da quantidade de questões restantes (quinta coluna) pela quantidade de questões fornecidas a cada semestre (sexta coluna). Nesse caso específico, como precisamos de apenas 8 questões para completar as 30 desejadas, e a cada semestre somaremos 9 questões, o tempo gasto será de 0,88 semestre (menos de 1 semestre). Porém, esse resultado não é linear para todos os módulos, pois veja que para o módulo e parte **M4R** (ultima linha) o tempo necessário é de 3 semestres.

Portanto há a perspectiva de que em três semestres, isto é, em julho de 2003, já tenhamos um banco de itens calibrados e prontos para o uso em uma avaliação diagnóstica. Tal perspectiva será verdadeira se, é claro, todos os itens utilizados nas provas sejam aproveitados e que não sejam descartados durante o processo de calibração.

Existe uma desvantagem que deve ser considerada quando se aplicam testes adaptativos (i.e. baseados em *Testlets* e balanceamento de conteúdo). De acordo com estudos empíricos de Kingsbury and Zara (1991), que tratam da comparação destes métodos de seleção de itens, concluiu-se que testes adaptativos que utilizam o balanceamento de conteúdo possuem de 5% a 11% mais itens administrados aos alunos, enquanto que o uso de *Testlets* aumenta o teste entre 43% a 104%.

6.4 Modelagem do Sistema de Avaliação Diagnóstica TAEPI

A criação do banco de itens e o desenvolvimento do **SisBI**, permitindo a sua gerência, estimularam o trabalho de modelagem de um sistema adaptativo de avaliação diagnóstica da proficiência em inglês do programa de mestrado do ICMC-USP, batizado de **TAEPI** – ***Testes Adaptativos para o Exame de Proficiência em Inglês***. O principal objetivo dessa modelagem é descrever os ambientes e funcionalidades referentes a um teste adaptativo voltado para este exame de proficiência em inglês, permitindo a implementação e aplicação do mesmo no cenário do ICMC-USP. Assim como

foi feito na modelagem do **SisBI**, o mapeamento das classes, relacionamentos e funções do **TAEPI** também foram desenvolvidas em **UML** (*Unified Modeling Language*), cujo enfoque foi a definição de diagramas necessários para a concepção do sistema. Por meio dos diagramas da **UML** foi possível representar o sistema sob diversas perspectivas de visualização, facilitando o entendimento do seu comportamento. A documentação completa e todos os diagramas da modelagem desse sistema estão descritos no Apêndice C. As funcionalidades específicas oferecidas pelo **TAEPI** variam de acordo com os três tipos de usuários (Administrador, Professores ou Alunos) que podem utilizá-lo. Deste modo, pode ser entendido que o sistema disponibiliza diferentes *ambientes de operação*, dependendo somente do usuário que vai utilizar. A Figura 6.5 ilustra o diagrama de caso de uso geral do **TAEPI**.

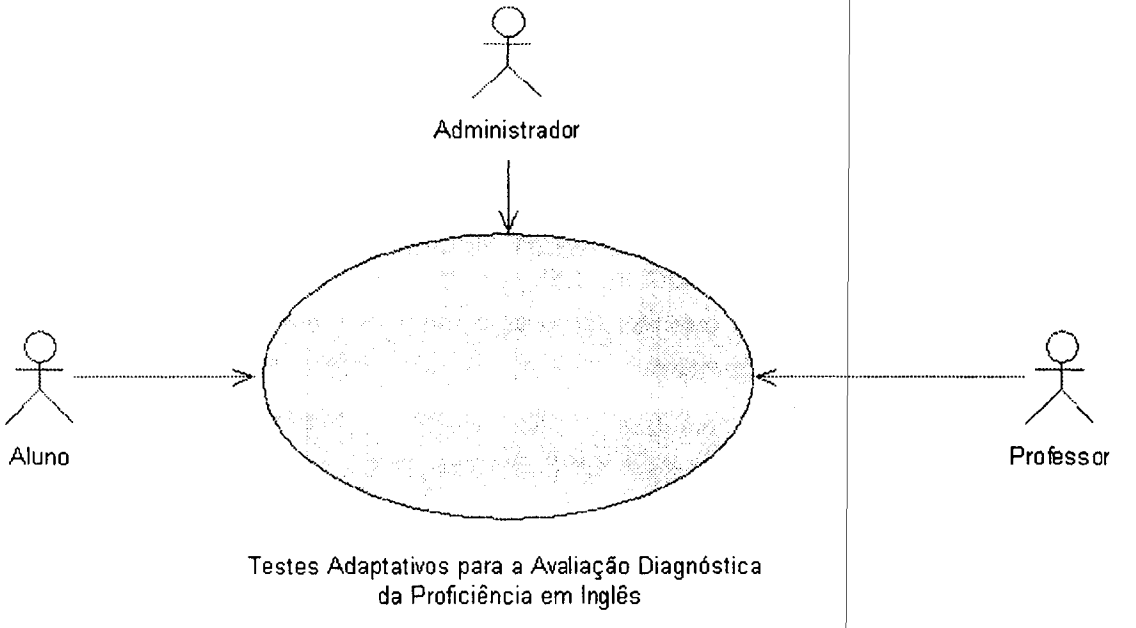

Figura 6.5: Diagrama de Caso de Uso geral do TAEPI

Dessa maneira, o **TAEPI** é composto por três ambientes distintos de operação, sendo estes, um ambiente para o *Administrador*, um para os *Professores* e outro para os *Alunos*. Cada ambiente é formado por tarefas (funcionalidades) que são de responsabilidade de cada usuário pertencente àquele ambiente, as quais seguem:

Administrador : responsável pelo cadastro de professores;

Professores : responsáveis principalmente pela criação e especificação de exames adaptativos que serão realizados pelos alunos, além da possibilidade de consulta dos mesmos;

Alunos : são responsáveis pelo seu próprio cadastro e sua principal tarefa é realizar um exame adaptativo especificado por um professor, além de poder consultar seus resultados e gabaritos de exames já realizados.

A seguir serão descritas as funcionalidades de cada ambiente.

6.4.1 O Ambiente do Administrador

O administrador, no contexto do sistema **TAEPI**, possui uma única função, que é o cadastramento do professor. A Figura 6.6 mostra o diagrama de caso de uso que descreve essa função.

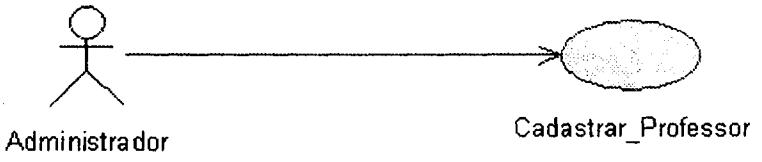

Figura 6.6: Diagrama de Caso de Uso do Ambiente do Administrador

O cadastramento de professores possibilita que professores interessados em utilizar o **TAEPI** criem e especifiquem exames adaptativos. Este cadastramento é iniciado quando o administrador acessa o sistema (função: *Efetuar_Login*) e seleciona a função de *Cadastrar_Professor* fornecendo as seguintes informações: nome do professor, *username*, senha, e data. A Figura 6.7 mostra o diagrama de atividades[7] que representa essa funcionalidade.

6.4.2 O Ambiente do Professor

Este ambiente contém todas as tarefas que um determinado professor (previamente cadastrado) pode realizar dentro do **TAEPI**. Os relacionamentos entre tais tarefas e o professor estão ilustrados na Figura 6.8, que representa o diagrama de caso de uso deste ambiente.

Conforme pode ser visto na Figura 6.8 as seguintes tarefas podem ser desempenhadas pelo professor:

Efetuar Login : corresponde à permissão ou não de acesso do professor ao sistema. Acontece quando um professor abre a tela de acesso do sistema e fornece um *username* e *Senha* que são verificados;

Criar Exame : possibilita a um determinado professor criar um exame adaptativo que será disponibilizado aos alunos. Dentre as informações que o professor deve fornecer para a criação de um exame, estão as partes (módulo e parte) que irão compor o exame e os pesos relacionados a cada uma delas (utilizados para calcular o resultado final);

Especificar Critérios de Aplicação do Exame : após a criação de um exame, o professor pode especificar seus critérios de acordo com as partes inseridas no mesmo. Tal especificação é bastante flexível e permite a definição de valores máximo e mínimo de habilidade do exame, a quantidade de questões que se deseja administrar e ainda a quantidade de *testlets* (caso alguns dos módulos inseridos contenha) requeridos, bem como os valores máximo e mínimo

[7] A atividade "Validar Dados Lidos" corresponde a uma validação sintática dos dados. O mesmo vale para todos diagramas no Apêndice C que tal atividade aparece.

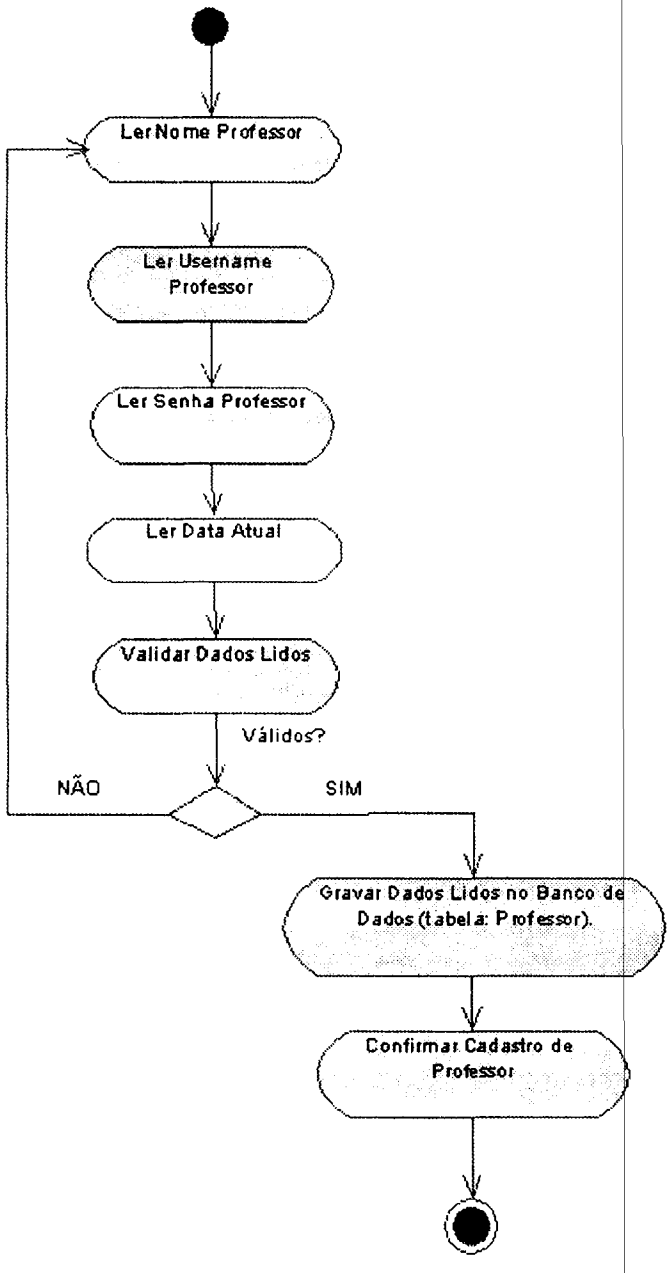

Figura 6.7: Diagrama de Atividades que representa a função Cadastrar Professor

da habilidade de cada *testlet*. Para especificar os critérios de aplicação, o professor deve fornecer um número de identificação do exame previamente cadastrado;

Consultar Resultado Global : a consulta do resultado global acontece quando o professor fornece a identificação do(s) aluno(s) ou um determinado exame, permitindo que um professor verifique os resultados de um determinado exame que já foi realizado. O **TAEPI** oferece as opções de verificar o resultado individual ou coletivo dos alunos. Caso esta última opção seja

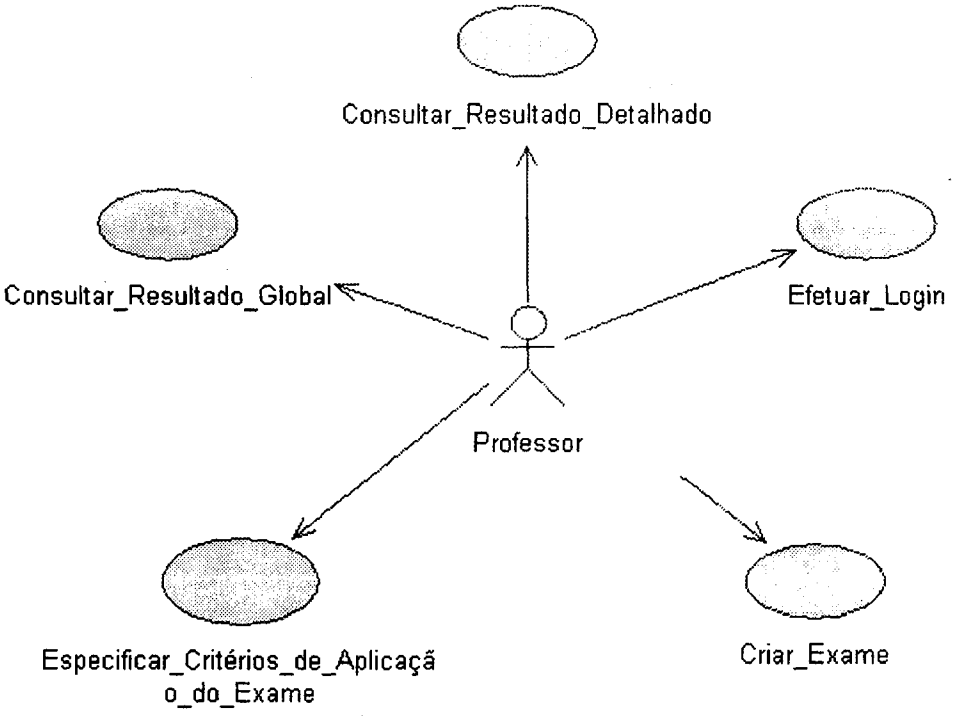

Figura 6.8: Diagrama de Caso de Uso do Ambiente do Professor

selecionada, o professor também tem acesso à média do valor da habilidade e à pontuação alcançada em cada parte pelo aluno;

Consultar Resultado Detalhado : o professor, por meio desta função, pode verificar o resultado detalhado de um determinado aluno em um exame já realizado. O resultado detalhado contém informações de todas as questões fornecidas ao aluno durante um exame adaptativo, juntamente com o módulo e parte a que pertence, a resposta fornecida, os cálculos parciais dos valores de habilidade e pontuação obtida. O professor deve fornecer as identificações do exame e do aluno para obter estes resultados.

A Figura 6.9 mostra o diagrama de atividades da função *Criar Exame*. Os demais diagramas de atividades desse ambiente podem ser vistos no Apêndice C.

6.4.3 O Ambiente do Aluno

O ambiente do aluno possui todas as funções para resolução de exames adaptativos previamente cadastrados pelo professor. As funcionalidades disponibilizadas pelo ambiente do aluno estão ilustradas na Figura 6.10. Observe que além da possibilidade de realizar um exame (função *Realizar_Exame*) os alunos também podem se cadastrar, consultar resultados de exames e consultar o gabarito de um determinado exame realizado anteriormente.

A Figura 6.10 representa o diagrama de caso de uso do ambiente do aluno, em que são descritas todas as atividades que os alunos podem desempenhar no **TAEPI**. Uma explicação de cada uma delas é feita a seguir:

Cadastrar Aluno : possibilita o cadastramento de alunos que irão realizar um determinado exame de proficiência em inglês sob o método adaptativo informatizado. O cadastro é realizado pelo próprio aluno quando este acessa pela primeira vez o sistema e escolhe a opção *cadastrar*, fornecendo algumas informações como nome, *username*, senha entre outros. O cadastro de aluno permite ao mesmo realizar qualquer exame que esteja disponível no sistema;

Efetuar Login : permite o acesso do aluno ao sistema de maneira que este possa desempenhar as tarefas mostradas na Figura 6.10. Semelhante ao acesso do professor, consiste no fornecimento do *username* e senha que são verificados;

Realizar Exame : é a principal funcionalidade que este ambiente proporciona, pois permite que um aluno previamente cadastrado realize um exame especificado pelo professor. Resumidamente, a realização de um exame é a execução pura da avaliação adaptativa, e nele contém todas os procedimentos de seleção de itens, estimação de habilidade e verificação dos critérios de parada que permitem a tarefa de adaptação exercida pelo exame. Acontece quando um determinado aluno acessa o sistema e seleciona um exame para realizar, de forma que a cada questão fornecida, o sistema atualiza as informações sobre aquele exame e aluno específicos, como valor da habilidade, módulos e partes avaliados e pontuação obtida;

Consultar Resultado : permite que um determinado aluno verifique os resultados de um exame já realizado por ele. O aluno deve fornecer a identificação do exame realizado e sua senha de acesso, de maneira que após a autenticação, o sistema exiba as seguintes informações: número do exame, módulos e partes avaliadas, quantidade de questões feitas, valor da habilidade e pontuação finais;

Consultar Gabarito : depois que um determinado aluno realizou um exame, é permitido que ele verifique o gabarito do mesmo, de maneira que ele possa conhecer suas falhas e erros cometidos. Assim, o aluno fornece o número do exame realizado e recebe como resposta as informações de identificação da parte, a questão fornecida, a resposta dada e a resposta correta. Entretanto, tal função deve ser muito planejada antes de ser disponibilizada, pois ela expõe os itens do banco, o que pode ser muito prejudicial para uma avaliação adaptativa formal.

A Figura 6.11 mostra o diagrama de atividades correspondente à tarefa *Realizar Exame* que é a principal função desempenhada por este ambiente. Observe que nessa função estão representados os principais procedimentos inerentes aos testes adaptativos como por exemplo a especificação da habilidade inicial, e a aplicação dos diferentes conteúdos do banco, contendo ou não *testlets*.

6.4.4 A Importância do TAEPI

A modelagem do protótipo do sistema de avaliação adaptativa (TAEPI) procurou explorar todas as vantagens e potencialidades proporcionadas pelo balanceamento de conteúdo do banco de itens. Conforme pode ser visto nos diagramas da modelagem no Apêndice C, é possível criar testes adaptativos contendo um único ou vários módulos pertencentes ao banco de itens e, conseqüentemente, especificar seus critérios de parada de maneira individualizada. Portanto, com intuito de se beneficiar da sensibilidade de conteúdo do banco de itens, a modelagem do sistema **TAEPI** prevê a criação de testes adaptativos multidimensionais, por meio da especificação de diversos módulos e partes que podem estar presentes em uma avaliação, além da definição de variados critérios de parada para cada conteúdo inserido em um exame adaptativo. Dessa maneira, podemos dizer que a liberdade de especificar exames contendo um ou vários "conteúdos" (representados pelos módulos e partes) e em cada um destes determinar seus critérios de parada, dá a possibilidade da criação de exames totalmente flexíveis e de acordo com a necessidade de uma avaliação. Logo, o que torna o sistema **TAEPI** interessante é a autonomia de produzir exames maleáveis, condizentes com a realidade do ambiente de aplicação do mesmo.

Observe que a possibilidade de criar tais exames é especialmente importante para a avaliação diagnóstica (à qual esse sistema se destina), pois podem existir situações variadas no contexto do exame no qual se queira medir diferentes habilidades dos alunos, de forma individualizada ou não. Suponha que queiramos avaliar apenas duas habilidades do EPI, sendo estas representadas pelas partes PURPOSE (**M1P**) e REVIEW (**M4R**). Podemos então criar um exame contendo apenas os módulos e partes **M1P** e **M4R** respectivamente, definindo para cada um deles seus critérios de parada desejados. Perceba que, na configuração deste exame, apenas avaliamos os conhecimentos referentes às partes de PURPOSE e REVIEW sem nos preocuparmos com as outras.

Outro exemplo que pode ocorrer no contexto do EPI é priorizar a avaliação de alguns módulos ditos mais "importantes" em detrimento de outros. Por exemplo, as partes INTRODUCTION e ABSTRACT contidas no módulo 2 (**M2I** e **M2A** respectivamente), que tratam da estruturação de textos (veja Tabela 1.1), podem pertencer a uma única avaliação, de maneira que se o diagnóstico do aluno for satisfatório nessa avaliação, ele pode ser dispensado de outras. Isso porque a solução das questões desse módulo contempla o conhecimento inerente a outros, já que se um aluno consegue estruturar bem um determinado texto, com certeza ele conhece as convenções da língua e as estratégias de escrita, contidas nos módulo 1 e 4, respectivamente.

Assim, a grande vantagem do **TAEPI** está na capacidade de criação de exames adaptativos flexíveis e condizentes com as necessidades atuais do domínio do conhecimento ao qual pertence, se traduzindo no alcance dos objetivos que a avaliação diagnóstica propõe.

6.5 Trabalhos Relacionados

Neste ponto, vale destacar os trabalhos desenvolvidos por Huang (1996) e Rios et al. (1998), que tiveram como resultado o **Algoritmo CBAT-2** e o **Sistema SIETTE**, respectivamente. Nesses

trabalhos os autores optaram por não aplicar o processo de calibração propriamente dito e sim por definir explicitamente os parâmetros dos itens. O parâmetro de *Discriminalidade* (A) foi fixado no valor constante de 1,2 porque, segundo um estudo de Kingsbury and Weiss, (1979) citado por Huang (1996), é um valor próximo da média dos parâmetros dos bancos de itens. O parâmetro de *Dificuldade* (B) foi definido com um valor entre 0 e 1 pelos projetistas do teste, com base em seus conhecimentos sobre cada item. Por sua vez, o parâmetro de *Adivinhação* (C) foi diretamente definido pela seguinte divisão: *1 / número de alternativas do item.* Assim, uma questão que possui 4 alternativas de resposta terá o valor de c igual a 0,25. Tais atribuições dispensaram a condução de um estudo empírico sobre as questões para a atribuição dos parâmetros. Entretanto, tanto no **CBAT-2** quanto no **SIETTE**, à medida que as questões são usadas em uma avaliação, acontece a recalibração do parâmetro b, de acordo com as respostas fornecidas em diferentes testes segundo a seguinte fórmula:

$$diff_i = \frac{20 \cdot init_i + \Theta_i}{20 + R_i + W_i}$$

em que $init_i$ é o valor do parâmetro da *dificuldade inicial* da questão i, a constante 20 é um fator de normalização, R_i e W_i são a quantidade de vezes que a questão i foi respondida de forma correta e incorreta, respectivamente. Por sua vez, Θ_i é um acumulador de dificuldade da questão i e é equivalente a:

$$\Theta_i = \sum_{j=1}^{n} k_j \cdot f(\theta'_j)$$

em que n é o número de respostas fornecidas ($n = R_i + W_i$), θ_j é o valor da habilidade do aluno que forneceu a *j-ésima* resposta para i; k_j representa a resposta da questão (correta ou incorreta) e $f(\theta_j)$ denota uma função linear que converte os valores de habilidade θ para a escala de 0 a 1 (Huang, 1996). Outros trabalhos que podem ser citados são o **COMTEX** (Olea et al. 1999) e o **CAT-ASVAB** (Sands et al. 1997). O **COMTEX** é um sistema de avaliação adaptativa de conhecimentos de Geografia, História e Ciências Sociais para alunos do ensino médio. A calibração do banco de itens do **COMTEX** foi desenvolvida com a aplicação dos modelos logísticos da TRI sob o método da Estimativa de Máxima Verossimilhança (EMV), por meio da utilização do XCALIBRE, exatamente como foi desenvolvido no nosso trabalho. Por sua vez o **CAT-ASVAB**[8], que é uma bateria de testes adaptativos vocacionais do Departamento de Defesa Americano, também aplicou os modelos logísticos sob a EMV para a calibração do banco de itens utilizando o programa LOGIST (Sands et al. 1997). Igualmente importante no contexto dos testes adaptativos são os métodos de estimação de habilidade dos alunos (ocorridos durante o teste) e os métodos de seleção de itens. Apesar de ser difícil de implementar, segundo Kingsbury and Zara, (1989), o método *Bayesiano* (Owen, 1975) foi utilizado para estimar os valores da habilidade dos alunos tanto no **CBAT-2** quanto

[8]Computer Adaptive Testing - Armed Services Vocational Aptitude Battery

no **CAT-ASVAB**. A estratégia de seleção de itens nesses sistemas foi implementada utilizando o procedimento de Máxima Informação aplicando o modelo logístico da Estimativa de Máxima Verossimilhança. O **SIETTE** por sua vez, adotou os modelos Bayesianos tanto para a estimação de habilidade quanto para a seleção de itens, mesmo que esta última não seja aconselhada (Kingsbury and Zara, 1989) conforme descrito na subseção 4.4.2. Já o **COMTEX** adotou a EMV e os modelos logísticos em todos os seus processos de desenvolvimento, desde a calibração do banco (visto anteriormente) até à estimação de habilidade. Os critérios de parada também assumem formas variadas nesses programas. No **COMTEX** o critério de parada adotado foi o número de itens fornecido, que devia ser de no mínimo 5 e no máximo 35 itens. No sistema **CBAT-2** foram aplicados dois critérios de parada: o nível de habilidade do aluno associado a cada conteúdo do exame e um número mínimo de questões que deve ser fornecido de cada módulo. O **SIETTE** estabeleceu três critérios: número máximo e mínimo de questões em cada conteúdo e o nível de habilidade do aluno em cada conteúdo estabelecido pelo projetista do teste. O **CAT-ASVAB** por sua vez é terminado depois que um aluno realizou um número fixo de itens ou o tempo destinado ao teste tenha se esgotado. A Tabela 6.10 resume os critérios de parada e os métodos de estimação e seleção de itens de cada sistema.

Apesar de alguns sistemas de testes adaptativos usarem modelos bayesianos para seus métodos de seleção e estimação de parâmetros, os modelos logísticos sob a Estimativa de Máxima Verossimilhança e os procedimentos de máxima informação, possuem uma ampla documentação e são certamente mais simples e fáceis de implementar. Além disso, a ausência da variância nos resultados (que está presente no modelo bayesiano) os torna mais precisos e condizentes com a realidade da avaliação e dos alunos (Olea et al. 1999; Hamblenton and Swaminathan, 1985 e Kingsbury and Zara 1989). Assim, trazendo esta discussão para a avaliação diagnóstica no contexto do EPI (domínio do nosso exame), talvez a configuração mais indicada para uma posterior implementação dos testes adaptativos seja a adoção destes modelos, tanto para a calibração do banco (estimação de parâmetros), como para seleção de itens e a estimação de habilidade.

Tabela 6.8: Resultado final da estimação dos parâmetros com os respectivos ajuste dos itens ao modelo da TRI

N./M.	M1G			M1P			M2I			M2A			M3			M4S			M4R		
	Q	NA	M	Q.	NA	M	Q	NA	M	Q.	NA	M	Q.	NA	M	Q.	NA	M	Q.	NA	M
1	17.27	X	D	29.27	54	2P	01.27	102	2P	05.29	73	2P	14.27	63	2P	08.27	71	2P	12.29	116	2P
2	18.27	59	2P	30.27	53	2P	02.27	89	3P	06.29	73	2P	15.27	63	2P	09.27	70	2P	13.29	114	2P
3	19.27	58	3P	31.27	51	2P	03.27	89	2P	07.29	72	2P	16.27	63	2P	10.27	70	2P	14.29	112	2P
4	20.27	58	2P	32.27	50	2P	04.27	87	2P	08.29	72	2P	15.29	82	2P	11.27	68	2P	$0,81 \leq a \leq 0,99$		
5	21.27	58	2P	25.29	70	2P	05.27	83	2P	06.30	X	D	16.29	82	2P	13.27	69	2P	$1,63 \leq a \leq 3$		
6	22.27	58	2P	26.29	68	2P	06.27	82	2P	07.30	27	2P	17.29	82	2P	09.29	65	2P	3 q.		
7	24.27	57	2P	24.30	22	2P	07.27	75	2P	08.30	27	2P	18.29	59	2P	10.29	65	2P			
8	25.27	57	2P	25.30	21	2P	01.29	98	2P	09.30	26	2P	15.30	X	D	11.29	63	2P			
9	27.27	56	2P	26.32	24	2P	02.29	93	2P	07.32	31	2P	16.30	51	2P	10.30	26	2P			
10	19.29	104	2P	27.32	21	2P	03.29	92	2P	08.32	29	2P	17.30	51	2P	11.30	26	2P			
11	20.29	103	2P	20.41	X	NE	04.29	90	2P	09.32	28	2P	12.38	X	D	11.32	50	2P			
12	21.29	103	2P	$0,58 \leq a \leq 1,02$			01.30	31	2P	10.32	28	2P	13.38	X	D	12.32	50	2P			
13	22.29	79	2P	$-0,87 \leq b \leq 3$			02.30	31	2P	05.38	15	2P	1738	35	2P	13.32	50	2P			
14	23.29	57	2P	10 q.			03.30	31	2P	06.38	15	2P	13.41	X	D	$0,44 \leq a \leq 1,01$					
15	24.29	57	2P				04.30	31	2P	07.38	15	2P	14.41	14	2P	$-0,87 \leq b \leq 3$					
16	18.30	X	D				05.30	X	D	08.38	15	2P	15.41	14	2P	13 q.					
17	19.30	23	2P				01.32	42	2P	05.39	X	NE	$0,58 \leq a \leq 1,18$								
18	22.30	22	2P				02.32	40	2P	06.39	X	NE	$-1,34 \leq b \leq 3$								
19	23.30	22	2P				03.32	39	2P	07.39	X	NE	12 q.								
20	20.32	25	2P				04.32	39	2P	08.69	X	NE									
21	21.32	25	2P				05.32	38	2P	05.40	X	NE									
22	22.32	24	2P				06.32	33	3P	06.40	X	NE									
23	23.32	24	2P				01.38	15	2P	07.40	X	NE									
24	24.32	24	2P				02.38	15	2P	08.40	X	NE									
25	25.32	X	D				03.38	X	D	05.41	14	2P									
26	$0,55 \leq a \leq 1,18$						04.38	15	2P	06.41	14	2P									
27	$0,60 \leq b \leq 3$						$0,59 \leq a \leq 1,11$			07.41	14	2P									
28	22 q.						$-2,10 \leq b \leq 3$			08.41	14	2P									
29							24 q.			$0,60 \leq a \leq 1,8$											
30										$-0,64 \leq b \leq 3$											
31										19 q.											

Tabela 6.9: Perspectiva de tempo para o preenchimento ideal do banco de itens do EPI

	N. de questões da Prova	N. de questões mínimo necessário	N. de questões Atual	N. de questões restantes necessário	N. de questões fornecidas por semestre (3 provas no mín.)	Perspectiva de Tempo (em semestres)
M1G	3	3 x 30=30	22	30 - 22=8	3 x 3=9	8 / 9 = 0,88
M1P	3	30	10	20	9	2,22
M2I	4	40	24	16	12	1,33
M2A	4	40	19	21	12	1,75
M3	6	60	12	48	18	2,66
M4S	3	30	13	17	9	1,88
M4R	3	30	2	27	9	3

Tabela 6.10: Critérios de parada, métodos de estimação e seleção de itens nos sistemas adaptativos avaliados

Sistema	Método de Seleção	Método de Calibração dos Itens	Método de Estimação da Habilidade	Critérios de Parada
CBAT-2	Máxima Informação Logístico (EMV)	Pré-determinado	Bayesiano	Nível de Hab./ Número Mín. de Itens
SIETTE	Bayesiano	Pré-determinado	Bayesiano	Nível. de Hab./ Número Mín./ Máx. de Itens
COMTEX	Máxima Informação Logístico (EMV)	Automático Logístico (XCALIBRE)	EMV	Número de Itens
CAT-ASVAB	Máxima Informação Logístico (EMV)	Automático Logístico (LOGIST)	Bayesiano	Número de Itens/ Limite de Tempo

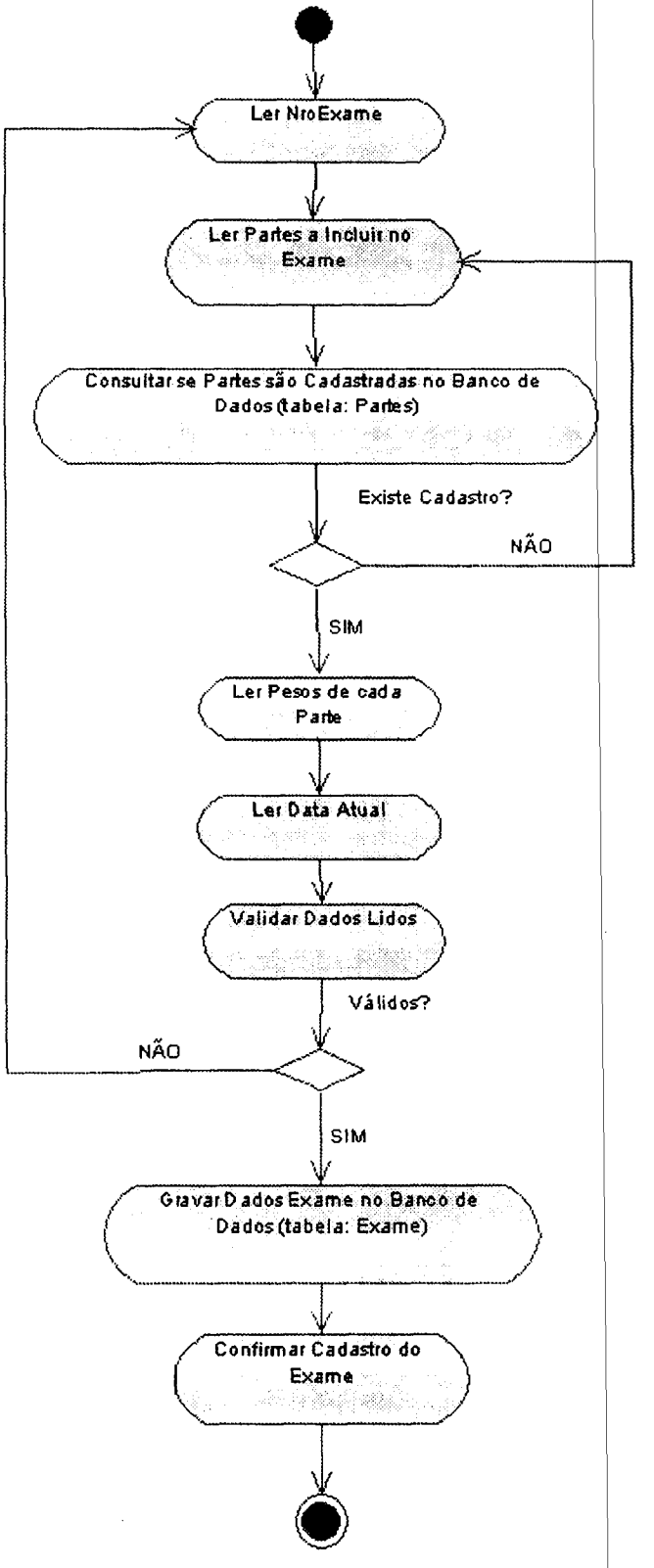

Figura 6.9: Digrama de Atividades da função Criar Exame

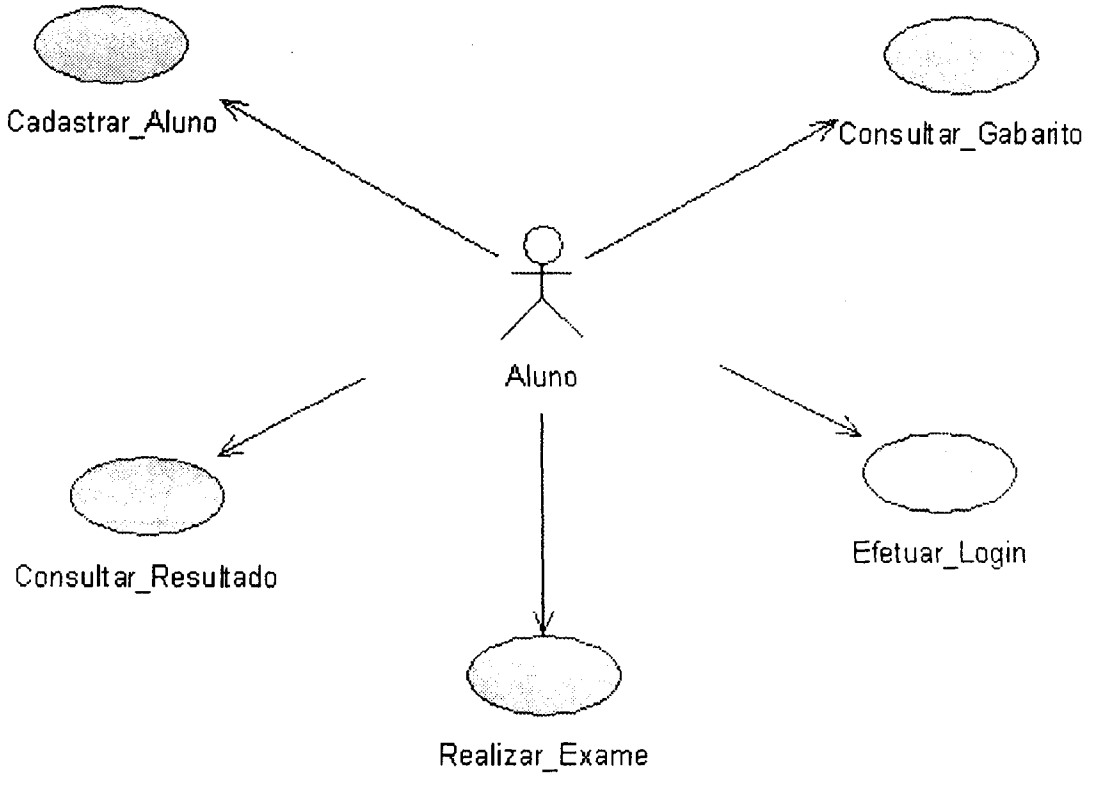

Figura 6.10: Diagrama de Caso de Uso do Ambiente do Aluno

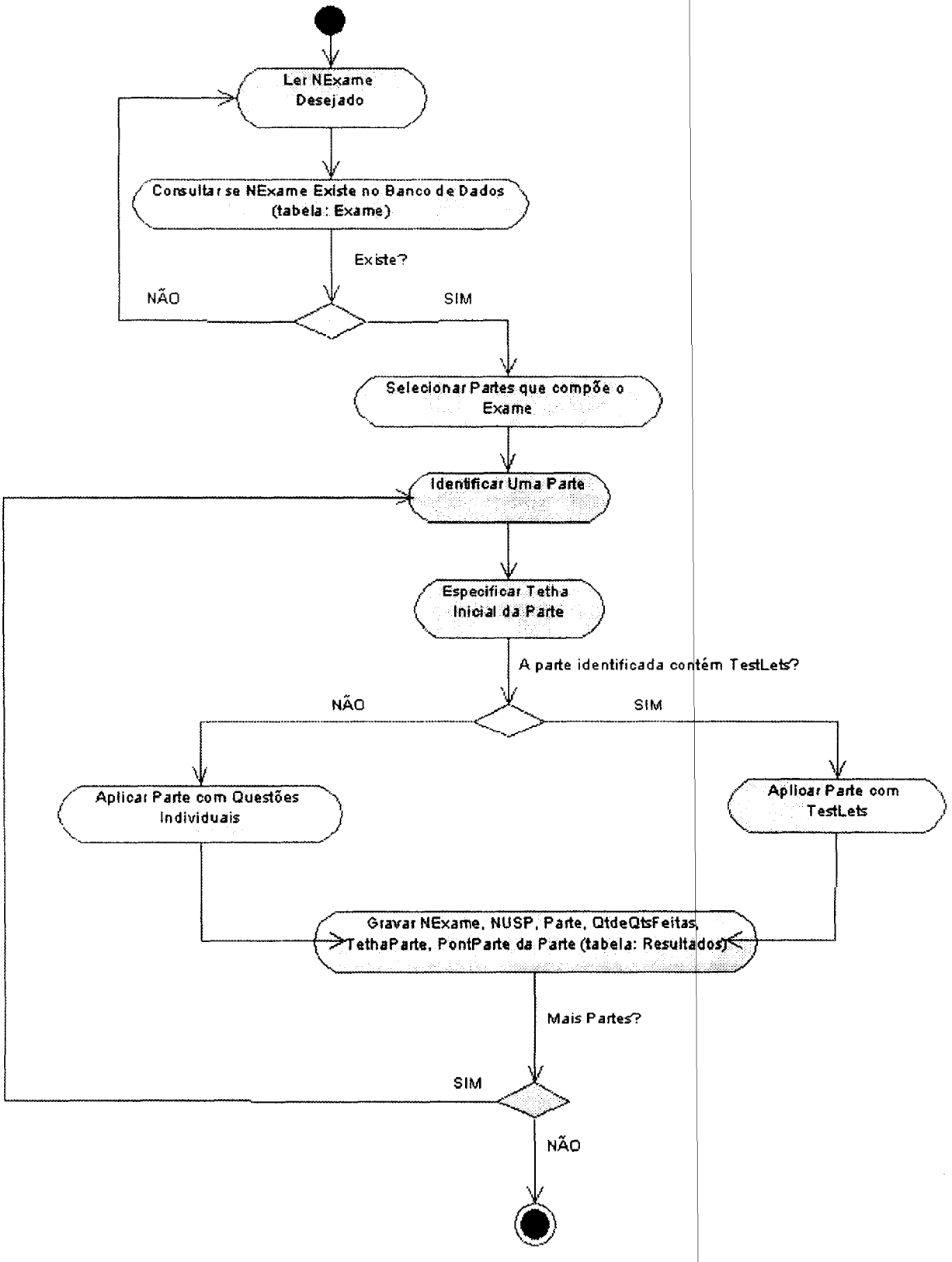

Figura 6.11: Diagrama de Atividades da função Realizar Exame

CAPÍTULO 7

Conclusão e Trabalhos Futuros

Esta dissertação apresenta a teoria dos testes adaptativos, destacando principalmente seus formalismos, métodos, modelos, programas relacionados e as principais características que envolvem a criação de uma avaliação adaptativa. Neste contexto, o principal componente inerente a um sistema de avaliação desse tipo é o banco de itens, considerado como elemento central do mesmo. Desse modo, podemos concluir que a vigência e a implantação de um teste adaptativo fica prejudicada sem a existência de um banco de itens robusto e calibrado de acordo com as necessidades e objetivos da avaliação pretendida.

Assim, a principal contribuição advinda deste trabalho de mestrado é o domínio e explanação de forma detalhada da teoria dos testes adaptativos, com ênfase especialmente no processo de criação e calibração de um banco de itens sensível ao conteúdo (referente ao domínio da proficiência em língua inglesa do ICMC-USP), bem como os aspectos envolvidos na estimação de parâmetros e nas características dos modelos de respostas pertencentes à Teoria de Respostas de Itens (TRI).

A experiência da criação de um banco de itens voltado para o Exame de Proficiência em Inglês deste Instituto é o principal resultado deste trabalho. Foi criado um banco de itens com 103 questões calibradas, especificamente com 32 questões pertencentes ao módulo **M1** (convenções da língua), 43 questões ao módulo **M2** (estruturação de textos), 12 questões do módulo **M3** (compreensão de textos) e 10 questões ao módulo **M4** (estratégia de escrita). O que se pôde aprender dessa experiência é que há a necessidade constante (por parte das pessoas envolvidas neste processo) de *cuidado, planejamento, disciplina* e *organização*. O desenvolvimento de um banco de itens exige *cuidado* no que tange às decisões iniciais do projeto de construção do banco, correspondente àquelas que definem a estrutura e conteúdo do mesmo. *Disciplina* e *organização* são indispensáveis para a fase de calibração dos itens (que também envolve a estimação de parâmetros), pois é necessário ser coerente com relação à seleção do modelo de resposta mais adequado aos itens. Por sua vez, é preciso um *planejamento* para o desenvolvimento do banco, já que seu crescimento e manutenção são processos que exigem tempo e dedicação.

A principal dificuldade encontrada durante o desenvolvimento deste trabalho foi, sem dúvida, a reunião de itens e sua posterior calibração e análise para inclusão no banco de itens. A tarefa de

conseguir "dados de estimação" a partir da colaboração de terceiros, essenciais para calibrar o banco de itens rapidamente foi bastante difícil, pois tivemos que agrupar questões duplicadas dos exames já realizados e para isso organizar conjuntos de amostras para saber o número exato de questões diferentes a serem incluídas no banco.

O desenvolvimento do **SisBI** trouxe à tona a dificuldade de criar bancos de itens sensíveis ao conteúdo, pois as restrições impostas na divisão do mesmo exigem um conhecimento empírico sobre o domínio de conhecimento do banco. Segundo Huang (1996) tal fato é o que mais dificulta a implantação dos testes adaptativos em pequenas instituições ou institutos como o ICMC. A modelagem do **TAEPI**, por sua vez, apresentou dificuldades no que diz respeito à decisão das particularidades de aplicação do exame adaptativo no cenário deste Instituto. O fato de não termos ainda uma experiência anterior/piloto da avaliação adaptativa nos obrigou a aceitar várias opções de definição do exame (p.e. critério de parada, seleção de itens, estimação da habilidade etc.), provocando a criação da modelagem de um sistema que prevê a existência de várias combinações desses. Esta flexibilidade tornou-se uma grande vantagem no final.

Entretanto, vale ressaltar que este trabalho apresenta alguns pontos que devem ser melhorados e como trabalhos futuros podemos citar:

1. A definição do método de estimação de habilidade mais adequado para o contexto do exame de proficiência em inglês baseado em um estudo piloto;

2. A investigação das questões técnicas relacionadas com os testes adaptativos, como a taxa de exposição e conflito de itens do banco;

3. A seleção, calibração e inclusão de novos itens no banco, aumentando seu volume, com intuito de torná-lo disponível em um futuro próximo para a implantação da avaliação diagnóstica adaptativa (conforme Tabela 6.9);

4. A definição e estudo do procedimento de pontuação mais adequado para o cenário deste exame, permitindo que a escala de habilidade medida pelos testes adaptativos seja traduzida para uma escala comumente usada;

5. A implementação do **TAEPI**;

6. A definição dos critérios de parada mais apropriados em relação à avaliação diagnóstica, considerando que a mesma é sensível ao conteúdo;

7. Uma comparação dos resultados da avaliação diagnóstica realizada pelos testes adaptativos com o atual método de avaliação do EPI do ICMC.

Tendo em vista as dificuldades encontradas ao longo deste trabalho e considerando o fato real de que a implantação dos testes adaptativos em pequenas instituições é custosa, podemos afirmar que os resultados alcançados e as contribuições deste trabalho descritas nesta dissertação são considerados

bastante satisfatórios, já que produzimos: um banco de itens calibrado (para um modelo da TRI de 2 parâmetros) e sensível ao conteúdo, um sistema de gerência e manutenção do mesmo, a modelagem de um sistema de avaliação adaptativa para o cenário deste Instituto e também a exposição teórica dos modelos e métodos que regem o comportamento dos testes adaptativos. Assim, ao final da pesquisa acreditamos ter alcançado o nosso objetivo: o domínio da teoria e tecnologia para a utilização de testes adaptativos para avaliar a proficiência em inglês para o programa de mestrado.

Bibliografia e Referências

(Anquetil, 2000) Anquetil, N. (2000). Desenvolvimento de software orientado a objetos. [http://www.cos.ufrj.br/~nicolas/UML/index.html] acessado em (08/01/2002).

(Aquino, 2001) Aquino, V. T. (2001). Avaliação automática de exames de profeciência em inglês. Dissertação de Mestrado, Universidade de São Paulo - USP São Carlos. Dissertação de Mestrado.

(Baker, 1992) Baker, F. (1992). *Item Response Theory Parameter Estimation Techniques.* MARCEL DEKKER, INC.

(Bennet, 1998) Bennet, E. R. (1998). Reinventing assessment. [www.ets.org/research/pic/bennet.html] acessado em (20/08/2000).

(Bilmes, 1998) Bilmes, J. A. (1998). A gentle tutorial of the em algotithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report TC-97-021, International Computer Science Institute.

(Bock and Liberman, 1970) Bock, R. D. and Liberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35(2):18.

(Brown, 1997) Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology*, 1(1):15.

(Brown et al., 1999) Brown, S., Race, P., and Bull, J. (1999). *Computer-Assisted assessment.* Staff and Educational Development Series. Kogan Page.

(Bruno, 1987) Bruno, J. E. (1987). Admissible probability measures in instructional management. *Journal of Computer Bases Instruction*, 14(1):7.

(Cabral and Araújo, 2002) Cabral, A. M. O. and Araújo, L. G. G. (2002). Uml - unified modeling language. [http://www.geocities.com/tutprog1/Uml1.zip] acessado em (08/01/2002).

(Cardoso and Domingues, 2000) Cardoso, E. and Domingues, L. (2000). Gestão de produtos informáticos: Diagrama de classes-uml. [http://www.iscte.pt/GSI/document/GSI_Aula5_6_DiagramaClasses.pdf] acessado em (18/01/2002).

(Cardoso and Souza, 2001) Cardoso, R. F. and Souza, R. D. (2001). Ferramenta para geração de avaliações baseadas em níveis de dificuldade. *Artigo apresentado no I Workshop de Informática do Sul de Minas*, page 8.

(Collins et al., 1996) Collins, J. A., Greer, J. E., and Huang, S. X. (1996). Adaptive assessment using granularity hierarchies and bayesian nets. In *ITS'96 3th International Conference in ITS*, Montreal.

(Craig, 1998) Craig, L. (1998). *Applying UML and Patterns - An Introduction to Object Oriented Analysis and Design.* Prentice-Hall, Inc. NJ.

(Cumming et al., 2000) Cumming, A., Kantor, R., Powers, D., Santos, T., and Taylor, C. (2000). Toefl 2000 writing framework: A working paper. [`ftp://etsis1.ets.org/pub/toefl/253719.pdf`] acessado em (22/05/2001).

(Dowling et al., 1996) Dowling, C. E., Hockemeyer, C., and Ludwig, A. H. (1996). Adaptive assessment and training using the neighbourhood of knowledge states. In *ITS'96 3th International Conference in ITS*, Montreal.

(Dunkel, 1999) Dunkel, P. A. (1999). Considerations in developing or using second foreign language proficiency computer-adaptive tests. *Language Learning & Technology*, 2(2):16.

(Gomez, 1999) Gomez, V. (1999). Avaliação formativa e continuada da educação baseada na internet. In *VI Congresso Internacional de Educação a Distância.*

(Green et al., 1984) Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., and Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 4:13.

(Hambleton and Swaminathan, 1985) Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Kluwer Nijhoff Publishing.

(Hanson, 1996) Hanson, B. (1996). Estimating parametric continuous latent distributions. Author's Note `http://www.b-a-h.com/papers/note9601.html` acessado em 22/02/2002.

(Huang, 1996) Huang, S. X. (1996). A content-balanced adaptive testing algorithm for computer-based training systems. In *ITS'96 3th International Conference in ITS*, Montreal.

(Kingsbury and Zara, 1989) Kingsbury, G. G. and Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied measurement in education*, 2(4):16.

(Kingsbury and Zara, 1991) Kingsbury, G. G. and Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied measurement in education*, 4(3):20.

(Kreitzberg et al., 1978) Kreitzberg, C. B., Stocking, M. L., and Swanson, L. (1978). Computerized adaptive testing: principles and directions. *Computer & Education*, 2:10.

(Lucena et al., 1998) Lucena, C. J. P., Fuks, H., Milidiú, R., Laufer, C., Blois, M., Choren, R., Torres, V., Ferraz, F., Carvalho, G. R., and Daflon, L. (1998). O aulanet e as novas tecnologias de informação aplicadas à educação baseada na web. [`http://anauel.cead.puc-rio.br/aulanet`] acessado em (22/05/2001).

(Meyer, 1973) Meyer, P. L. (1973). *Probabilidade: aplicações à estatística.* Ao Livro Técnico S.A.

(Miller et al., 1998) Miller, A. H., Imrie, W. B., and Cox, K. (1998). *Student Assessment in higher education.* Kogan Page.

(Morgan and O'Reilly, 1999) Morgan, C. and O'Reilly, M. (1999). *Assessing Open and Distance Learners.* Open and Distance Series. Kogan Page.

(Olea et al., 1999) Olea, J., Ponsoda, V., and Prieto, G. (1999). *Tests Informatizados Fundamentos y Aplicaciones.* Ediciones Pirámede.

(Owen, 1975) Owen, R. J. (1975). A baysesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70(350).

(Ponsoda et al., 1994) Ponsoda, V., Olea, J., and Revuelta, J. (1994). Adtest: A computer-adaptive test based on the maximum information principle. *Educational and Psychological Measurement*, 54(3):6.

(Revuelta and Ponsoda, 1997) Revuelta, J. and Ponsoda, V. (1997). Una solução a la estimación inicial en los tests adaptativos informatizados. *Revista Electronica de Metodologia Aplicada*, 2(2):6. [`www.uniovi.es/~psi/REMA/v2n2/a1/`] acessado em (27/08/2000).

(Righetto, 2001) Righetto, V. L. (2001). Provanet: Um sistema de avaliação de aprendizado na internet. Dissertação de Mestrado, Universidade Estadual de Campinas - Unicamp Campinas-SP. Dissertação de Mestrado.

(Rios et al., 1998) Rios, A., de la Cruz, J. L. P., and Conejo, R. (1998). Siette: Intelligent evaluation system using tests for teleeducation. In *ITS'98 4th International Conference in ITS*, Texs.

(Romani et al., 2000) Romani, L., Rocha, H., and Silva, C. (2000). Ambientes para educaçãp a distância baseada na web: Onde estão as pessoas. In *Anais do IHC'2000 - Workshop sobre fatores humanos em sistemas e computação*, Gramado-RS.

(Rudner, 1998) Rudner, M. L. (1998). An on-line, interactive, computer adaptative testing mini-tutorial. [www.ericae.net/scripts/cat/catdemo.html] acessado em (14/02/2001).

(Sahu, 1998) Sahu, S. K. (1998). Bayesian estimation and model choice in item response models. Technical report, University if Wales.

(Sands et al., 1997) Sands, W. A., Waters, B. K., and McBride, J. R. (1997). *Computerized Adaptive Testing: From Inquiry to Operation.* American Psychological Association.

(Scott and Fowler, 1998) Scott, K. and Fowler, M. (1998). *UML Distilled: Appling the Standard object Modeling Language.* Addison-Wesley.

(Serrão, 2000) Serrão, C. (2000). Fundamentos de bases de dados. [`http://gab01-3.iscte.pt/19992000/fbd/Slides_02.pdf`] acessado em (18/01/2002).

(Swales, 1990) Swales, J. (1990). *Genre Analysis - English in academic and research settings.* Cambridge University Press.

(Veldkamp and der Linden, 2000) Veldkamp, B. P. and der Linden, W. J. V. (2000). Designing item pools for computerized adaptive testing. *Chapter in Computerized Adaptive Testing: Theory and Practice*, (1):12.

(Wainer and Kiely, 1987) Wainer, H. and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal od Educational Measurement*, 24(3):16.

(Wang et al., 2001) Wang, X., Bradlow, E. T., and Wainer, H. (2001). User's guide for scoright (version 1.2): A computer program for scoring tests built of testlets. Technical Report RR-01-06, Educational Testing Service - ETS.

(Weiss, 1985) Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and clinical Psychology*, 53(6):15.

(Weissberg and Buker, 1990) Weissberg, R. and Buker, S. (1990). *Writing Up Research - Experimental Research Report Writing for Students of English.* Prentice-Hall, Inc.

(Wise and Kingsbury, 2000) Wise, S. L. and Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica*, 21:20.

(Woodruff and Hanson, 1997) Woodruff, D. and Hanson, B. (1997). Estimation for item responde models using the em algorithm for finite mixtures. In *Annual Meeting of the Psychometric Society*, Tennessee.

(XCalibreManual, 1997) XCalibreManual (1997). User's manual for the xcalibre marginal maximum-likelihood estimation. Assessement Systems Corporation [`http://www.assess.com`] acessado em (22/02/2002).

(Yourdon, 1990) Yourdon, E. (1990). *Análise Estruturada Moderna.* Editora Campus.

(Zinn, 2000) Zinn, Y. L. (2000). Considerações sobre exames de proficiência e admissão para universidades dos estados unidos da américa. *Estudos & Dados*, 8(1):6.

APÊNDICE A

Exemplo de uma Prova do Exame de Proficiência em Inglês – EPI

Este Apêndice traz o exemplo de uma prova modelo do Exame de Proficiência em Inglês do ICMC. Vale ressaltar que a prova aqui apresentada (Código 29) não é uma prova formal, igual a realizada no instituto, e sim uma prova modelo, que serve principalmente pra o treinamento e familiarização dos alunos com o tipo da mesma. Todos os módulos e partes previstas no EPI estão contidas nesta prova, perfazendo um total de 26 questões. A prova formal possui apenas 20 questões.

CÓDIGO 29

Teste de Computação

Part 1 – M2I

Divide and classify the stages of an introduction.

The description of the structure of introductions, together with an introduction to a paper in the field of Computer Science, is reproduced here. First read them, then answer the questions.The structure of introductions. The introduction of a research paper can be divided into 8 stages.

In STAGE I (setting), the writer establishes a context, or frame of reference to help readers understand how the research fits into a wider field of research.

In STAGE II (review of the literature), the writer reviews the findings of other researchers who have already published in his/her area of interest.

In STAGE III (gap), the writer: indicates that previous literature described in stage II is inadequate because an important aspect of the research area has been ignored by other authors; OR indicates that there is an unresolved conflict among authors of previous studies concerning the research topic; OR indicates that an examination of the previous literature suggests an extension of the topic; OR raises a new research question not previously considered by other workers in the field. In STAGE IV (purpose), the writer formally announces the purpose(s) of the research.

In STAGE V (methodology), the writer describes either the steps followed in conducting the research or some method considered in the study.

In STAGE VI (main results), the writer presents the main findings of the research.

In STAGE VII (value of the research), the writer indicates possible benefits or applications of the research.

Finally, in STAGE VIII (layout of the article), the writer presents the structure or layout of the paper, indicating the sections and briefly commenting on them.

Some of the stages are OPTIONAL and may depend on the journal/conference style and length of the article. However, the general plan given in the reading frame is very common.

Texto

<http://ieeexplore.ieee.org/lpdocs/epic03/> Computer Graphics International 2001, Hong Kong, July 3-6 2001 A HYBRID APPROACH TO THE RECOVERY OF DEFORMABLE SUPERQUADRIC MODELS FROM 3D DATA James Sinnott and Toby Howard

Introduction

1) The problem of recovering the shape of objects from unstructured 3D data is important in many areas of computer graphics and computer vision, including robotics, medical imaging and the automatic construction of virtual environments. 2) In the last 25 years, much work has focussed on finding suitable models for the recovery of objects from 3D data. 3) This work has largely proposed the use of some form of parametric model, most commonly the superquadrics.

4) Superquadrics are simple parametric models that can represent a large range of shapes with a small number of parameters, and have mathematical properties that make them particularity suited to efficient model recovery. 5) The shape of an object is recovered by fitting a superquadric model as closely as possible to the 3D data. 6) Pentland [1] was first to propose the use of superquadrics as a model for object recovery, and in recent years, many researchers [2], [3], [4], [5], [6], [7] have reported success in the recovery of superquadrics models, often combined with global or local deformations. 7) The most commonly used method of fitting superquadric models to 3D data is through nonlinear least-squares minimisation of an error-of-fit function. 8) However, least-squares minimisation approaches often perform poorly when searching complex parameter spaces and can only guarantee convergence to a local minimum. 9) In this paper we present an alternative hybrid approach to the recovery of deformable superquadric models from 3D data. 10) We propose a two-stage process for fitting a deformable superquadric to a set of points, based on a genetic algorithm and a nonlinear least-squares minimisation routine.

11) The rest of the paper is organised as follows. 12) Section 2 provides a formal definition of our object model, superquadrics with global parametric deformations. 13)Section 3 defines the problem of fitting deformable superquadric models to 3D points and describes the standard

approach to solving this problem through the use of nonlinear least-squares minimisation. 14) Section 4 investigates the possibility of using genetic algorithms as an alternative. 15) Following this discussion, Section 5 proposes a novel hybrid approach to fitting deformable superquadrics to 3D points, and Section 6 presents the results of testing this hybrid approach. 16) Finally, Section 7 concludes the paper and discusses areas for future work.

Questões

01) Which sentence(s) in the introduction presented here correspond(s) to the STAGE I (setting)?

1 *

1 and 2

1 to 3

02) Which sentence(s) in the introduction presented here correspond(s) to the STAGE II (review of the literature)?

2 to 7 *

2

6 and 7

03) Which sentence(s) in the introduction presented here correspond(s) to the STAGE III (gap)?

7 and 8

8 *

7

04) Which sentence(s) in the introduction presented here correspond(s) to the STAGE IV (purpose)?

9

9 and 10 *

10

Part 2 – M2A

Divide and classify the stages of an abstract

Instruções

The description of the structure of abstracts, together with an abstract of a paper in the field of Computer Science, is reproduced here. First read them, then answer the questions. The structure of abstracts

The abstract of a research paper can be divided into 5 stages.

In STAGE I (background), the writer presents some background information to help readers understand the rest of the abstract.

In STAGE II (purpose), the writer establishes the principal activity (or purpose) of the study and its scope.

In STAGE III (methodology), the writer provides some information about the methodology used in the study.

In STAGE IV (results), the most important results of the study are presented by the writer.

In STAGE V (conclusion), the writer formulates a statement of conclusion or recommendation.

However, abstracts are usually written to be as brief and concise as possible. In order to shorten an abstract, it is possible to eliminate or combine much of the information shown in the format previously shown.

The typical order of the stages in reduced abstracts is shown below:

The STAGE II (purpose) and STAGE III (methodology) are presented as one and are the first to appear.

The STAGE IV (results) is presented (typically with more emphasis than the others).

The last and optional stage is the STAGE V (conclusion).

Texto

<http://www.dcs.ex.ac.uk/ jamie/> IEEE Transactions on knowledge and data engineering, Vol. 13 No. 2 March/April 2001 INCREMENTAL SYNTACTIC PARSING OF NATURAL LANGUAGE CORPORA WITH SIMPLE SYNCHRONY NETWORKS Peter C.R. Lane and James B. Henderson

Abstract

1)This article explores the use of Simple Synchrony Networks (SSNs) for learning to parse English sentences drawn from a corpus of naturally occurring text. 2) Parsing natural language sentences requires taking a sequence of words and outputting a hierarchical structure representing how those words fit together to form constituents.3) Feed-forward and Simple Recurrent Networks have had great difficulty with this task, in part because the number of relationships required to specify a structure is too large for the number of unit outputs they have available. 4) SSNs have the representational power to output the necessary $O(n^2)$ possible structural relationships because SSNs extend the $O(n)$ incremental outputs of Simple Recurrent Networks with the $O(n)$ entity outputs provided by Temporal Synchrony Variable Binding. 5) This article presents an incremental representation of constituent structures which allows SSNs to make effective use of both these dimensions.6) Experiments on learning to parse naturally occurring text show that this output format supports both effective representation and effective generalization in SSNs.7) To emphasize the importance of this generalization ability, this article also proposes a short-term memory mechanism for retaining a bounded number of constituents during parsing. 8) This

mechanism improves the $O(n^2)$ speed of the basic SSN architecture to linear time, but experiments confirm that the generalization ability of SSN networks is maintained.

Questões

05) Which stage(s) is(are) included in sentence 1?
STAGE I (background)
STAGE II (purpose) + STAGE III (methodology) *
STAGE III (methodology)

06) Which stage(s) is(are) represented by the most number of sentences?
STAGE I (background) and STAGE II (purpose)*
STAGE III (methodology)
STAGE V (conclusion)

07) Which stage(s) is(are) included in sentence 8?
STAGE II (purpose)
STAGE IV (results)*
STAGE V (conclusion)

08) Which stage(s) has(have) been eliminated?
STAGE I (background)
STAGE II (purpose)
STAGE V (conclusion) *

Part 3 – M4S

Identify strategies to write the STAGES of an introduction.

Instruções

Different strategies to write the stages of an introduction, together with examples of the STAGE I of introductions to papers in the field of Computer Science are reproduced here. First read them, then answer the questions.

Strategies to Stage I

There are different approaches to write the stages of an introduction. For example, the STAGE I can be written by using three different strategies:

ARGUING ABOUT THE TOPIC PROMINENCE
FAMILIARIZING TERMS OR OBJECTS OR PROCESSES
INTRODUCING THE RESEARCH TOPIC FROM THE RESEARCH AREA

The first one uses arguments; the second follows one of the three patterns: description, definition or classification; and the third follows the general to particular ordering of details.

Texto /Questões

Example 1

<http://epubs.siam.org/sam-bin/dbq/article/35184> SIAM J. CONTROL OPTIM. Vol. 40, No. 1, pp 88-106 ON THE BOUNDEDNESS AND CONTINUITY OF THE SPECTRAL FACTORIZATION MAPPING BIRGIT JACOB AND JONATHAN R. PARTINGTON

Spectral factorization is the process by which a (possibly matrix-valued) function G is written as $G = W \cdot W$; that is, $G(e^I \cdot \theta) = W(e^I \cdot \theta) \cdot W(e^I \cdot \theta)$ for $\theta E[0, 2\pi]$, or $G(iw) = W(iw) \cdot W(iw)$ for wER.

09) Which strategy does follow example 1?
Arguing about the topic prominence
Familiarizing terms or objects or processes *
Introducing the research topic from the research area

Example 2

<http://www.amstat.org/publications/jse/v9n2/hirsch.html> Journal of Statistics Education Volume 9, Number 2 (2001) REPRESENTATIVENESS IN STATISTICAL REASONING: IDENTIFYING AND ASSESSING MISCONCEPTIONS Linda S. Hirsch and Angela M. O'Donnell

"Probability is the study of likelihood and uncertainty. It plays a critical role in all of the professions and in most everyday decisions (Halpern 1996, p. 242). Being able to reason effectively about probability is necessary for many practical concerns such as interpreting weather reports, understanding DNA evidence at trials, the risks of childbirth defects, and car insurance rates among others (Abelson 1995; Derry, Levin, Osana, and Jones 1998).

10) Which strategy does follow example 2?
Arguing about the topic prominence *
Familiarizing terms or objects or processes
Introducing the research topic from the research area

Example 3

<http://www.amstat.org/publications/jse/v5n3/giraud.html> Journal of Statistics Education v.5, n.3 (1997) COOPERATIVE LEARNING AND STATISTICS INSTRUCTION Gerald Giraud

Cooperative learning has been advocated as an instructional methodology because of its effect on achievement and on other attributes that accompany the acquisition of knowledge, including motivation, classroom socialization, the student's confidence in learning, and attitude toward the subject being learned (e.g., Johnson and Johnson 1985, 1986a, 1986b). Cooperative learning has been defined as the instructional use of small groups such that students work together to maximize their own and each other's learning (e.g., Johnson, Johnson, and Smith 1991). In this study, cooperative learning involved students working together in small groups to solve problems and complete assignments in a post-secondary introductory statistics course.

11) Which strategy does follow example 1?
Arguing about the topic prominence
Familiarizing terms or objects or processes
Introducing the research topic from the research area *

Part 4 – M4R

Identify strategies to write the STAGES of an introduction.

Instruções

Different strategies to write the stages of an introduction, together with examples of the STAGE II (review of the literature) of introductions are reproduced here. First read them, then answer the questions.

Strategies to Stage II

There are different approaches to write the stages of an introduction. For example, the STAGE II can be written by using three different strategies:

CITATIONS GROUPED BY APPROACH
CITATIONS ORDERED FROM GENERAL TO SPECIFIC
CITATIONS ORDERED CHRONOLOGICALLY

The first one is better suited for reviews of the literature which encompasses different approaches; in the second, citations are organized in order from those most distantly related to the study to those most closely related; and the third is used, for example, when describing the history of research in an area.

Texto / Questões

Example 1

www.elsevier.com/locate/tcs Theoretical Computer Science 253 (2001) 287–309 MEASURE AND PROBABILITY FOR CONCURRENCY THEORISTS Prakash Panangaden

In the area of semantics, one of the earliest serious investigations is due to Kozen [28, 29]. His work uses measuretheoretic ideas in a serious way and, in particular, uses conditional probability distributions or Markov kernels as a central tool. In his work he gives probabilistic semantics of a language of while loops and also develops a "Stonetype" duality using the idea that measurable functions are probabilistic predicates. This analogue of "predicate transformer" semantics has been extensively developed by a group at Oxford [35]. In a different vein Gupta, Jagadeesan and Saraswat [18] have developed a modeling language for probabilistic systems based on the concurrent constraint programming paradigm. As soon as one adds recursion to the language [17] one is forced into the realm of continuous spaces and the ideas expounded in the present paper are relevant. The other main starting point for the use of probability theory in semantics was the work of SahebDjahromi [37, 38]. His work combined probability theory and domain theory and was the inspiration for probabilistic powerdomains [23, 22] by Jones and Plotkin. Ultimately this led to the enormously fruitful work of Edalat and others [12–14] on integration on domains. The work on verification of probabilistic systems has exploded in recent years - it is impossible to attempt a survey here. There are approaches based on automata theory, process algebra equivalences, logics and model checking. A very interesting development has been the use of a probabilistic process algebra as a compositional performance evaluation tool [20].

12) Which strategy does follow example 1?
Citations grouped by approach
Citations ordered from general to specific
Citations ordered chronologically *

Example 2

<http://citeseer.nj.nec.com/rowley96neural.html> Rowley, Baluja, and Kanade: Neural Network-Based Face Detection (PAMI, January 1998) NEURAL NETWORK-BASED FACE DETECTION Henry A. Rowley, Shumeet Baluja, and Takeo Kanade

Many face detection researchers have used the idea that facial images can be characterized directly in terms of pixel intensities. These images can be characterized by probabilistic models of the set of face images [4, 13, 15], or implicitly by neural networks or other mechanisms [3, 12, 14,

19,21,23,25,26]. The parameters for these models are adjusted either automatically from example images (as in our work) or by hand. A few authors have taken the approach of extracting features and applying either manually or automatically generated rules for evaluating these features [7, 11].

13) Which strategy does follow example 2?
Citations grouped by approach *
Citations ordered from general to specific
Citations ordered chronologically

Example 3

URL: <http://citeseer.nj.nec.com/bauer99empirical.html> Machine Learning, 36, 105-142 (1999) AN EMPIRICAL COMPARISON OF VOTING CLASSIFICATION ALGORITHMS: BAGGING BOOSTING, AND VARIANTS Eric Bauer , Ron Kohavi

Voting algorithms can be divided into two types: those that adaptively change the distribution of the training set based on the performance of previous classifiers (as in boosting methods) and those that do not (as in Bagging). Algorithms that do not adaptively change the distribution include option decision tree algorithms that construct decision trees with multiple options at some nodes (Buntine 1992b, Buntine 1992a, Kohavi & Kunz 1997); averaging path sets, fanned sets, and extended fanned sets as alternatives to pruning (Oliver & Hand 1995); voting trees using different splitting criteria and human intervention (Kwok & Carter 1990); and error-correcting output codes (Dietterich & Bakiri 1991, Kong & Dietterich 1995). Wolpert (1992) discusses "stacking classifiers into a more complex classifier instead of using the simple uniform weighting scheme of Bagging. Ali (1996) provides a recent review of related algorithms, and additional recent work can be found in Chan,Stolfo & Wolpert (1996). Algorithms that adaptively change the distribution include AdaBoost (Freund & Schapire 1995) and Arc-x4 (Breiman 1996a). Drucker & Cortes (1996) and Quinlan (1996) applied boosting to decision tree induction, observing both that error significantly decreases and that the generalization error does not degrade as more classifiers are combined. Elkan (1997) applied boosting to a simple NaiveBayesian inducer that performs uniform discretization and achieved excellent results on two real-world datasets and one artificial dataset, but failed to achieve significant improvements on two other artificial datasets.

14) Which strategy does follow example 3?
Citations grouped by approach *
Citations ordered from general to specific
Citations ordered chronologically

Part 5 – M3

www.computer.org/tkde/tk2001/k0079abs.htm IEEE Transaction on Knowledge anda data engineering, Vo. 13 No. 1 January/February MULTIPLE SIMILARITY QUERIES: A BASIC DBMS OPERATION FOR MINING IN METRIC DATABASES Bernhard Braunmüller, Martin Ester, Hans-Peter Kriegel, and Jörg Sander

1 INTRODUCTION

1) METRIC databases are databases where a metric distance function is defined for pairs of database objects. 2) A prominent special case is a database of objects from a vector space, that is, objects with numeric attributes. 3) For example, multimedia objects [10] are typically represented by a large number of numeric features, such as shape descriptors or color histograms. 4) In many scientific applications, 4a)e.g., in astronomy [15], 4)automatic facilities measure a large number of numeric values for each database object, such as the amplitude emitted in some frequency band. 5) On the other hand, in a database monitoring WWW accesses, the objects may model URLs and these objects are not from a vector space but a metric distance function can be supplied.

6) Similarity between database objects is expressed by the distance function such that a low distance corresponds to a high degree of similarity, whereas two objects with a large distance are considered to be rather dissimilar. 7) Similarity queries [29], 7a)e.g., range queries or k-nearest-neighbor queries, 7) are the most important queries in metric databases. 8) Such queries play a major role in applications, 8a)such as multimedia systems, decision support systems, and data mining.

9) A lot of research on analyzing large databases - manually and automatically - has been conducted. 10) Data exploration is the process of manually exploring a database [20]. 11) A user starts at a given database object and from there he or she interactively navigates through the database, 11a) for example, by iteratively retrieving all similar objects. 12) That is, the answers of previous queries may be used as query objects for new similarity queries. 13) Knowledge discovery in databases (KDD) has been defined as the nontrivial process of discovering valid, novel, potentially useful, and ultimately understandable patterns from data [11]. 14) The core step of the KDD process is the step of data mining, 14a) i.e., the application of appropriate algorithms that automatically produce a particular enumeration of patterns over the data. 15) For example, a density-based clustering algorithm such as DBSCAN [9] starts from some object and repeatedly retrieves the neighborhood of objects which have been retrieved by previous queries as long as the density in this neighborhood is large enough.

16) In traditional query processing, single queries are issued independently by different users. 17) In manual data exploration as well as in automatic data mining, however, many similarity queries must be answered in a single application. 18) We define multiple queries as sets of queries issued simultaneously. 19) Clearly, multiple queries provide much more potential for query optimization than single queries. 20) In the context of relational databases and SQL queries, common subqueries of several queries may be detected and their results reused to speed-up processing multiple queries. 21) A lot of papers have investigated this approach; recent work includes [28] and [25]. 22) Typically,

directed acyclic graphs are used to represent alternative query plans and a cost model is applied to estimate the cost for executing a given plan. 23) Subramanian and Venkataraman [28] consider sharing only among the best plans of each query and, thus, do not always obtain the cheapest overall solution. 24) Roy et al. [25] consider all query plans and propose a novel heuristic to efficiently optimize multiple SQL queries in this general case.

25) In the context of multimedia or metric databases, however, multiple queries have not yet received much attention. 26) Chakrabarti et al. [6] discuss efficient query processing for query refinement in multimedia databases. 27) A so-called multipoint query consists of a set of query objects and this query may be refined by adding or replacing query objects after having evaluated the answers of the previous query. 28) The basic idea of the proposed technique is to cache relevant information when processing k-nearest-neighbor queries on top of an index structure. 29) Thus, the I/O cost for refined queries is significantly reduced. 30) Introduced in [5] is the notion of a multiple similarity query defined by a set of query objects and some basic techniques for efficiently processing such queries are presented. 31) While query refinement of multipoint queries can be modeled by multiple similarity queries, the latter approach is much more general and effective for many other query types and applications, as described above.

32) In this paper, we investigate two orthogonal approaches to speed-up the processing of multiple similarity queries in metric databases: reduce I/O cost (32a) that is, the number of disk accesses) and reduce CPU cost (32b) that is, the number of distance calculations). 33) Furthermore, we explore the potential of parallelization. 34) The proposed techniques can be combined in order to obtain a maximum performance when processing multiple similarity queries.

Questões

15) Qual a relação entre as idéias/informações expostas nas orações 4 e 5 da introdução?
a) 5 é uma explicação da 4
b) contraste/oposição *
c) 5 é uma conclusão da 4

16) Qual a relação entre as idéias/informações expostas nas orações 11 e 12 da introdução?
a) 12 é uma reformulação da 11 *
b) contraste/oposição
c) 12 é uma conclusão da 11

17) Qual a relação entre as idéias/informações expostas nas orações 16 e 17 da introdução?
a) 17 é uma explicação da 16
b) similaridade
c) contraste/oposição *

18) Qual a relação entre as idéias/informações expostas nas orações 32 e 33 da introdução?
a) adição *
b) 33 é uma conclusão da 32
c) contraste/oposição

Part 6 M1G

Identify appropriated signal words used in the Stage III of introductions

Instruções

Different ways to write the Stage III of an introduction, together with six examples Stages III to papers in the field of Computer Science, are reproduced here. First read them, then answer the questions.

Texto / Questões

1) Elkan (1997) applied boosting to a simple NaiveBayesian inducer that performs uniform discretization and achieved excellent results on two real-world datasets and one artificial dataset, ———— failed to achieve significant improvements on two other artificial datasets.

19) Which is the word that fills in the blank of example 1?
but *
while
though

2) ——— knowledge-based approaches typically require an extensive manual knowledge engineering effort to create the knowledge base.

20) Which is the word that fills in the blank of example 2?
While
however *
though

3) ——— this may seem a glimpse of the obvious, programming languages seem to have been mainly devised for ease of generation rather than for comprehension (in so far as cognitive aspects have been considered at all).

21) Which is the word that fills in the blank of example 3?
although *
but
however

4) ——— these approaches only approximate phrasal recognition and provide a weak sense of context.

22) Which is the word that fills in the blank of example 4?
but *
although
while

5)——— the results have been disappointing, being limited to artificial domains and oversimplified subproblems (e.g. (Elman, 1991)).

23) Which is the word that fills in the blank of example 5?
although
while
however *

6 ——— there is very little work that considers both costs together.

24) Which is the word that fills in the blank of example 6?
while
however *
although

Part 7 – M1P

Identify appropriate verb tenses used in the Stage IV of introductions

Instruções

Different ways to write Stage IV of an introduction, together with six examples of Stages IV to papers in the field of Computer Science are reproduced here. First read them choose the best tense for each verb given in parentheses, then answer the questions.

Texto / Questões

1) In response to this, we ——— (design) a study to analyze people information capture activities.

2) The aim of this paper is to (demonstrate) ——— the applicability of information theory tools in an average-case learning model.

3) In this paper, we (investigate) ——— two orthogonal approaches to speed-up the processing of multiple similarity queries in metric databases: reduce I/O cost (that is, the number of disk accesses) and reduce CPU cost (that is, the number of distance calculations).

25) Which is the sequence that fills in the blanks of examples 1 to 3?
will design, demonstrate, will investigate
designed, demonstrate, investigate *
designed, demonstrates, investigate

4) This report (emphasize) ——— the new mechanisms in SAMUEL, and illustrate their utility in learning some interesting behaviors in multi-agent environments.

5) This paper (address) ——— one of reinforcement learning's biggest stumbling blocks: the curse of dimensionality [5], in which costs increase exponentially with the number of state variables.

6) This paper (study) ——— the pitfalls of discretization during reinforcement learning.

26) Which is the sequence that fills in the blanks of examples 4 to 6?
will emphasize, addressed, studied
emphasizes, addresses, study
will emphasize, addresses, studies *

APÊNDICE B

Modelagem e Interfaces do Sistema de Gerenciamento do Banco de Itens – SisBI

O propósito deste apêndice é apresentar todas as interfaces de uso do Sistema de Gerenciamento do Banco de Itens (SisBI), bem como os diagramas desenvolvidos na linguagem UML que representam a sua modelagem. Além da descrição das classes, também são mostrados os diagramas de caso de uso, de classes e de atividades, que descrevem o funcionamento geral deste sistema.

B.1 Interfaces de Uso

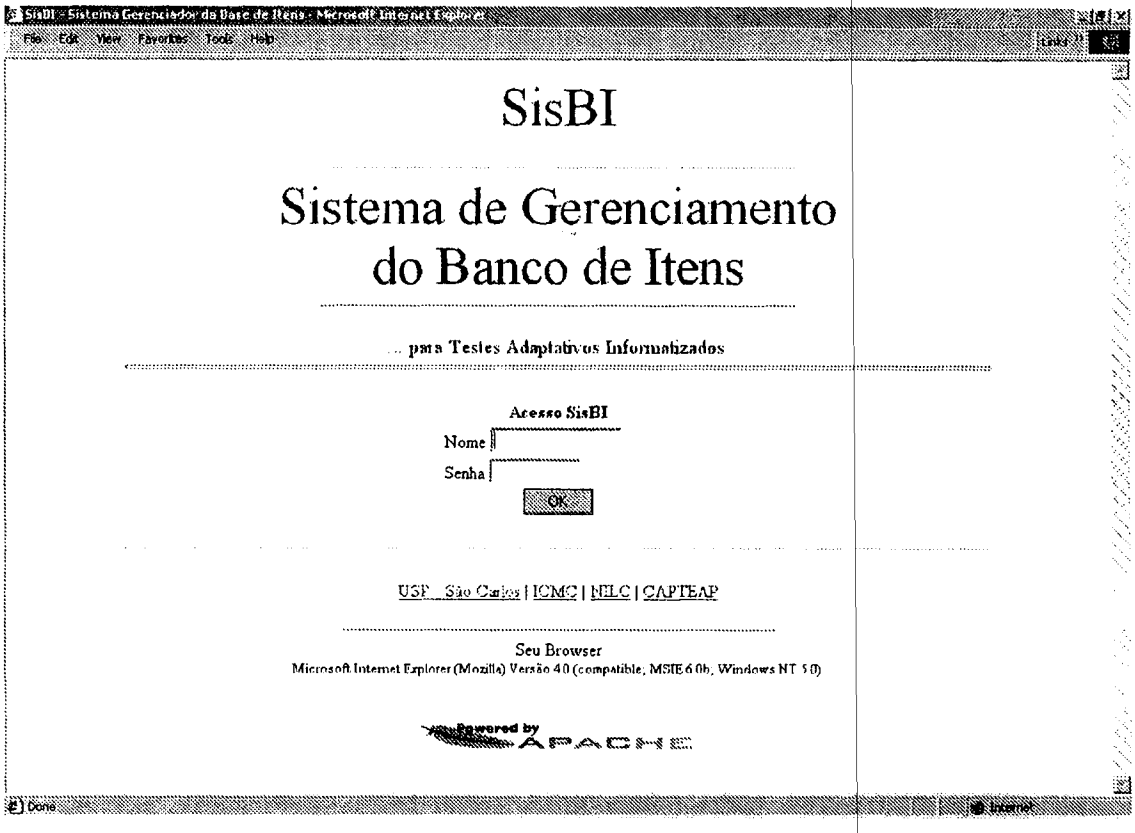

Figura B.1: Tela de entrada (acesso) do SisBI

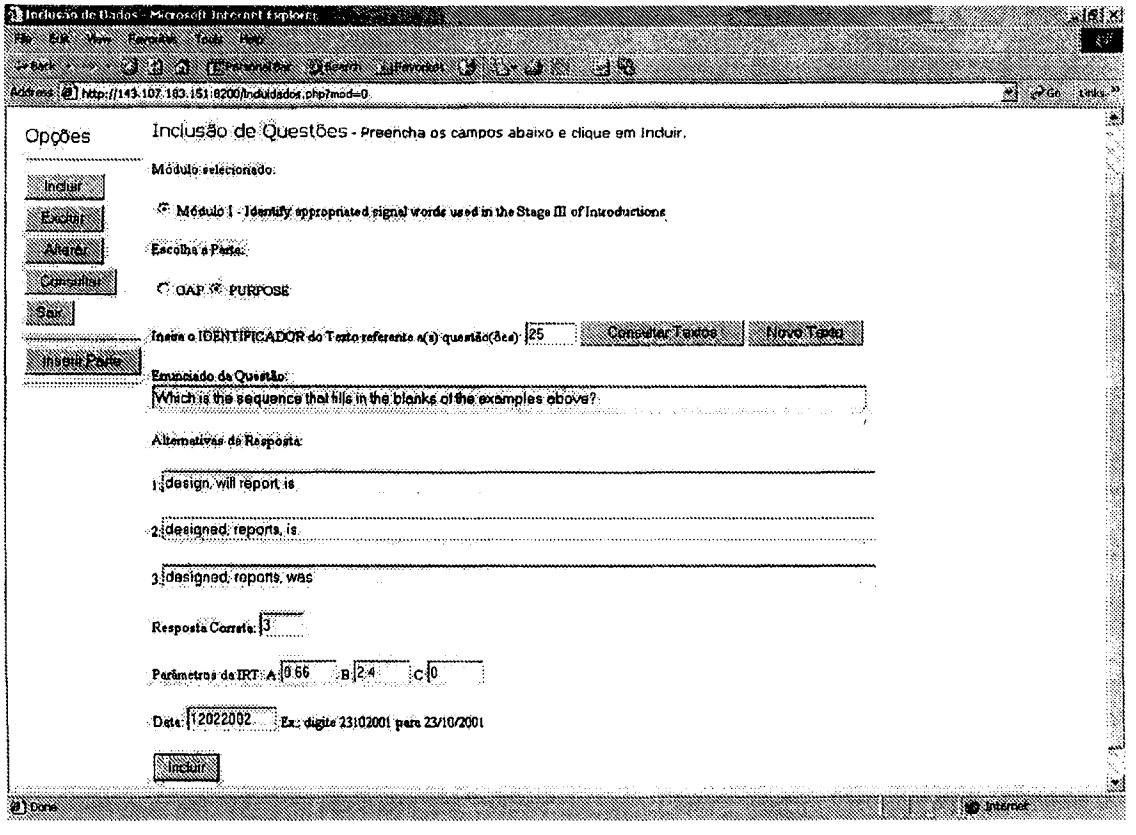

Figura B.2: Tela que representa o formulário de inclusão de uma nova questão.

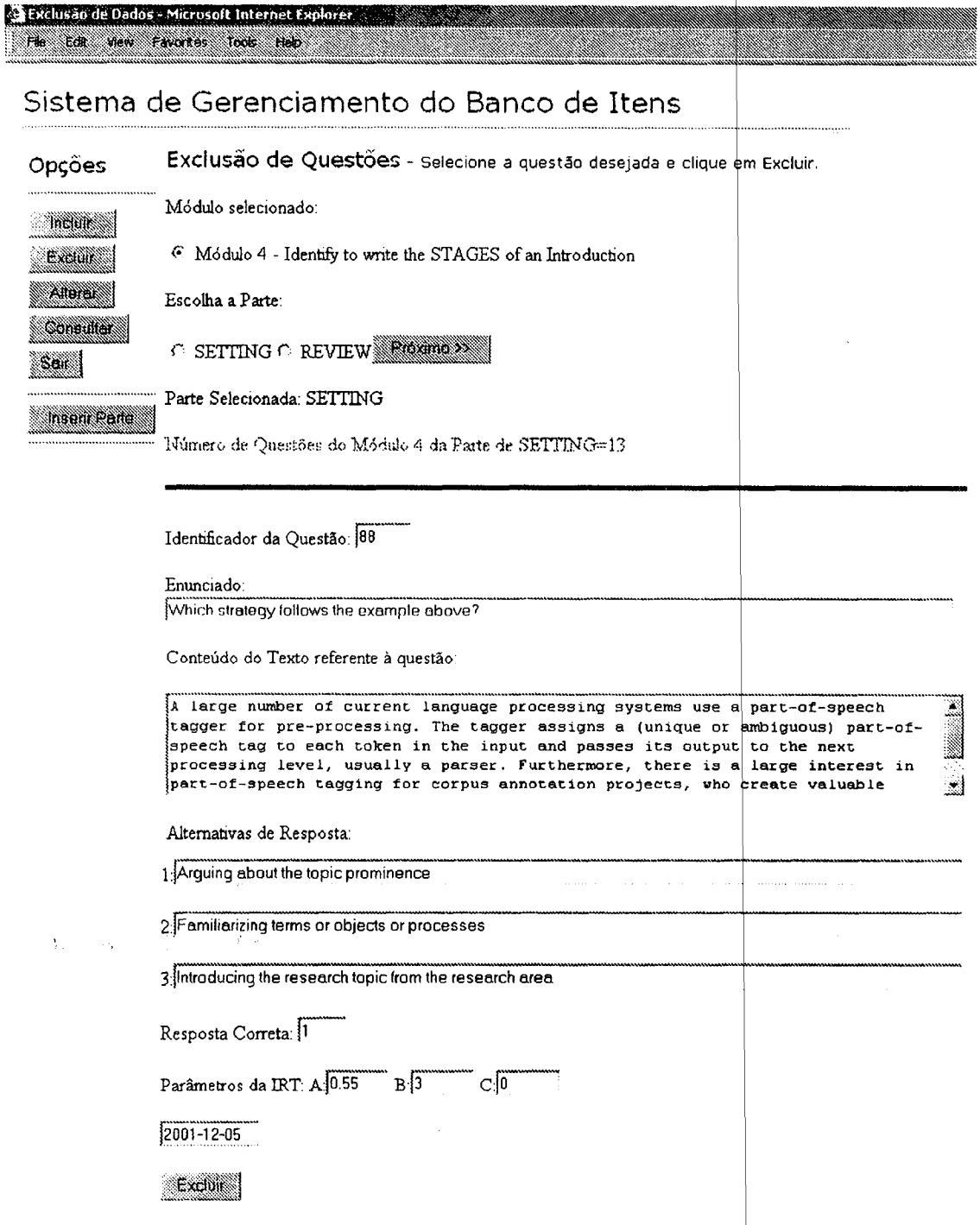

Figura B.3: Janela que representa a exclusão de questões.

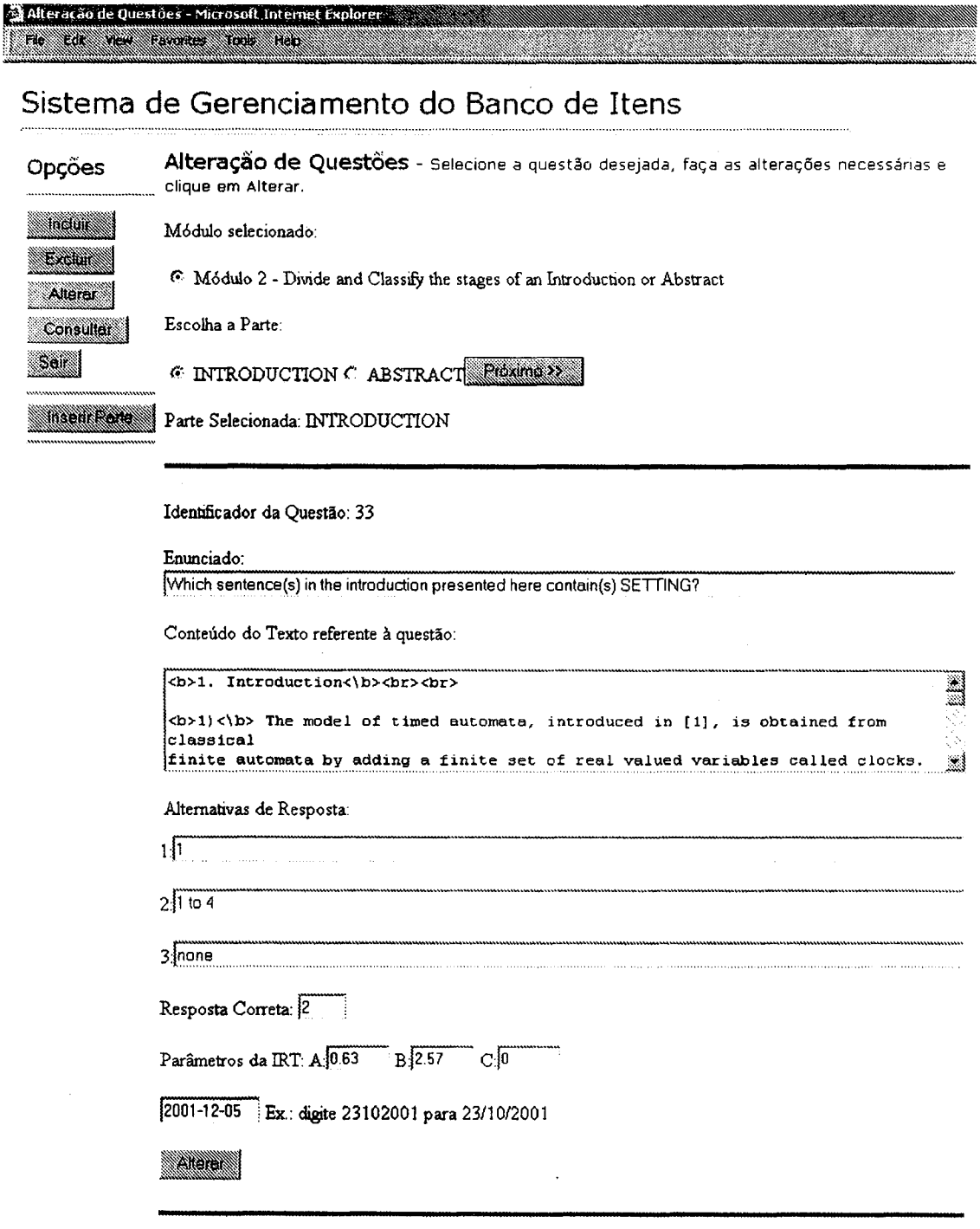

Figura B.4: Janela que representa a alteração de questões.

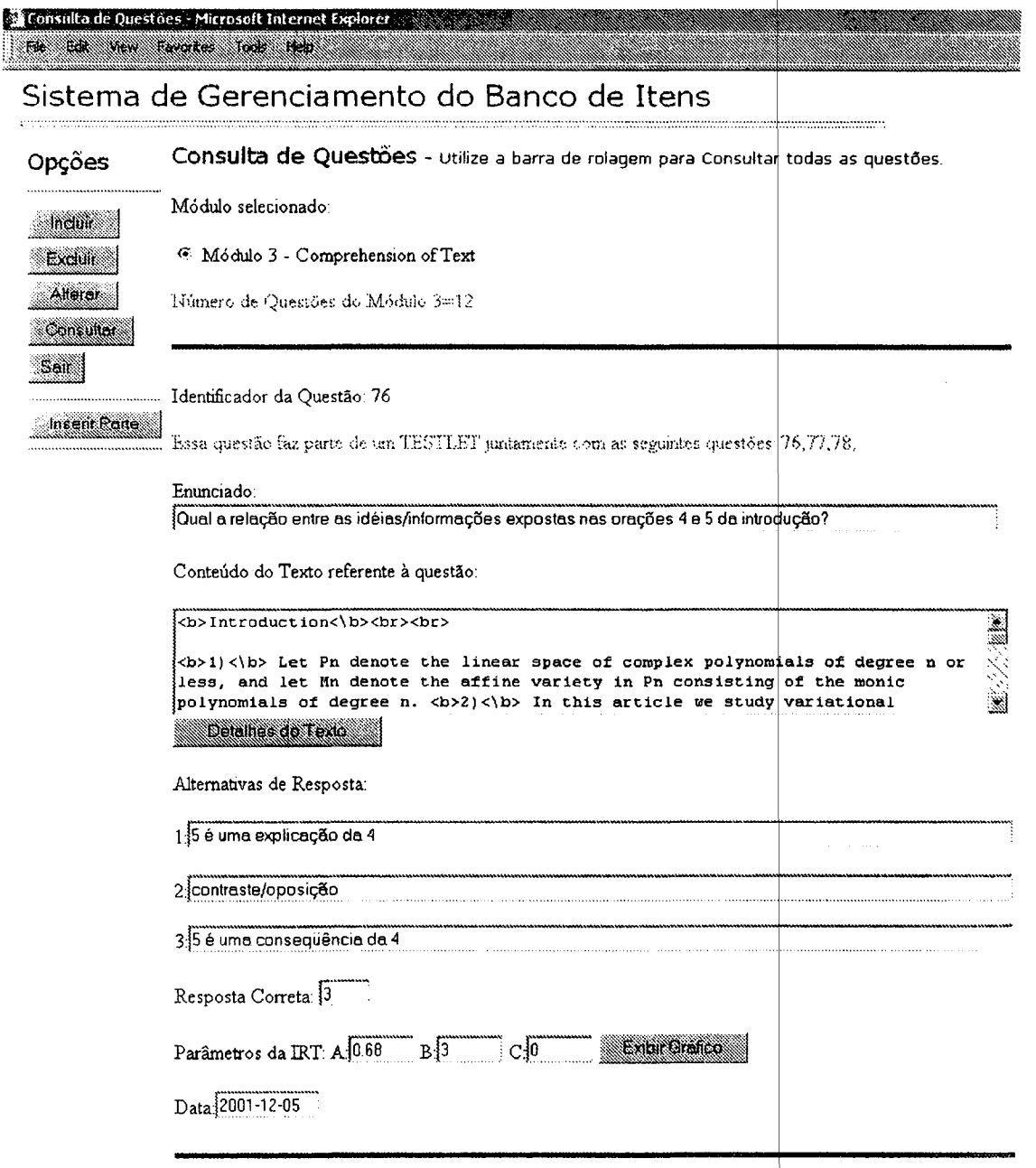

Figura B.5: Tela de consulta de questões.

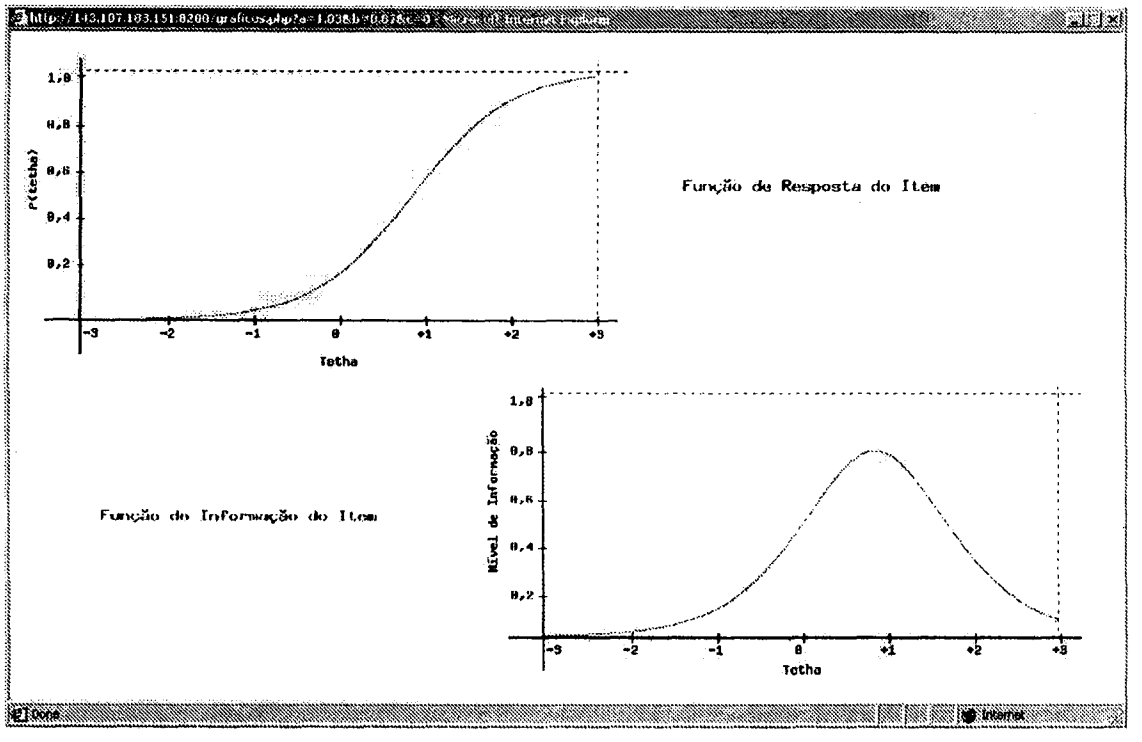

Figura B.6: Janela de gráficos da TRI de uma determinada questão com valores de a=1,03 e b=0,87.

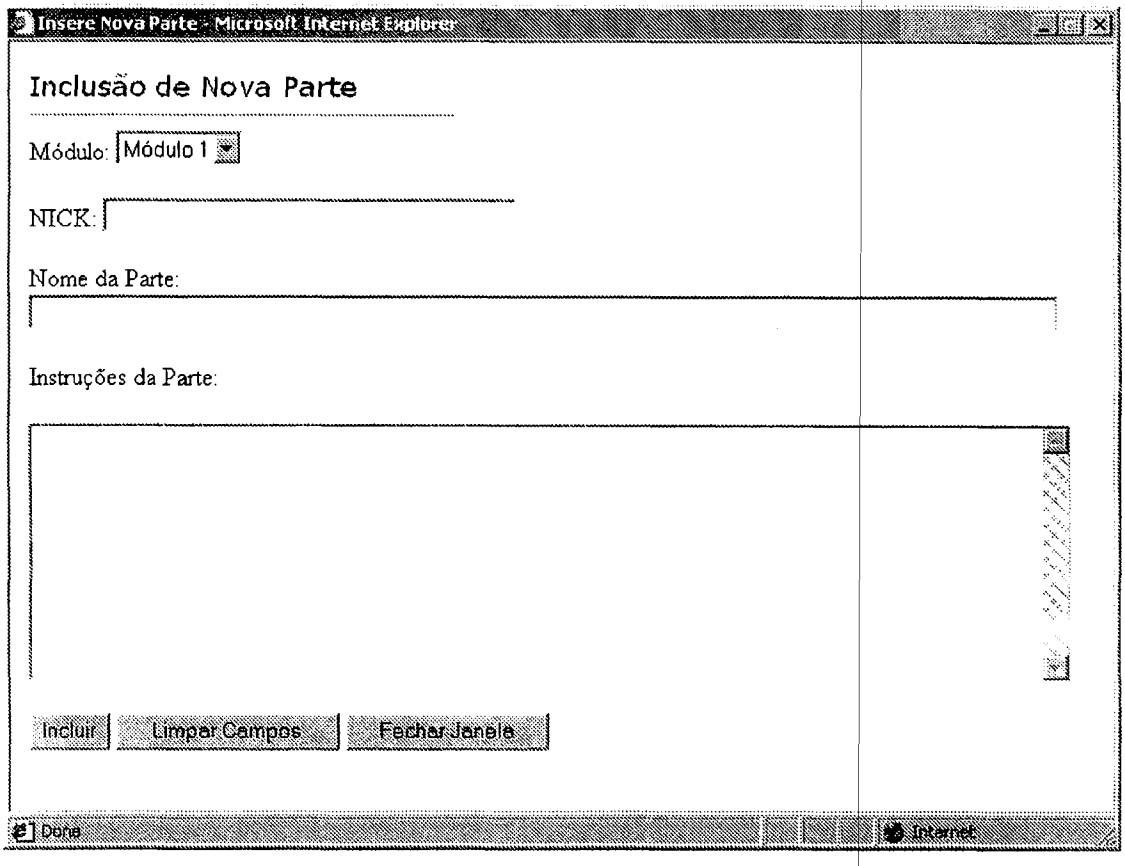

Figura B.7: Janela de inclusão de Nova Parte.

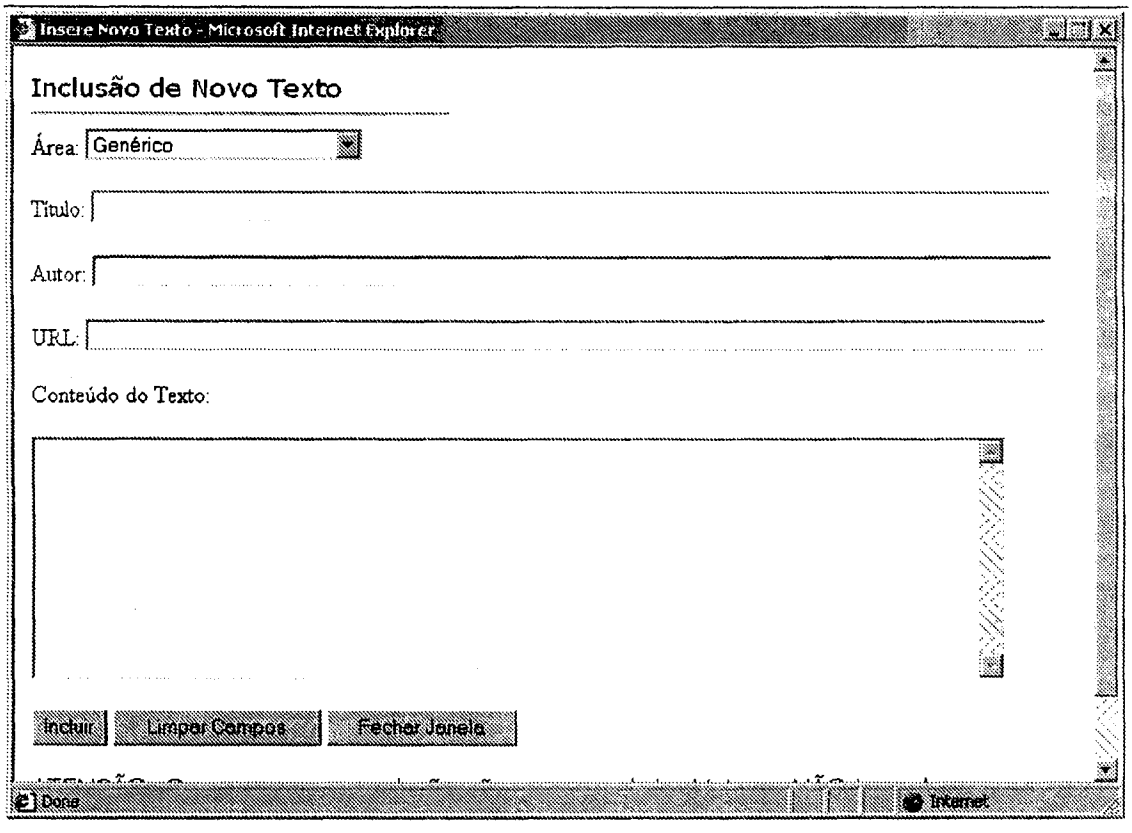

Figura B.8: Janela de inclusão de Novo Texto.

B.2 Diagramas de Casos de Uso

O único usuário do **SisBI** é o professor como indicado na Figura B.9. As operações disponíveis na Figura B.10 podem ser agrupadas de acordo com as três classes: *Questão* (Inserir_Questão, Excluir_Questão, Alterar_Questão, Consultar_Questão), *Texto* (Incluir_Texto, Consultar_Texto) e *Parte* (Incluir_Parte) além da operação Efetuar_Login. Novas operações, por exemplo, Excluir_Parte e Excluir_Texto podem ser implementadas no futuro. As operações são especificadas nos diagramas de atividades das Figuras B.13 a B.21.

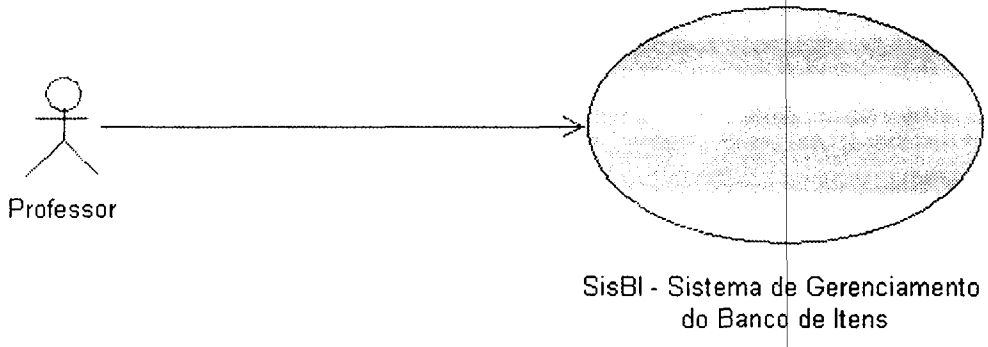

Figura B.9: Diagrama de Caso de Uso Geral.

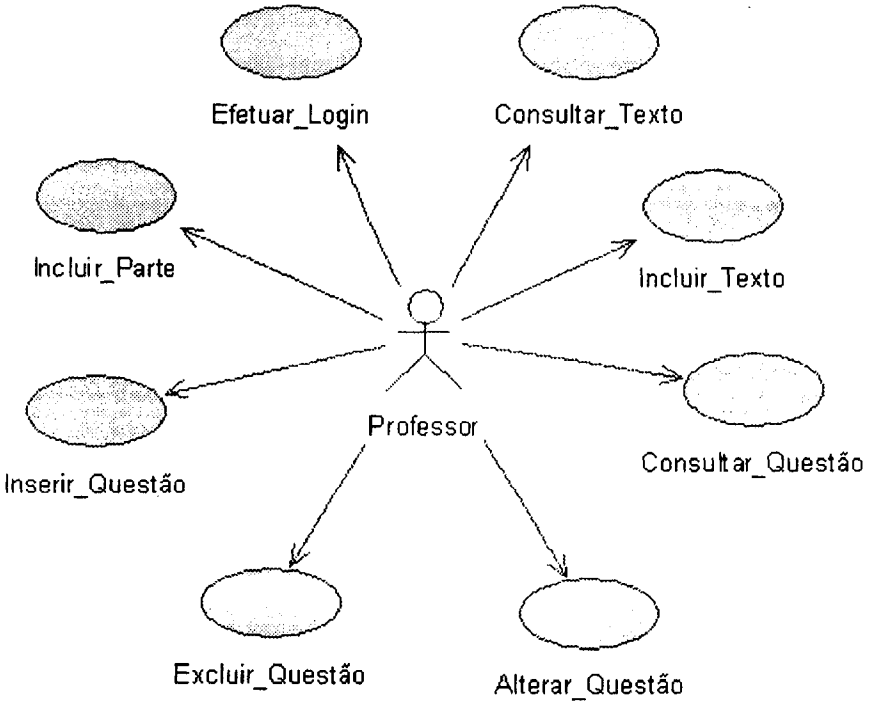

Figura B.10: Diagrama de Caso de Uso do Professor.

B.3 Descrição das Classes

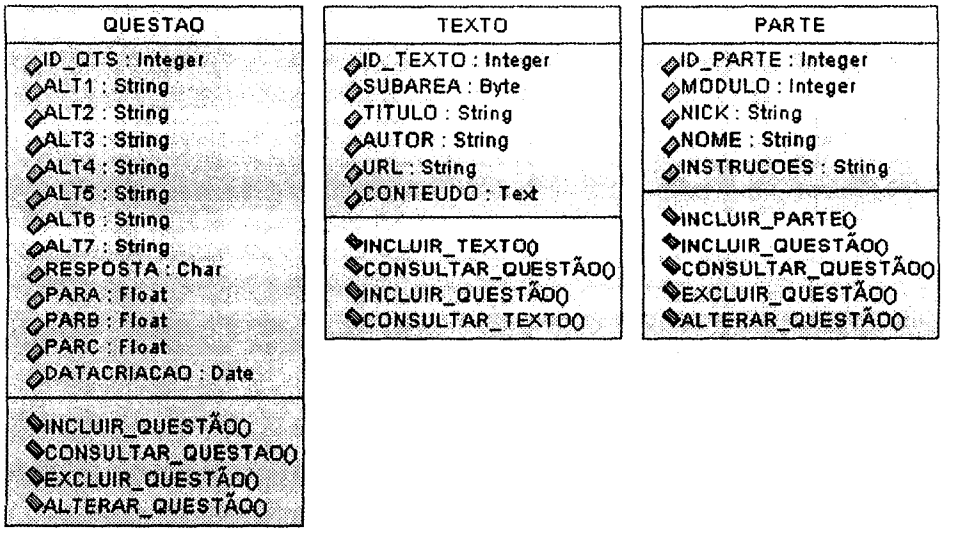

Figura B.11: Descrição das Classes, Atributos e Operações.

B.4 Diagrama de Classes

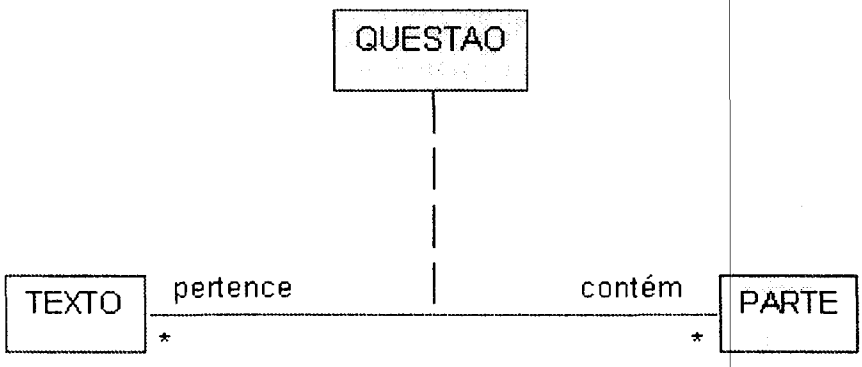

Figura B.12: Diagrama de Classes.

B.5 Diagramas de Atividades

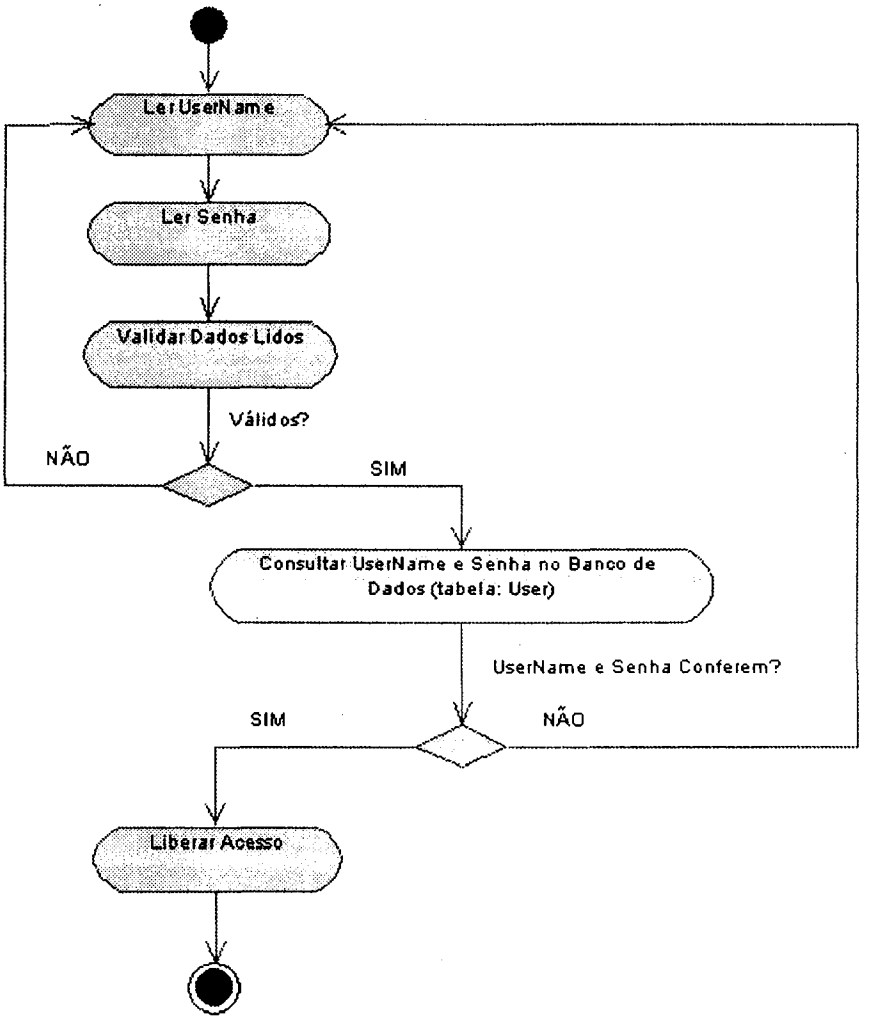

Figura B.13: Diagrama de Atividades – Efetuar Login.

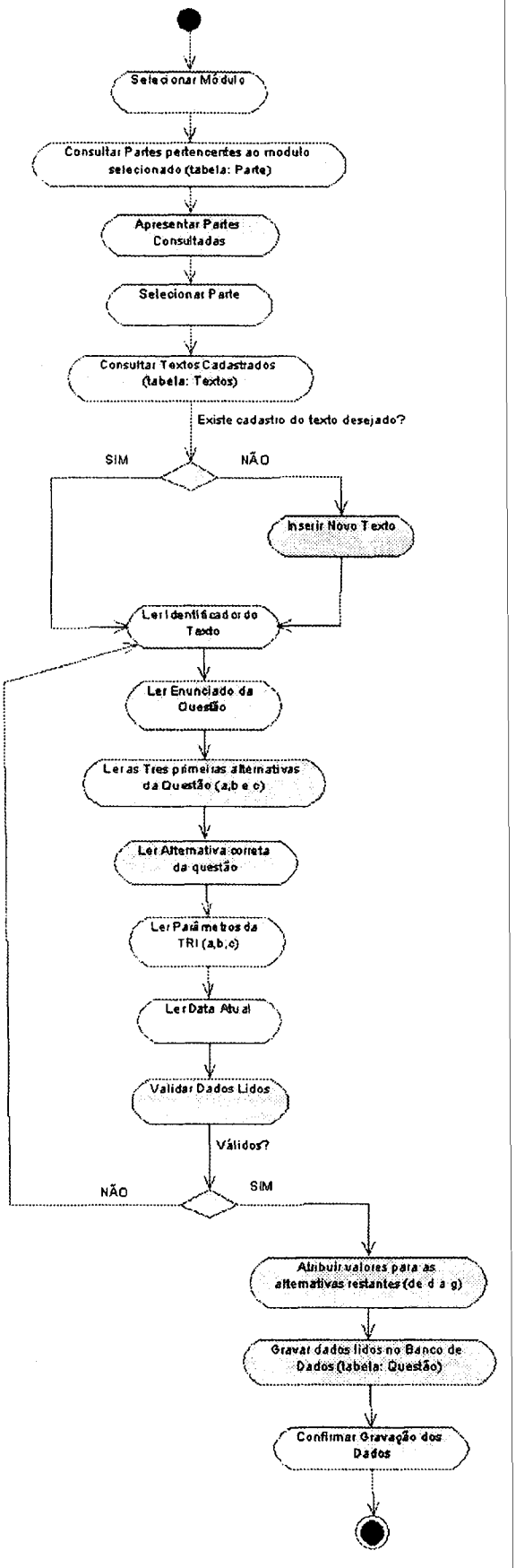

Figura B.14: Diagrama de Atividades – Incluir Questão.

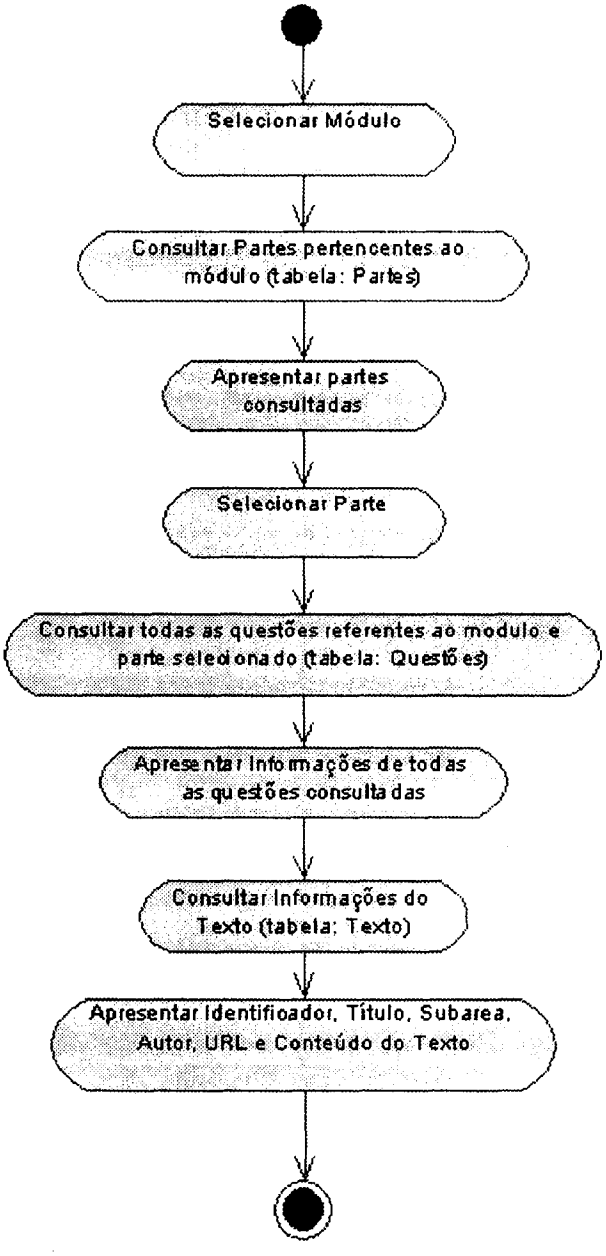

Figura B.15: Diagrama de Atividades – Excluir Questão.

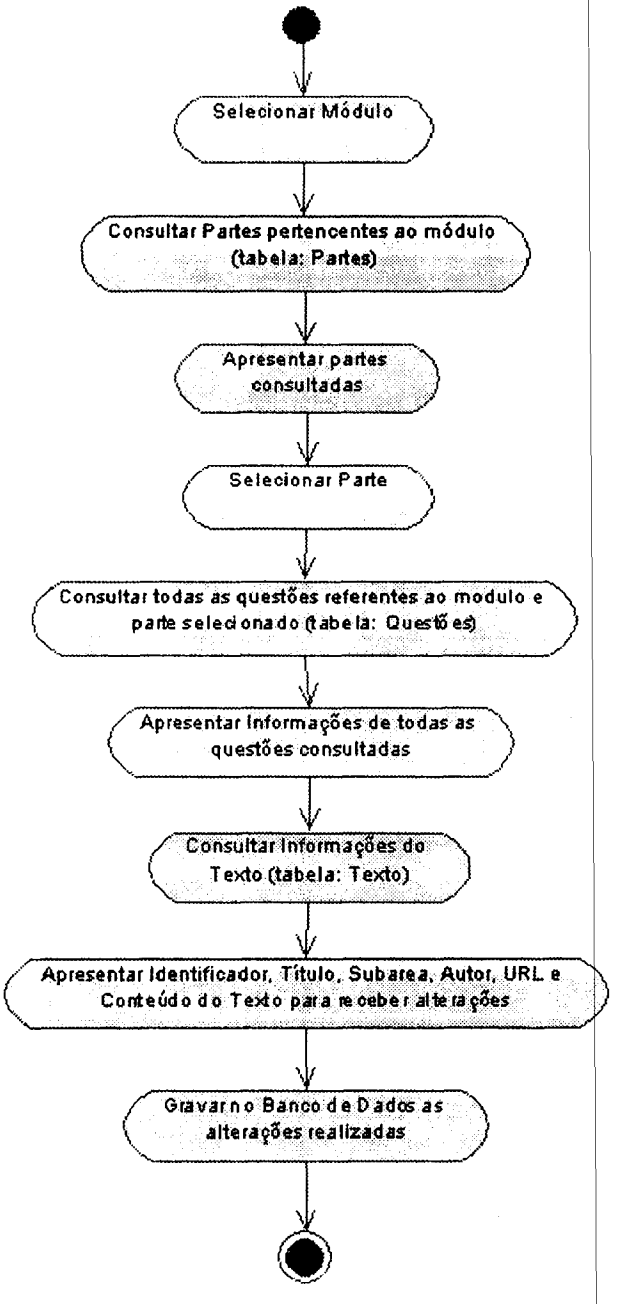

Figura B.16: Diagrama de Atividades – Alterar Questão.

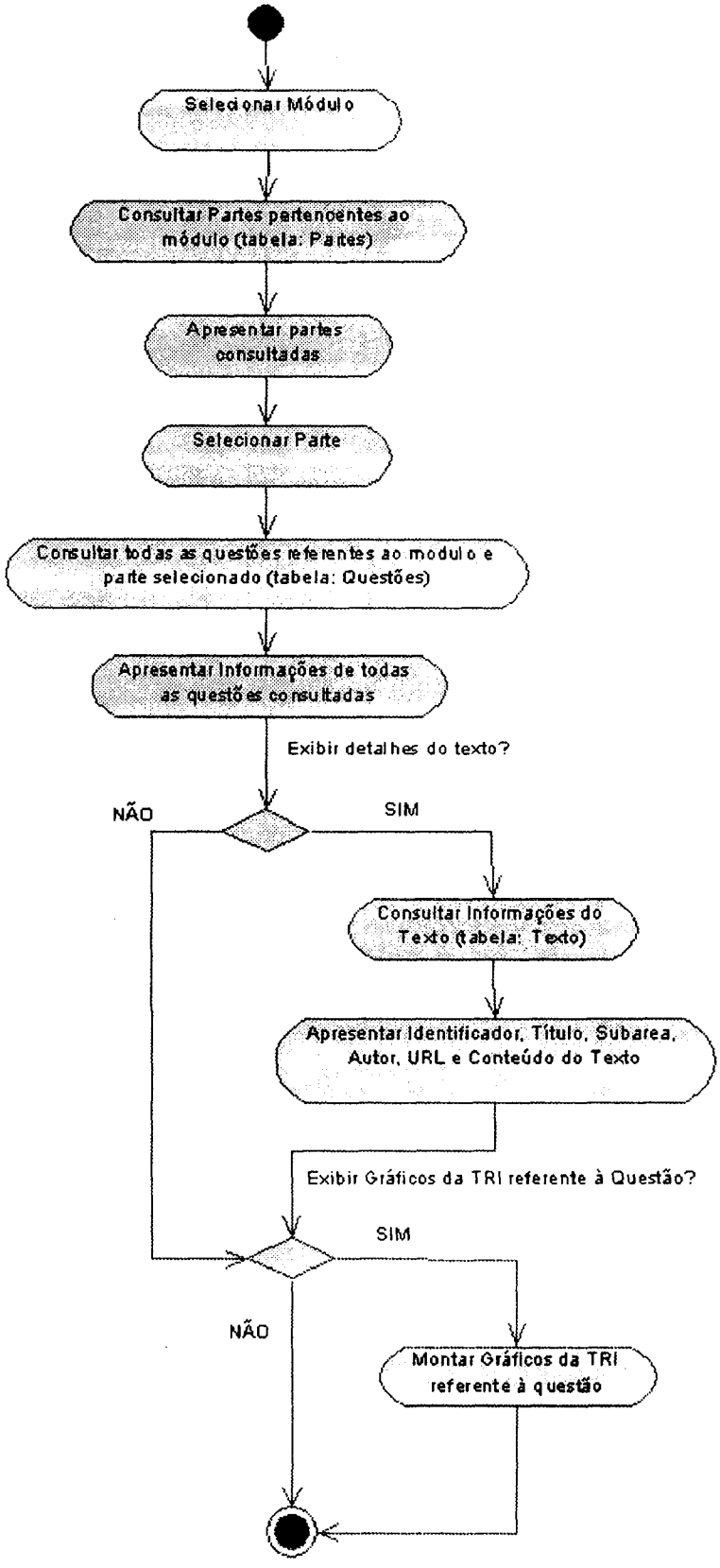

Figura B.17: Diagrama de Atividades – Consultar Questões.

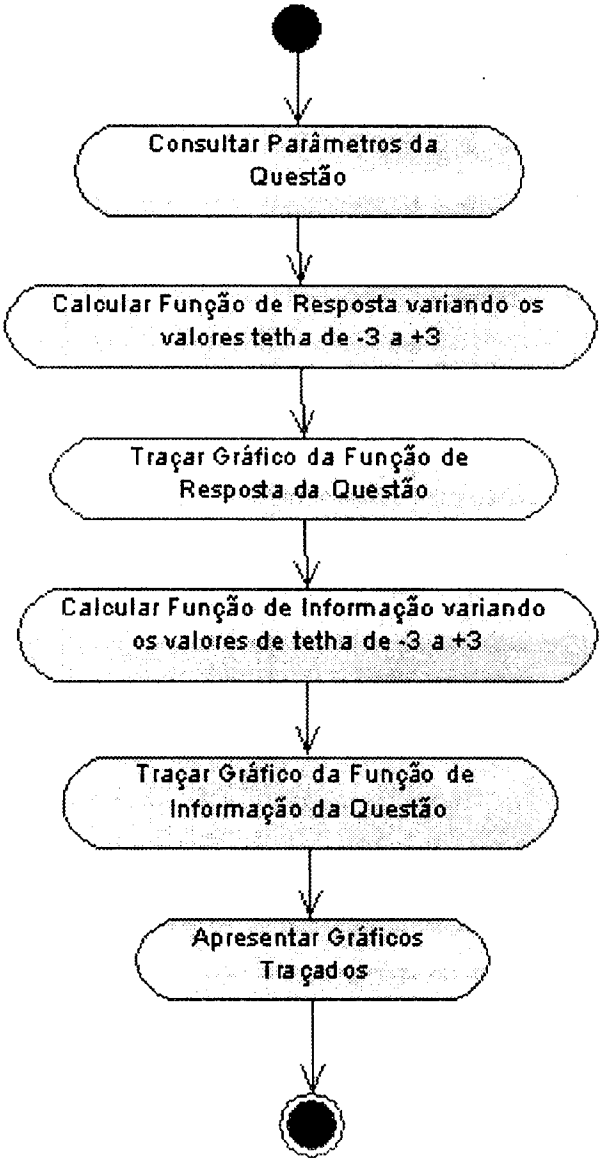

Figura B.18: Diagrama de Atividades – Traçar gráficos da TRI referente a uma questão.

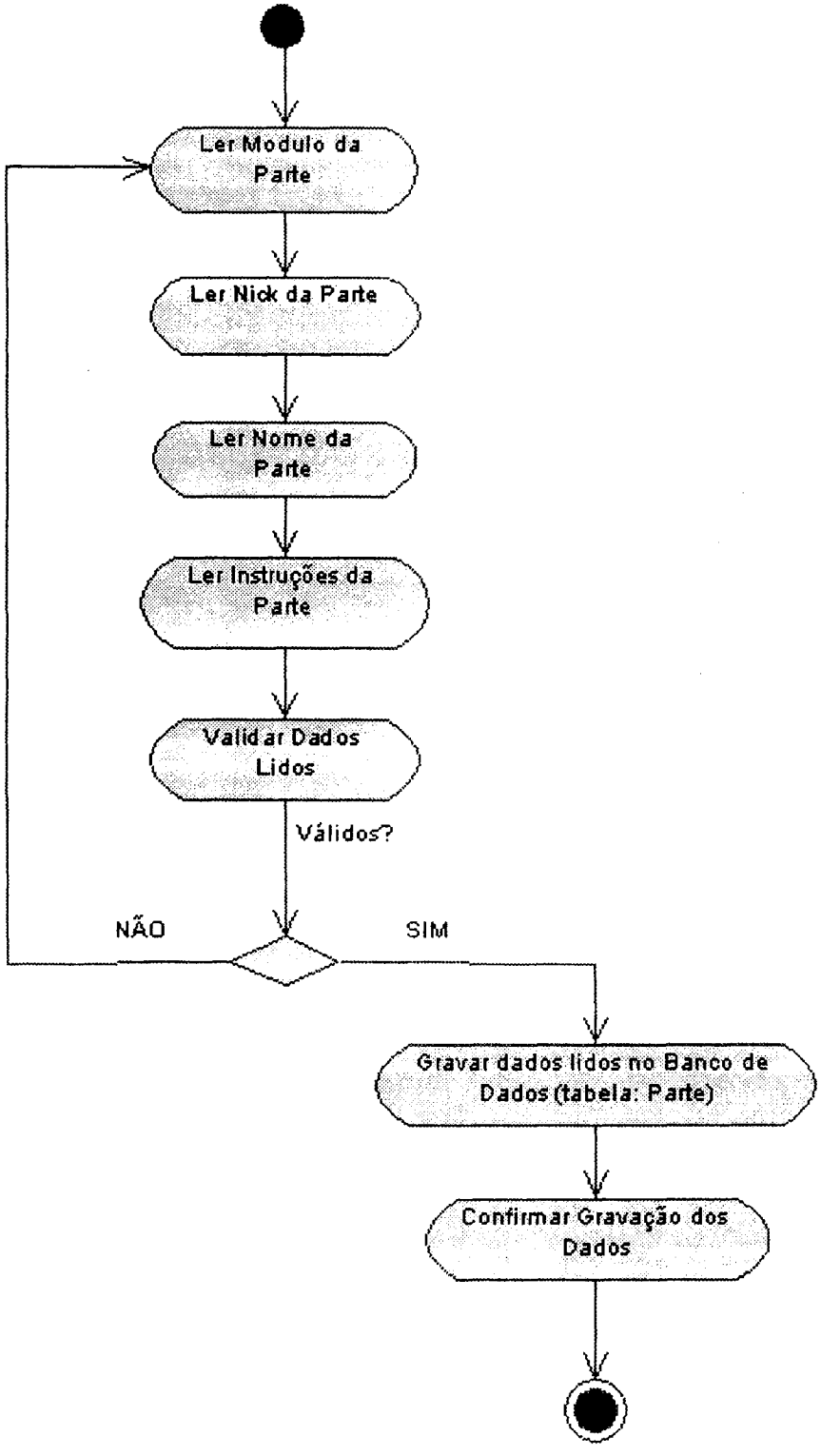

Figura B.19: Diagrama de Atividades – Incluir Nova Parte.

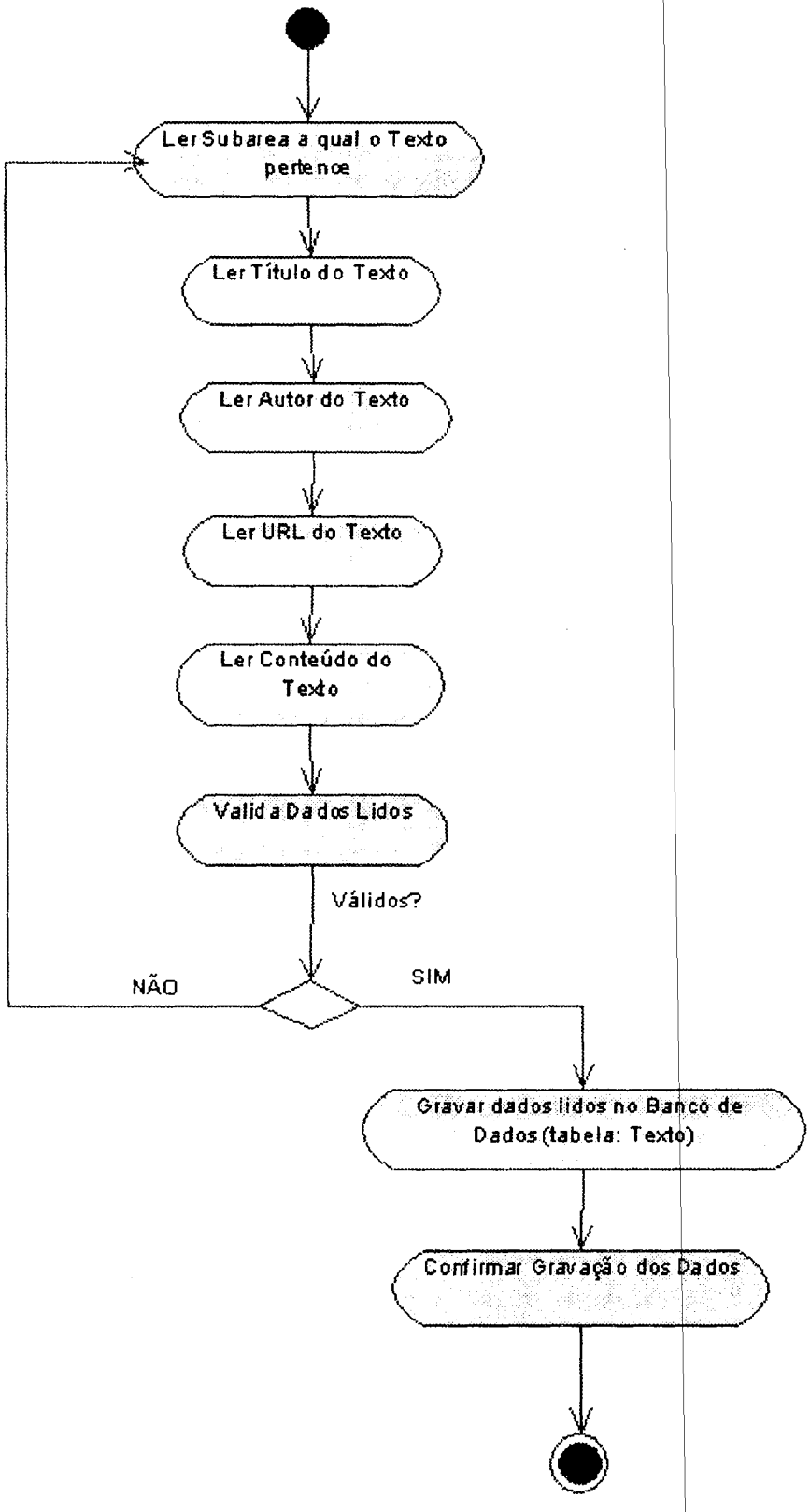

Figura B.20: Diagrama de Atividades – Incluir Novo Texto.

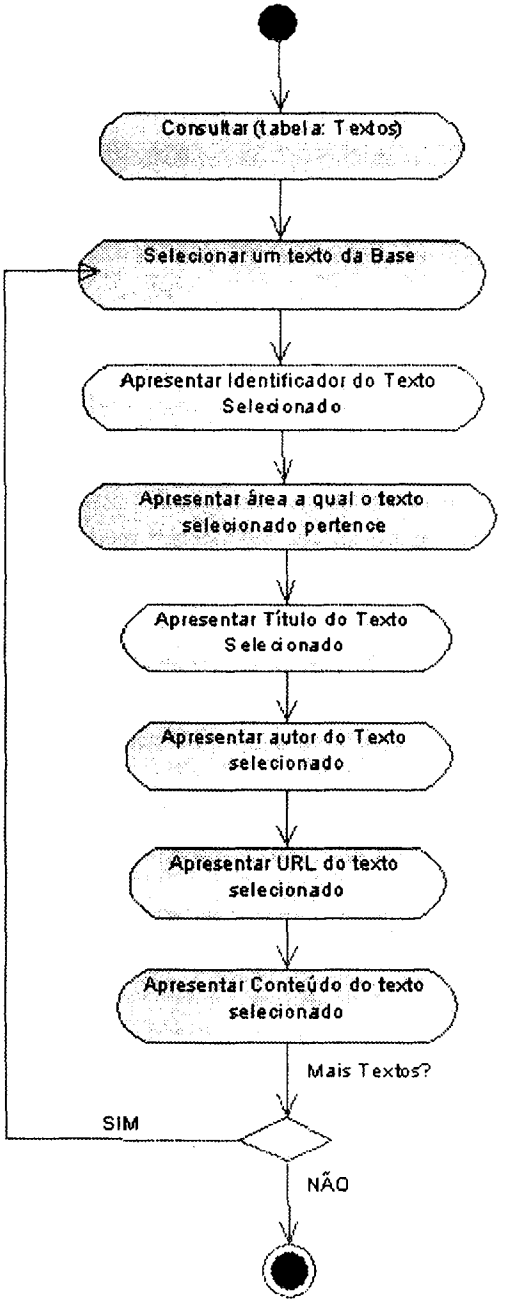

Figura B.21: Diagrama de Atividades - Consultar Textos.

APÊNDICE C

Documentação e Modelagem do Sistema TAEPI

O objetivo deste apêndice é apresentar a documentação e a modelagem do sistema **TAEPI (Testes Adaptativos para o Exame de Proficiência em Inglês)**. São apresentados todos os diagramas da UML que representam o comportamento do sistema, além da descrição completa das classes e operações referentes ao mesmo.

C.1 Documentação das Classes

A descrição das Classes na modelagem de um sistema busca especificar os conceitos importantes do domínio do mesmo, de forma a modelar o conceito de um mundo real. As propriedades e particularidades inerentes a esse conceito são representadas pelos atributos e operações (processos) pertencentes a cada Classe. Os objetos, no contexto de uma classe, são descritos pelo conjunto de atributos e operações que podem ser executadas no domínio que o sistema implementa.

C.1.1 Documentação das Classes e Atributos

Classe: ALUNO
Representa todos os alunos cadastrados no sistema.
Atributos
NUSP: atributo que armazena o número de matrícula de um determinado aluno da USP;
NOME: atributo que armazena o nome completo do aluno;
USERNAME: atributo que armazena o Nome de Acesso (usado, em conjunto com a Senha, para acessar o sistema) do aluno;
SENHA: atributo que armazena a Senha (usada para autenticar o acesso ao sistema) do aluno;
DATA: atributo que armazena a data de cadastro do aluno;
SUBÁREA: atributo que armazena a subárea de pesquisa do aluno.
Operações
CADASTRAR_ALUNO( );
CONSULTAR_RESULTADO( );

EFETUAR_LOGIN( ).

Classe:PROFESSOR

Representa todos os professores cadastrados no sistema.

Atributos

USERNAME: atributo que armazena o Nome do Usuário (usado, em conjunto com a Senha, para acessar o sistema) do professor;

SENHA: atributo que armazena a Senha (usada para autenticar o acesso ao sistema) do professor;

NOME: atributo que armazena o nome completo do professor;

DATA: atributo que armazena a data de cadastro do professor.

Operações

CADASTRAR_PROFESSOR( );

EFETUAR_LOGIN( )

Classe: EXAME

Representa todos os exames cadastrados e disponíveis pelo sistema para serem realizados.

Atributos

NEXAME: atributo de valor único que armazena o número de registro (identificador) de um determinado exame cadastrado;

PARTES: atributo representado por um vetor que armazena os identificadores das partes que compõem um determinado exame. Um exame pode conter uma ou várias partes;

PESOPARTES: atributo representado por um vetor que armazena os pesos respectivos a cada parte que compõe um determinado exame. Como um exame pode conter várias partes cada posição do vetor representa um peso associado a uma parte respectivamente;

DATA: atributo que armazena a data de criação do exame;

Operações

CRIAR_EXAME( );

ESPECIFICAR_CRITERIOS_DE_APLICAÇÃO_DO_EXAME( );

CONSULTAR_RESULTADO( );

CONSULTAR_RESULTADO_GLOBAL( );

REALIZAR_EXAME( );

Classe: CRITÉRIOS_EXAME

Representa os critérios de parada da(s) parte(s) que compõem um determinado exame já cadastrado.

Atributos

CRITERIOS_PARTE: atributo que representa uma matriz de n linhas e 3 (três) colunas que armazena os valores dos critérios de parada de uma determinada parte contida em um exame identificado por NEXAME. Cada linha da matriz representa uma parte incluída no exame. As colunas, por sua vez, representam respectivamente: o VALOR MAXIMO DE TETHA, o VALOR

MÍNIMO DE TETHA e o NÚMERO MÁXIMO DE QUESTÕES ou o NÚMERO MÁXIMO DE TESTLETS (caso a parte seja composta de TestLets) para uma dada parte.

Operações

ESPECIFICAR_CRITERIOS_DE_APLICAÇÃO_DO_EXAME( );
REALIZAR_EXAME( );

Classe:PARTE

Representa todas as partes cadastradas no sistema e que estão disponíveis para serem incluídas em um determinado exame.

Atributos

ID_PARTE: atributo de valor único que identifica o número de registro (identificador) de uma determinada parte;

MODULO: atributo que armazena o valor do módulo a qual a parte é associada;

NICK: atributo que armazena o apelido da parte. Apelido pode ser entendido nesse contexto como a maneira mais comum que uma determinada parte é identificada, por exemplo: GAP, REVIEW, ABSTRACT etc;

NOME: atributo que armazena o nome (propriamente dito) de uma determinada parte;

INSTRUÇÕES: atributo do tipo texto que armazena as instruções de uma determinada parte.

Operações

CRIAR_EXAME( ).

Classe: TEXTO

Representa todos os textos cadastrados no sistema e que estão disponíveis para serem associados a uma ou mais questões. A associação de um mesmo texto a duas ou mais questões diferentes representa um TestLet.

Atributos

ID_TEXTO: atributo de valor único que identifica o número de registro (identificador) de um determinado texto;

SUBÁREA: representa a área à qual o texto pertence. As áreas podem ser: Computação, Estatística, Matemática Computacional ou Genérica (no caso de o texto servir para avaliar conhecimentos comuns a todas as outras);

TÍTULO: atributo que armazena o título do texto;

AUTOR: atributo que armazena o nome do autor do texto;

URL: atributo que armazena a URL, ou endereço eletrônico completo, de um determinado texto;

CONTEÚDO: atributo do tipo texto que armazena conteúdo do texto;

Operações

REALIZAR_EXAME( )

Classe: QUESTÃO

Representa todas as questões cadastradas no sistema e que estão disponíveis para serem utilizadas em um determinado exame.

Atributos

ID_QTS: atributo de valor único que identifica o número de registro (identificador) de uma determinada questão;

ALT1 a ALT7: atributos que armazenam as alternativas de uma determinada questão (podem ser sete alternativas);

RESPOSTA: atributo que armazena a resposta correta da questão;

PARA: atributo que representa o valor do parâmetro de Discriminalidade da questão;

PARB: atributo que representa o valor do parâmetro de Dificuldade da questão;

PARC: atributo que representa o valor do parâmetro de Adivinhação da questão;

DATA: atributo que armazena a data de cadastro da questão.

Operações

REALIZAR_EXAME( ).

Classe: RESULTADO

Representa todos os resultados finais dos exames realizados pelos alunos.

Atributos

PARTE: atributo que identifica uma determinada parte que foi avaliada em um Exame;

QTDE_QTSFEITAS: atributo que representa o número de questões realizadas pelo aluno em um determinado exame identificado por NEXAME;

TETHA_FINAL: atributo que representa o valor final de Tetha (habilidade) obtido por um determinado aluno em um exame identificado por NEXAME;

PONTUAÇÃO_FINAL: atributo que representa o valor da pontuação final obtida por um determinado aluno em um exame identificado por NEXAME.

Operações

REALIZAR_EXAME( );

CONSULTAR_RESULTADO( );

CONSULTAR_RESULTADO_GLOBAL( ).

Classe: RESPOSTA

Representa as respostas dadas às questões, os valores de tetha parcial e a pontuação parcial de um determinado exame realizado por um aluno.

Atributos

PARTE: atributo que identifica uma determinada parte que foi avaliada em um Exame;

RESPOSTA_DADA: atributo que representa a resposta dada pelo aluno a uma questão que foi administrada. Pode ser correta/incorreta;

TETHA_PARCIAL: atributo que representa o valor parcial de Tetha (habilidade) obtido por determinado aluno durante a realização de um exame identificado por NEXAME;

PONTUAÇÃO_PARCIAL: atributo que representa o valor da pontuação parcial obtida por um determinado aluno durante a realização de um exame identificado por NEXAME.
Operações
REALIZAR_EXAME( );
CONSULTAR_RESULTADO_DETALHADO( ).

C.1.2 Documentação das Operações

- CADASTRAR_ALUNO: operação que realiza o cadastro de um determinado aluno no sistema;
- CONSULTAR_RESULTADO: operação que permite a um determinado aluno cadastrado, e que realizou um Exame, verificar seu resultado;
- EFETUAR_LOGIN: função que libera ou não o acesso ao sistema para um determinado usuário;
- CADASTRAR_PROFESSOR: operação que realiza o cadastro de um determinado professor no sistema;
- CRIAR_EXAME: operação que realiza a criação e cadastro de um determinado exame no sistema. Nela é possível declarar as partes que compõem um exame;
- ESPECIFICAR_CRITERIOS_DE_APLICAÇÃO_DO_EXAME: operação que permite a um professor cadastrado definir os critérios de parada de cada parte (individualmente) inserida em um determinado exame;
- CONSULTAR_RESULTADO: operação que permite a um determinado aluno verificar seu(s) resultado(s) em um exame realizado;
- CONSULTAR_RESULTADO_GLOBAL: operação que permite a um determinado professor cadastrado verificar os resultados de um ou vários alunos;
- REALIZAR_EXAME: operação que permite a um aluno realizar um determinado exame que foi cadastrado e especificado;
- CONSULTAR_RESULTADO_DETALHADO: operação que permite a um professor acompanhar o desempenho de um aluno em um determinado exame, verificando a resposta dada, e os valores de tetha e a pontuação parcial após cada questão fornecida e respondida pelo aluno;
- CONSULTAR_GABARITO: operação que possibilita ao aluno verificar o gabarito de um determinado exame que ele tenha realizado.

C.2 Documentação do Diagrama de Classes

O Diagrama de Classes lista todos os conceitos de um determinado domínio que serão implementados pelo sistema. Em outras palavras, ele define a relação estática entre os objetos pertencentes às classes e é muito importante porque define a estrutura do sistema. Além disso, a maioria dos outros diagramas da modelagem do sistema utiliza o diagrama de classes.

Veja a Figura C.6 para uma consulta mais detalhada do diagrama de classes.

C.3 Documentação dos Diagramas de Casos de Uso

Os Diagramas de Casos de Uso representam um conjunto de funcionalidades do sistema. Visto por esse ângulo, os casos de uso descrevem os cenários identificados da relação entre os usuários (administrador, aluno e professor) e o sistema, de maneira a mostrar como será a interação entre eles. Um caso de uso descreve as operações (funções) que o sistema deve cumprir para cada usuário, formalizando assim as tarefas que precisam ser feitas.

Tendo por base os diagramas de casos de uso, o sistema de testes adaptativos informatizados oferece funcionalidades específicas de acordo com os tipos de usuários que, no caso, podem ser: o administrador, os professores ou os alunos. Assim, a única tarefa que o administrador pode desempenhar é Cadastrar Professor (veja Figura C.2). O professor pode executar as seguintes operações: Consultar Resultado Detalhado, Efetuar Login, Criar Exame, Especificar Critérios de aplicação do Exame e Consultar Resultado Global (veja Figura C.3). Por sua vez, o aluno pode desempenhar as tarefas de: Cadastrar Aluno, Efetuar Login, Realizar Exame, Consultar Gabarito e Consultar Resultado (veja Figura C.4).

Segue abaixo uma descrição detalhada de cada tarefa que pode ser desempenhada pelos diversos usuários do sistema.

C.3.1 Tarefa: CADASTRAR PROFESSOR

Função: realizar o cadastro de professores que estejam interessados em utilizar o sistema, de maneira a liberar o acesso para realizarem as tarefas de Consultar Resultado Detalhado, Efetuar Login, Criar Exame, Especificar Critérios de Aplicação do Exame e Consultar Resultado Global. Acontece quando o administrador cadastra o professor interessado no sistema, fornecendo as seguintes informações: Nome, UserName, Senha, e Data.

C.3.2 Tarefa: EFETUAR LOGIN

Função: permitir ou não o acesso do usuário (professor ou aluno) ao sistema. Ocorre quando o usuário, previamente cadastrado, abre a página de acesso ao sistema e fornece um UserName e uma Senha, os quais serão verificados quanto à correção, de forma que o acesso possa ser liberado ou não.

C.3.3 Tarefa: CRIAR EXAME

Função: possibilita que o professor previamente cadastrado crie um exame que possa ser aplicado aos alunos usando o método adaptativo. O professor, quando da criação do exame, fornece um identificador do exame (único), define as partes (dependendo da habilidade/conhecimento que se quer avaliar) que serão incluídas no mesmo e a data de criação.

C.3.4 Tarefa: ESPECIFICAR CRITÉRIO DE APLICAÇÃO DO EXAME

Função: permite que o professor previamente cadastrado especifique os critérios de parada de um determinado exame de acordo com as partes pertencentes ao mesmo. É possível definir os valores máximo e mínimo dos níveis de habilidade desejados (valor de Tetha) da parte, a quantidade de questões que se deseja administrar ao aluno, e, para as partes do exame compostas de TestLets, é possível declarar a quantidade de TestLets requeridos.

C.3.5 Tarefa: CONSULTAR RESULTADO GLOBAL

Função: permite que o professor previamente cadastrado verifique os resultados de um determinado exame já realizado. O professor tem duas opções: 1) escolher entre consultar o resultado de um aluno individualmente, e 2) verificar os resultados de todos os alunos que realizaram um determinado exame, podendo saber ainda a média do valor da habilidade (tetha) e da pontuação. Na visualização dos resultados são apresentadas as seguintes informações: número USP do aluno (NUSP), número do exame requerido (NEXAME), identificação da parte avaliada, quantidade de questões feitas em cada parte, valor de habilidade (tetha) final e a pontuação final.

C.3.6 Tarefa: CONSULTAR RESULTADO DETALHADO

Função: permite que o professor previamente cadastrado verifique o desempenho de um determinado aluno em um exame já realizado. O professor pode verificar as partes avaliadas, o conjunto de questões fornecidas ao aluno durante o exame, suas respostas fornecidas e seus resultados parciais, como o valor de tetha (habilidade) e pontuação. O professor deve fornecer o número do exame desejado (NEXAME) e o número USP de cadastro do aluno (NUSP). As informações de desempenho são as seguintes: número do exame realizado, número USP do aluno, identificador da parte avaliada, identificador da questão fornecida, a resposta dada, o tetha e a pontuação parcial de cada questão durante o exame.

C.3.7 Tarefa: CADASTRAR ALUNO

Função: tem função de realizar o cadastro de alunos que utilizarão o sistema para realizar um determinado exame sob o método adaptativo informatizado. O cadastro do aluno libera o acesso para o mesmo realizar as tarefas de Efetuar Login, Consultar Resultado e Realizar Exame. Acontece quando um aluno abre a página de acesso do sistema (primeira página) e escolhe a opção cadastrar,

fornecendo as seguintes informações: Número USP (NUSP), Nome, UserName, Senha, Curso e Data.

C.3.8 Tarefa: REALIZAR EXAME

Função: tem função de aplicar um determinado exame usando o método adaptativo aos alunos cadastrados no sistema. Acontece quando o aluno acessa o sistema e escolhe a opção realizar exame. A função aplica todas as partes que compõem um exame criado anteriormente, respeitando todos os critérios de parada previamente definidos pelo professor e implementando todos os procedimentos e métodos exigidos pelo modelo adaptativo. Durante o exame podem ser aplicadas partes que contenham tanto questões individuais (que não pertencem a um TestLet) quanto os grupos de questões relativas a um texto (TestLet), dependendo somente das partes incluídas em um determinado exame.

C.3.9 Tarefa: CONSULTAR RESULTADO

Função: tarefa que permite a um aluno previamente cadastrado verificar seus resultados em determinado exame. O aluno dever fornecer o número do exame desejado (NEXAME), e sua Senha de acesso ao sistema. Depois da autenticação da senha fornecida, o sistema apresenta o resultado do exame com as seguintes informações: número USP do aluno (NUSP), identificador da parte avaliada, número do exame requerido (NEXAME), quantidade de questões feitas, valor de habilidade (tetha) final e a pontuação final em cada parte.

C.3.10 Tarefa: CONSULTAR GABARITO

Função: tarefa que possibilita a um aluno previamente cadastrado, e que realizou um determinado exame, verificar o gabarito do mesmo. O aluno deve fornecer o Número do Exame realizado, o número USP. As informações apresentadas ao aluno são: o identificador da parte, a questão fornecida, a resposta dada e a resposta correta da mesma.

C.4 Descrição dos Diagramas da UML

C.4.1 Diagramas de Casos de Uso

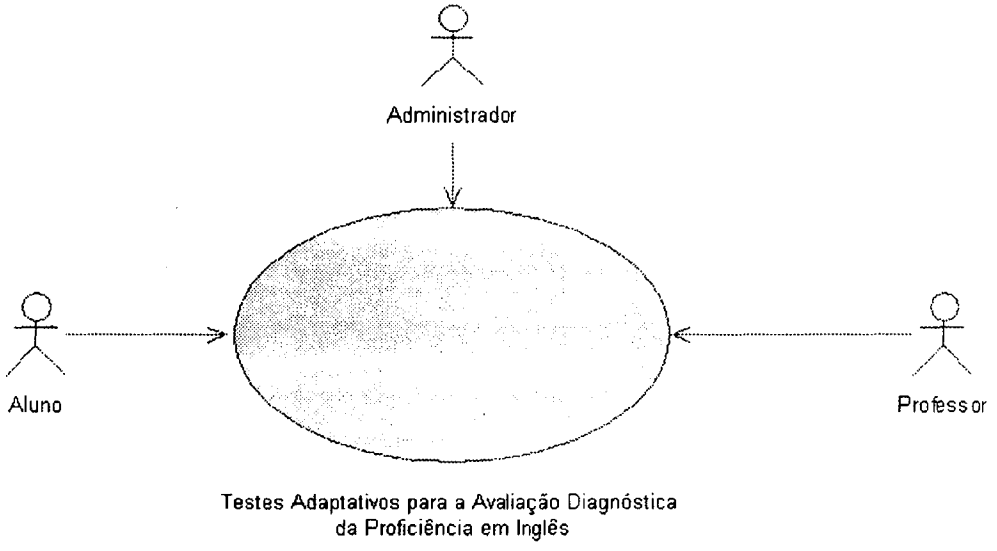

Figura C.1: Diagrama de Caso de Uso Geral do TAEPI.

C.4.1.1 Diagrama de Caso de Uso do Administrador

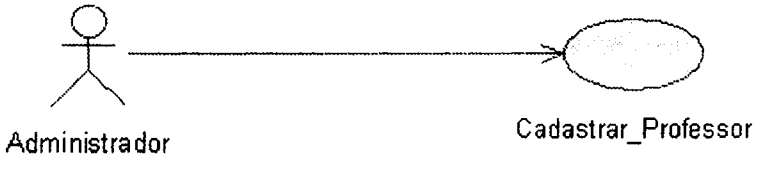

Figura C.2: Diagrama de Caso de Uso do Administrador.

C.4.1.2 Diagrama de Caso de Uso do Professor

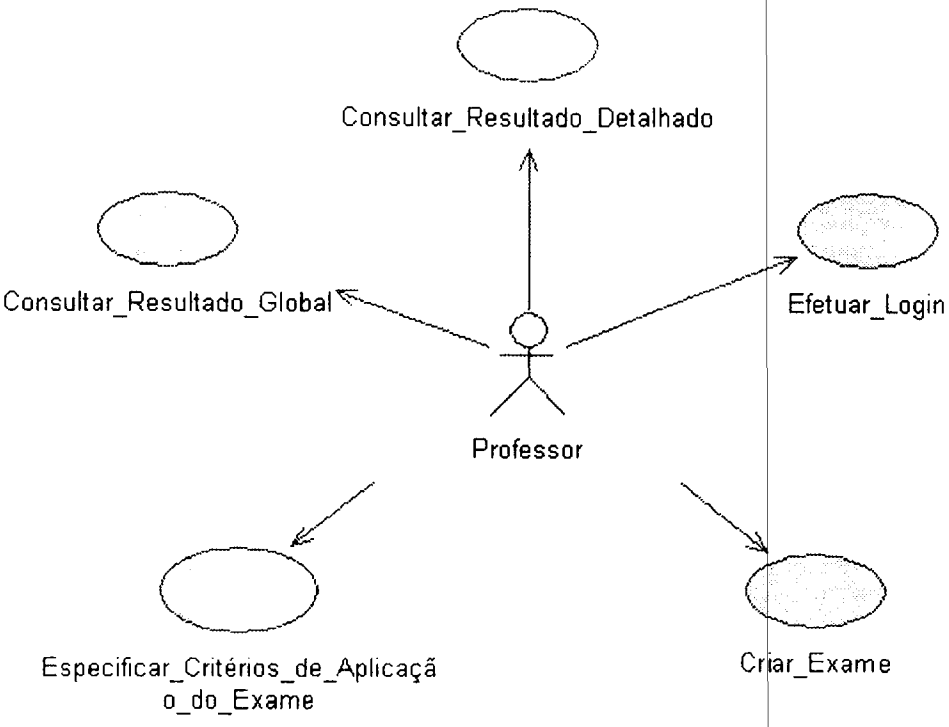

Figura C.3: Diagrama de Caso de Uso do Professor.

C.4.1.3 Diagrama de Caso de Uso do Aluno

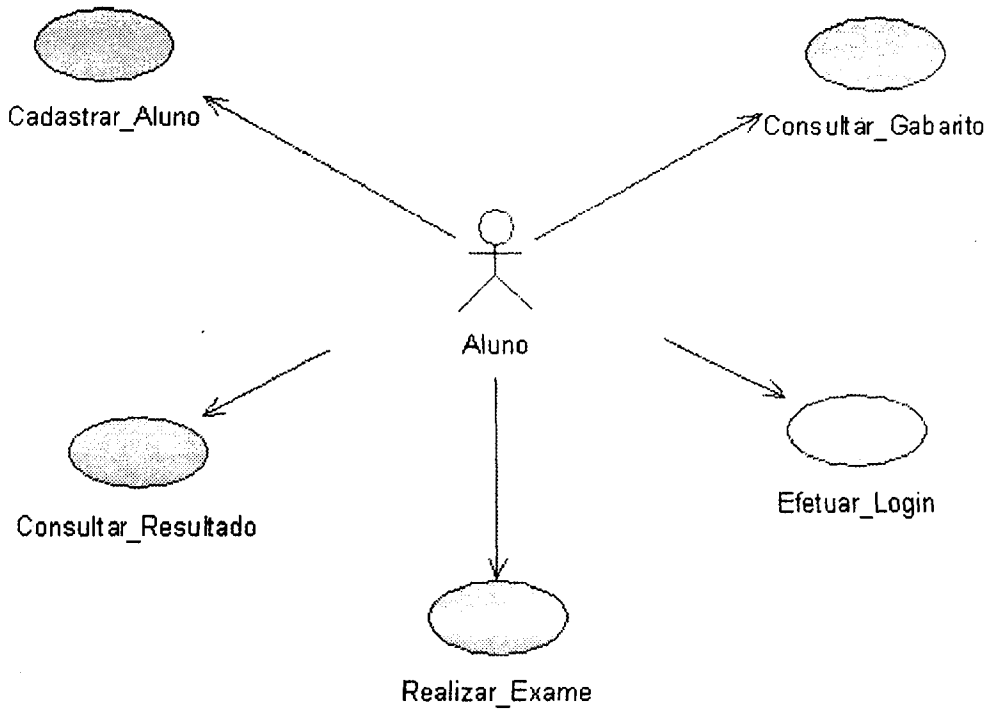

Figura C.4: Diagrama de Caso de Uso do Aluno.

C.4.2 Descrição das Classes

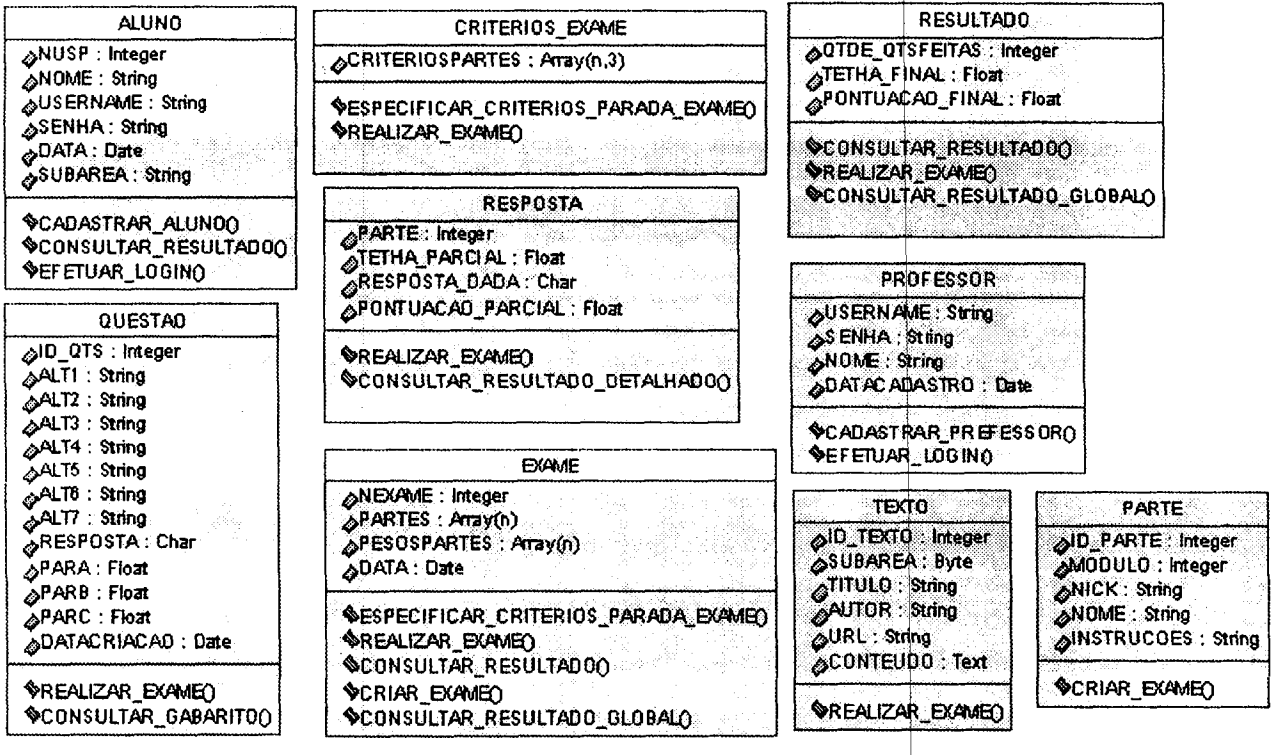

Figura C.5: Descrição da Classes, Atributos e Operações.

C.4.3 Diagrama de Classes

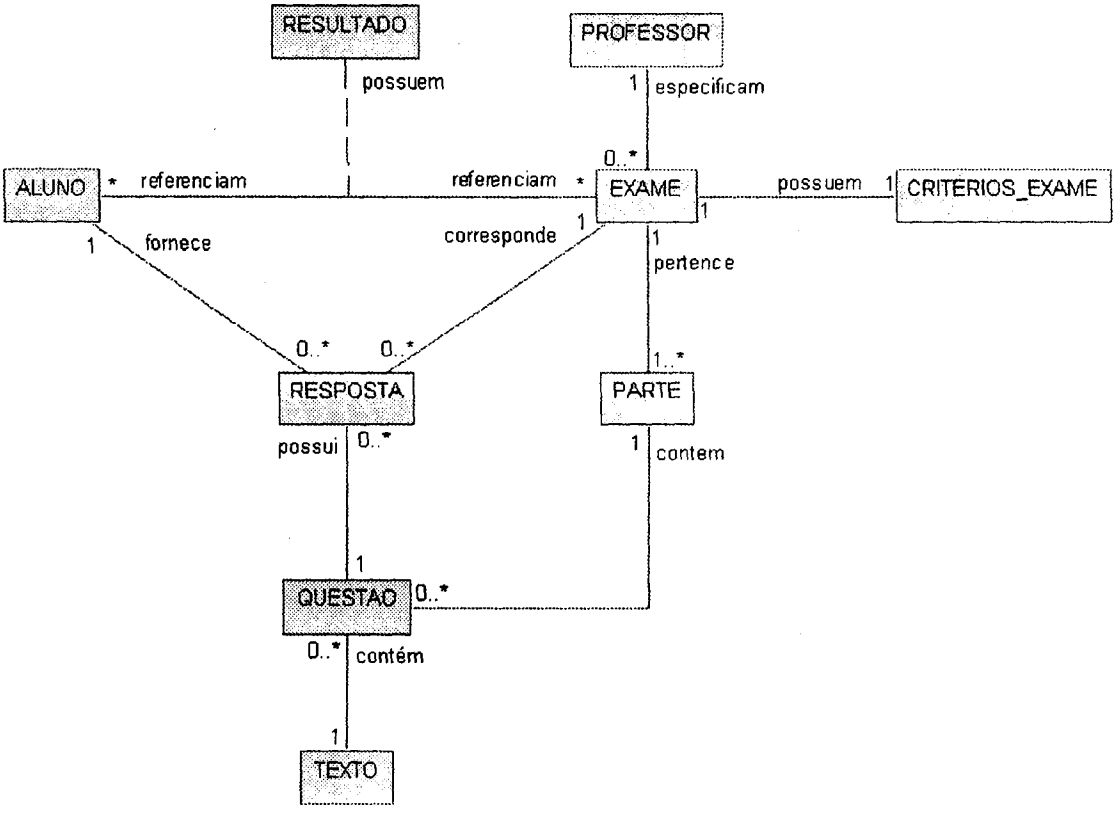

Figura C.6: Diagrama de Classes.

C.4.4 Diagramas de Atividades

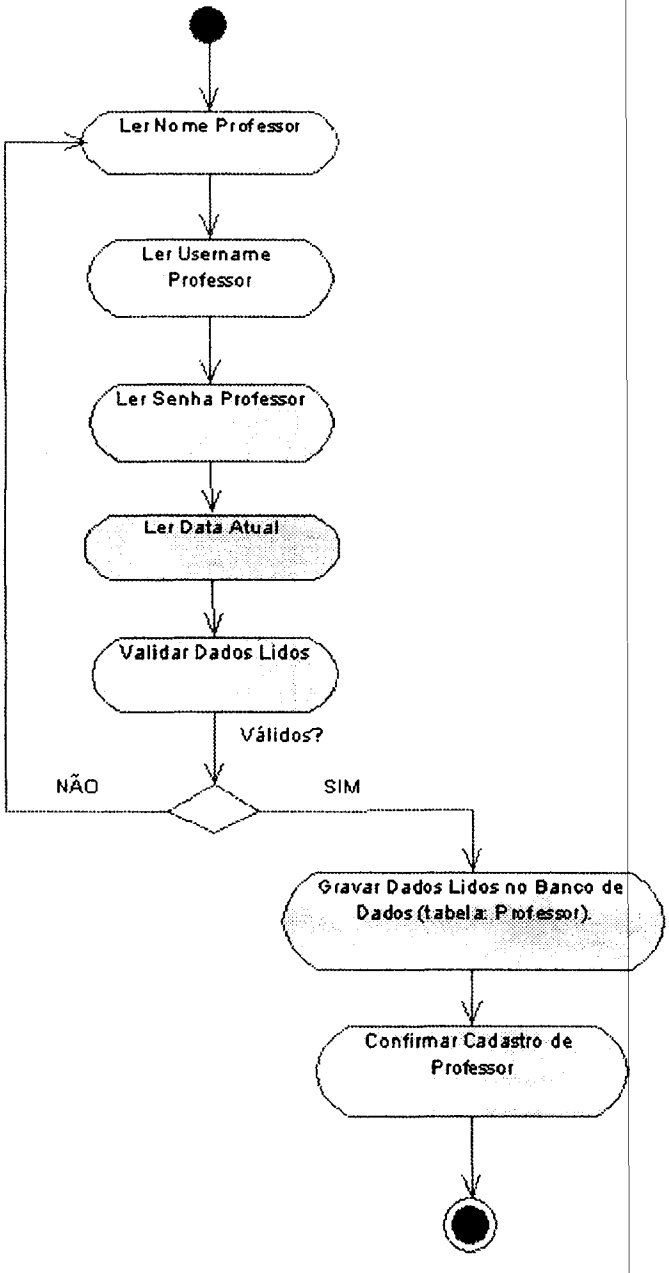

Figura C.7: Diagrama de Atividades – Cadastrar Professor.

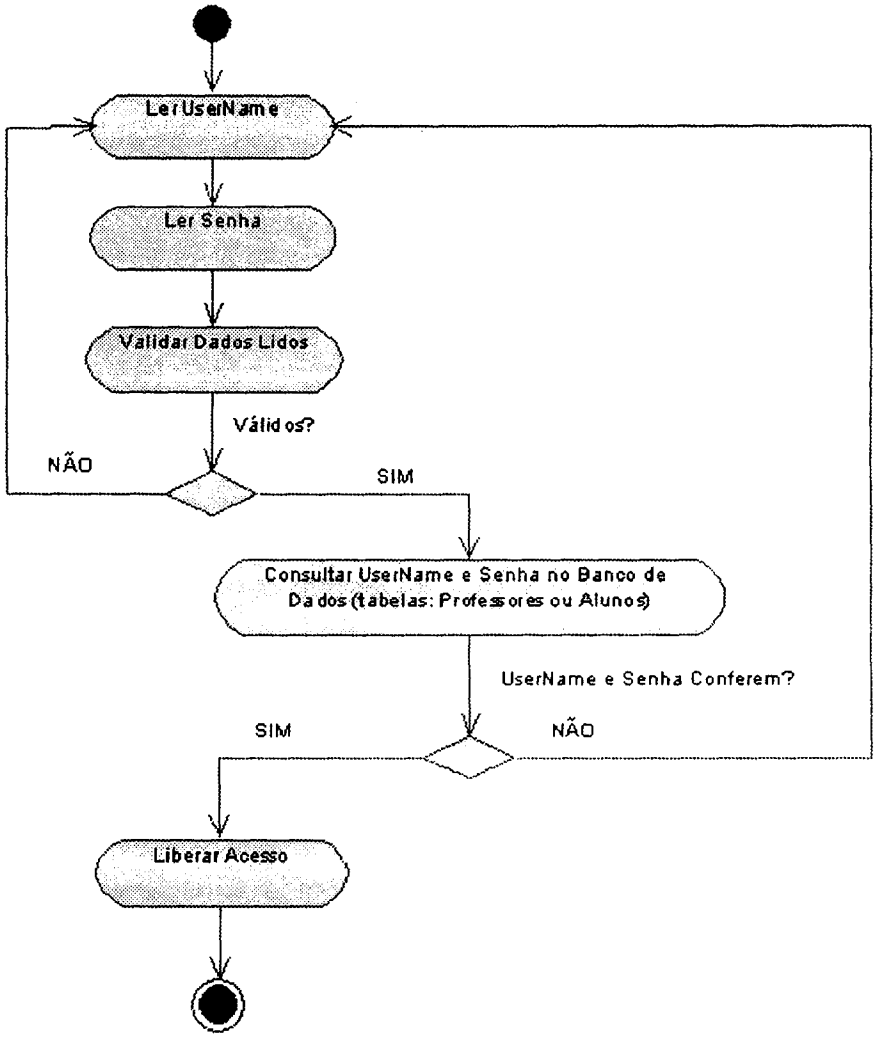

Figura C.8: Diagrama de Atividades – Efetuar Login.

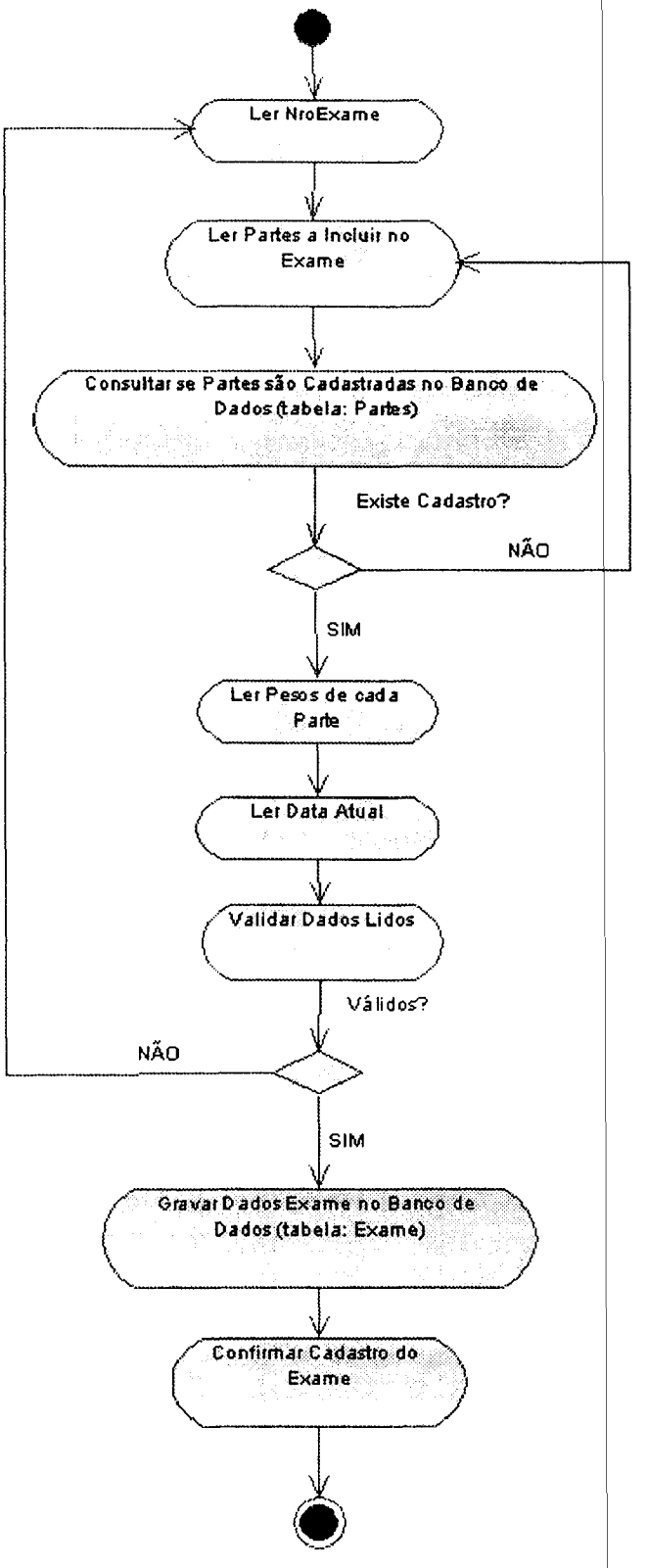

Figura C.9: Diagrama de Atividades – Criar Exame.

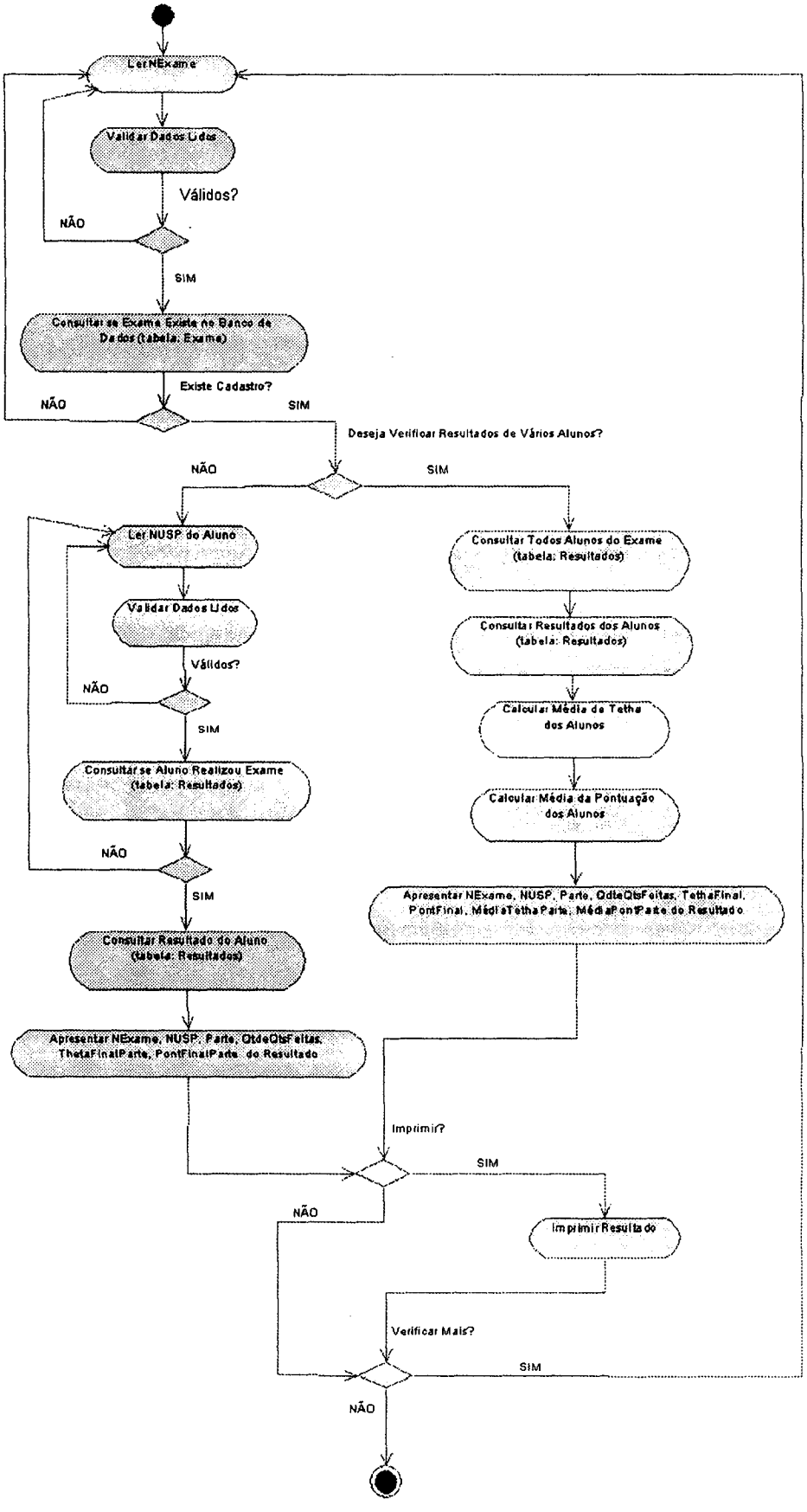

Figura C.10: Diagrama de Atividades – Consultar Resultado Global.

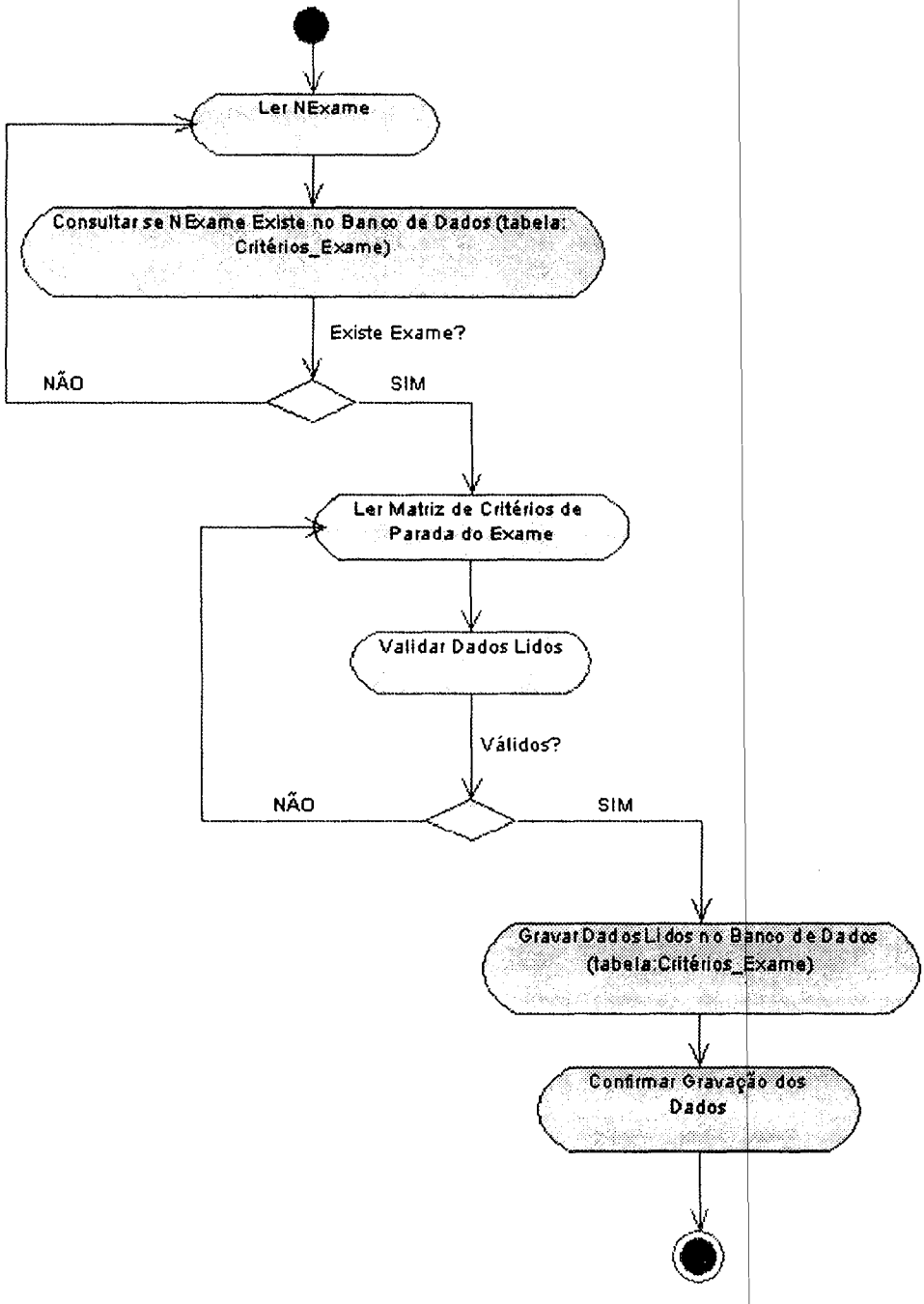

Figura C.11: Diagrama de Atividades – Especificar Critérios de Aplicação do Exame.

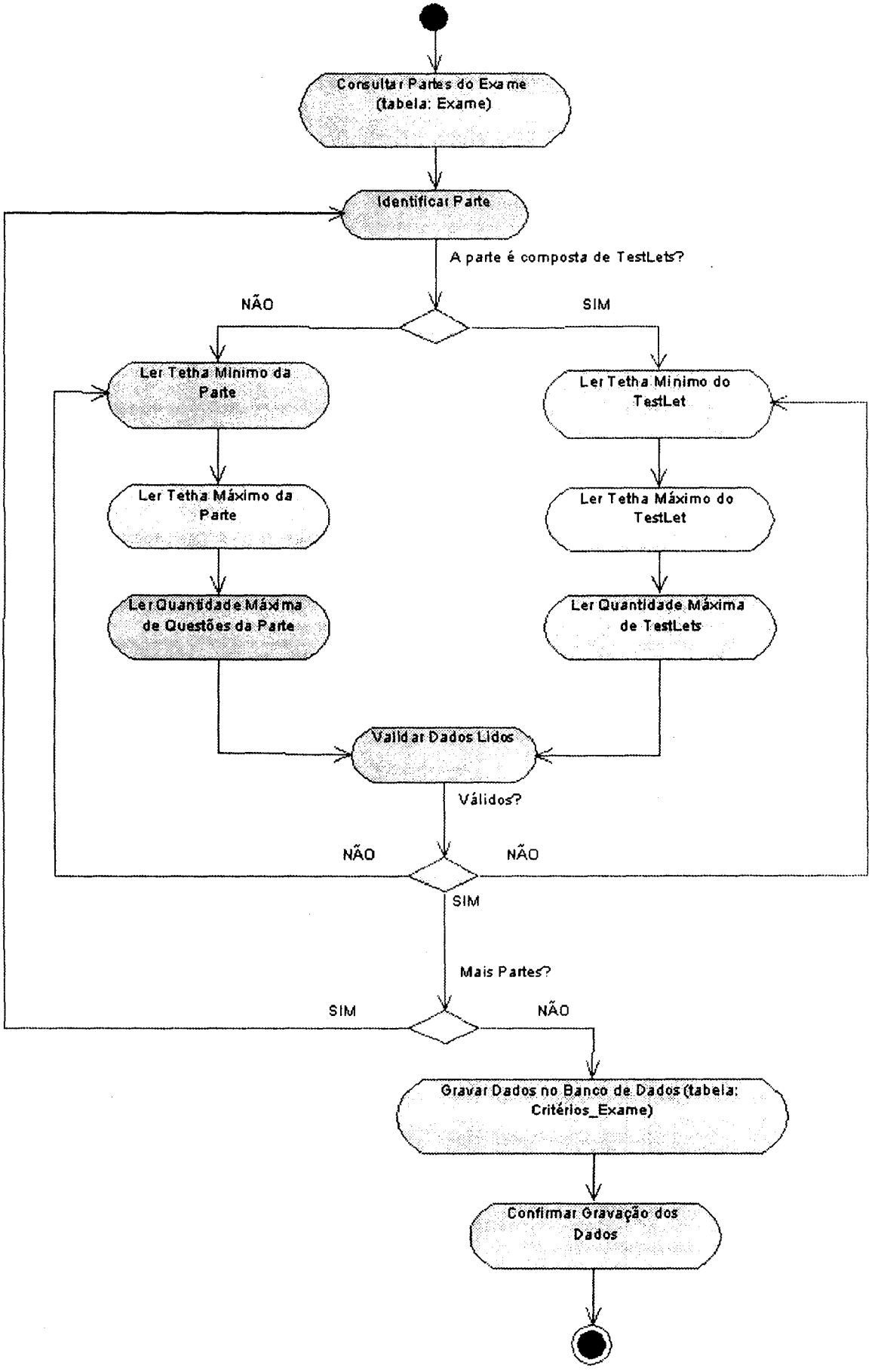

Figura C.12: Diagrama de Atividades – Ler Matriz de Critérios de Parada de Exame

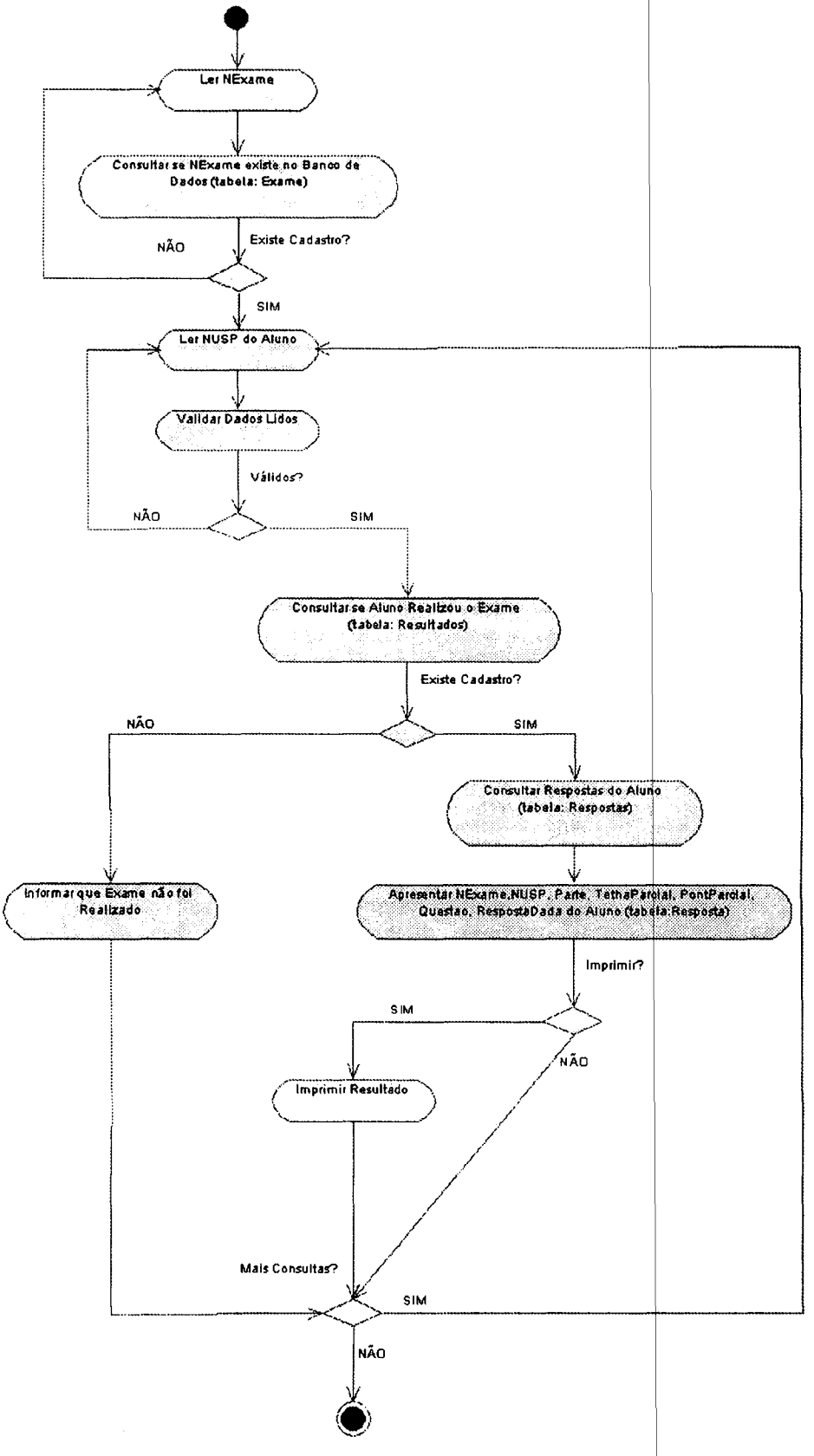

Figura C.13: Diagrama de Atividades – Consultar Resultado Detalhado.

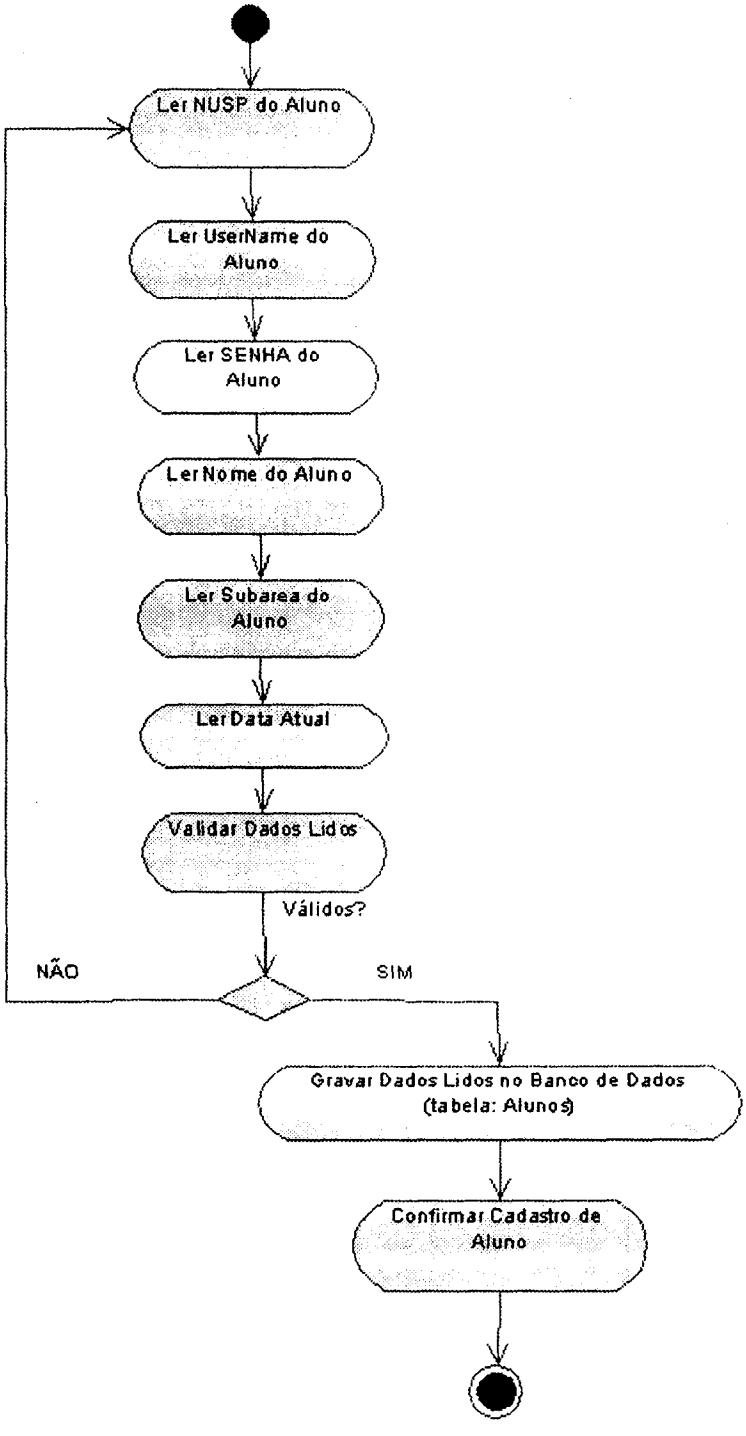

Figura C.14: Diagrama de Atividades – Cadastrar Aluno.

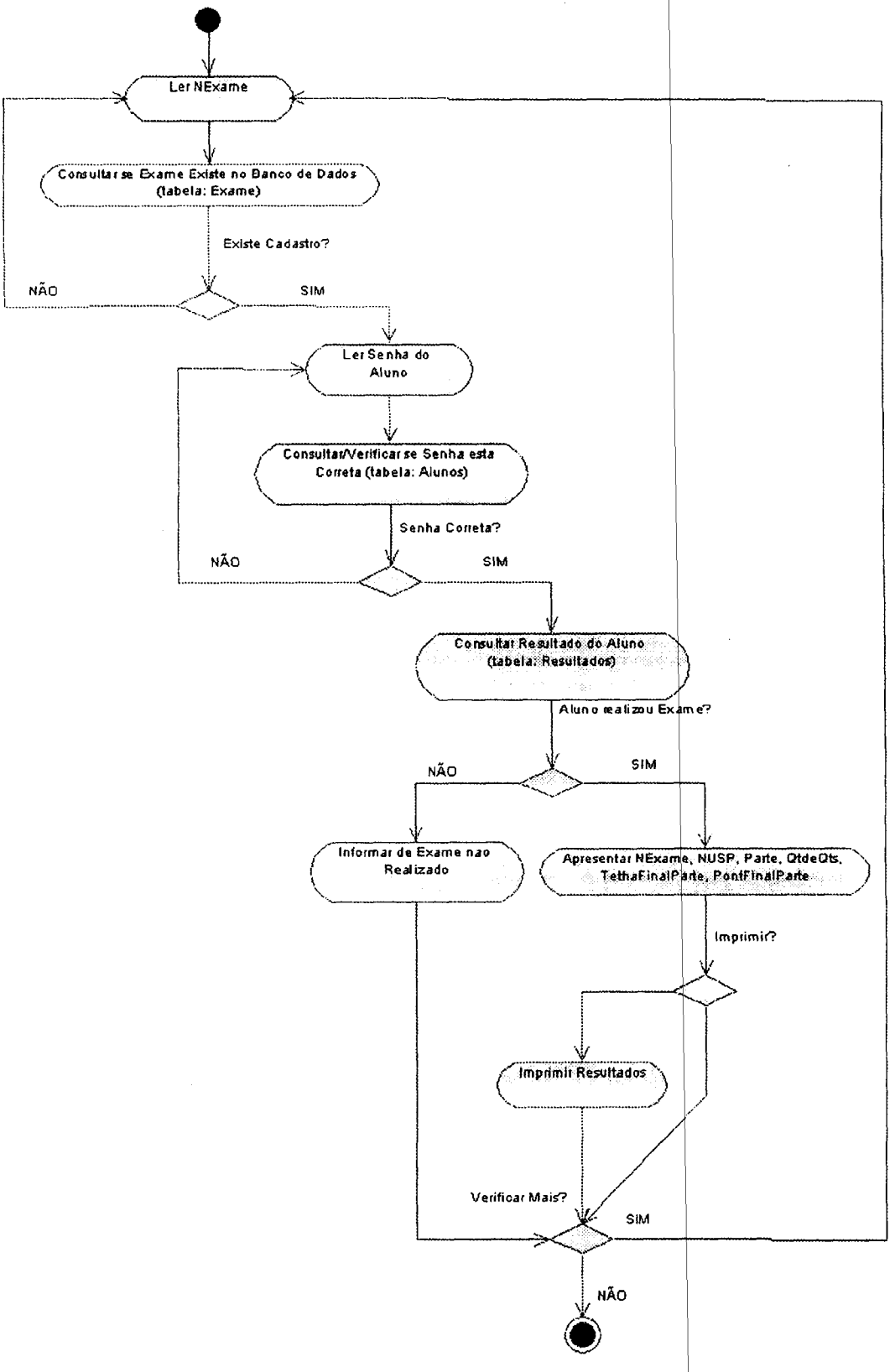

Figura C.15: Diagrama de Atividades – Consultar Resultado.

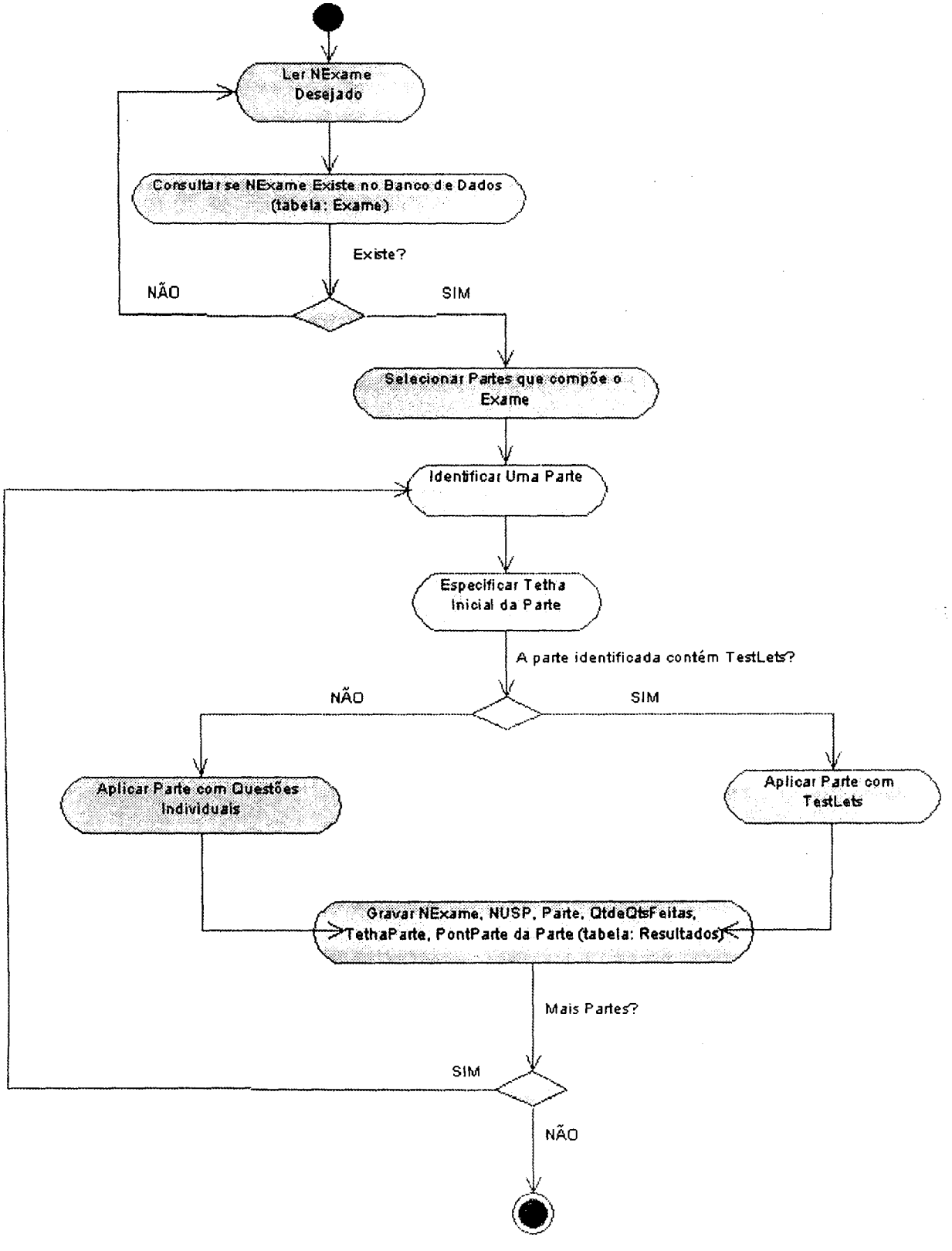

Figura C.16: Diagrama de Atividades – Realizar Exame.

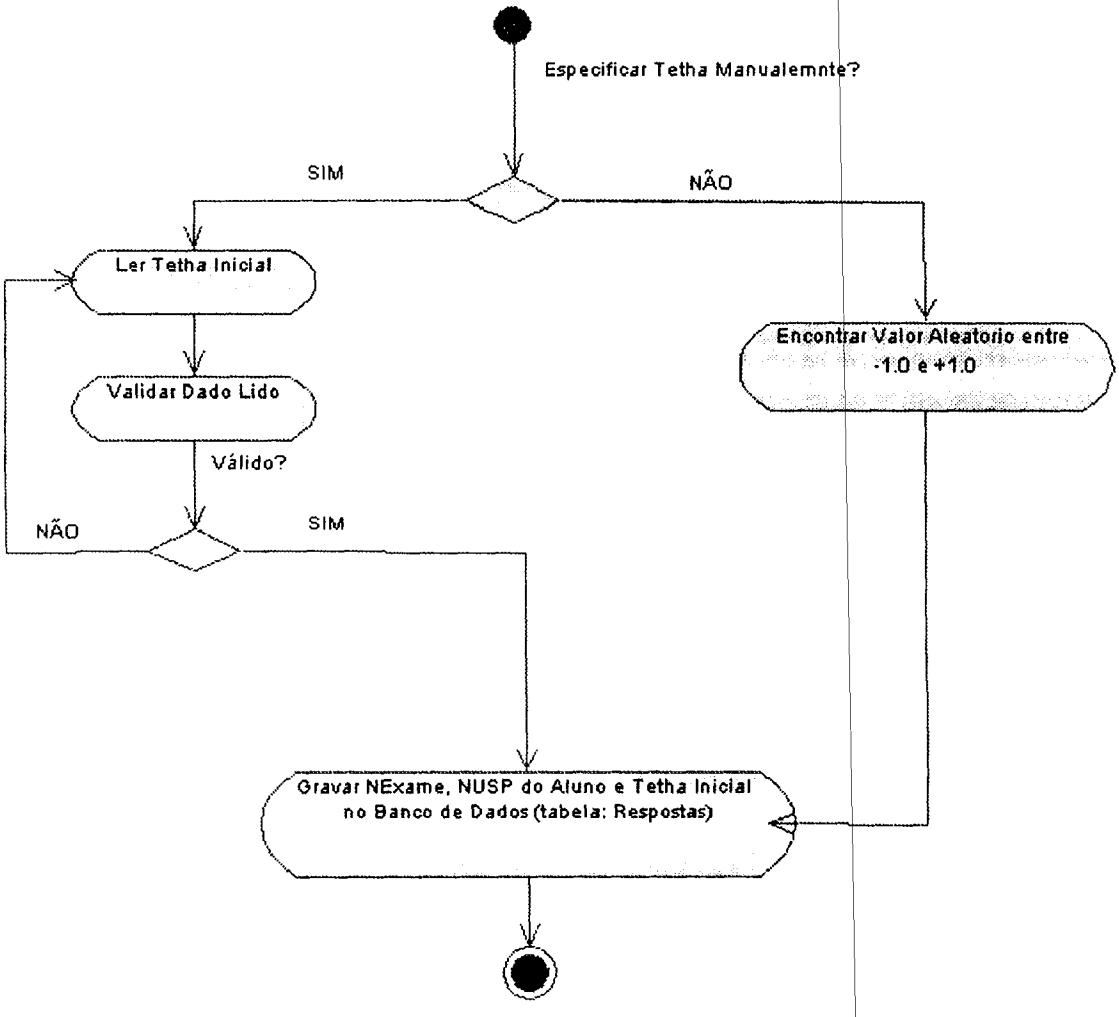

Figura C.17: Diagrama de Atividades – Especificar Tetha Inicial.

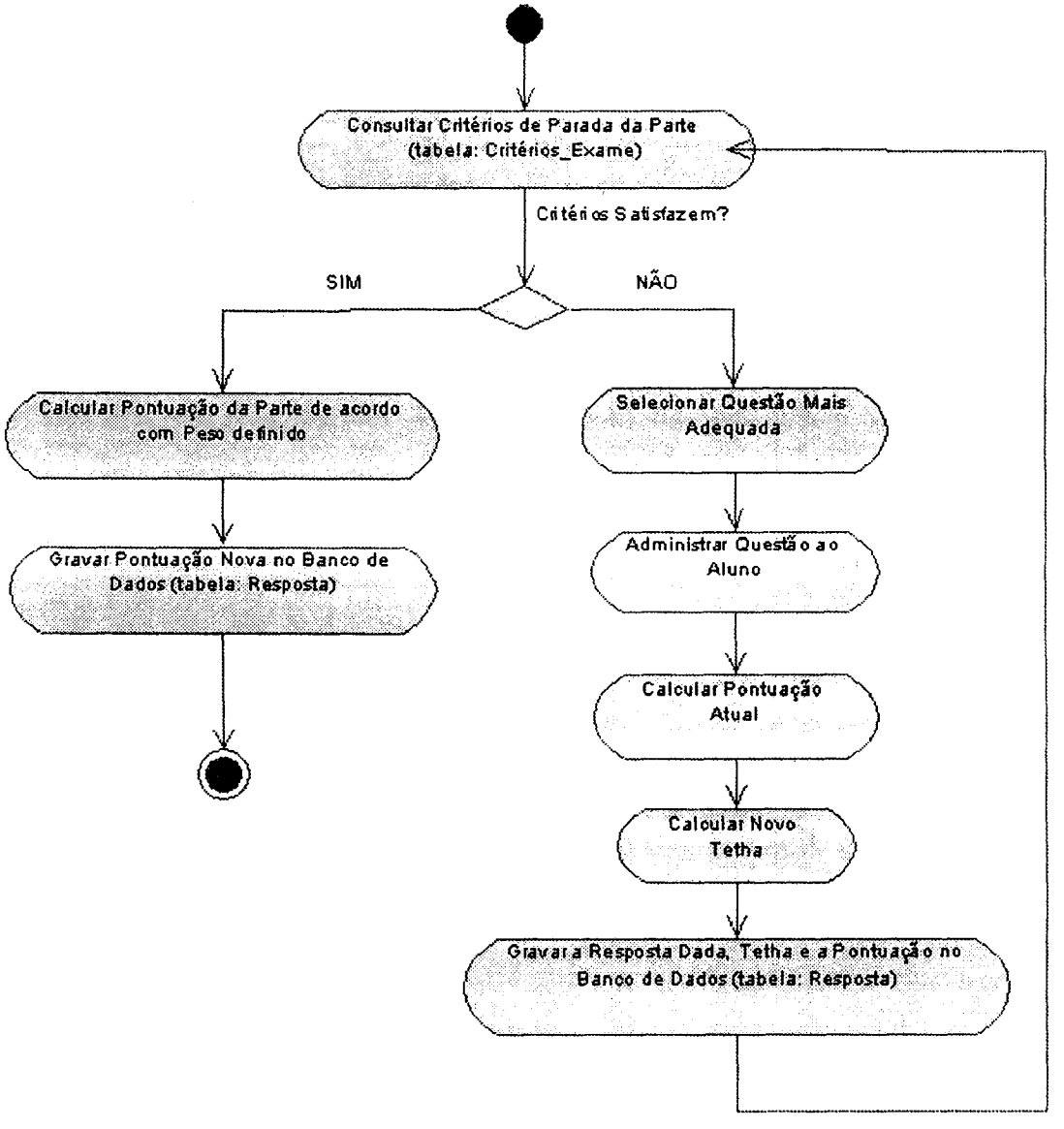

Figura C.18: Diagrama de Atividades – Aplicar Parte com Questões Individuais.

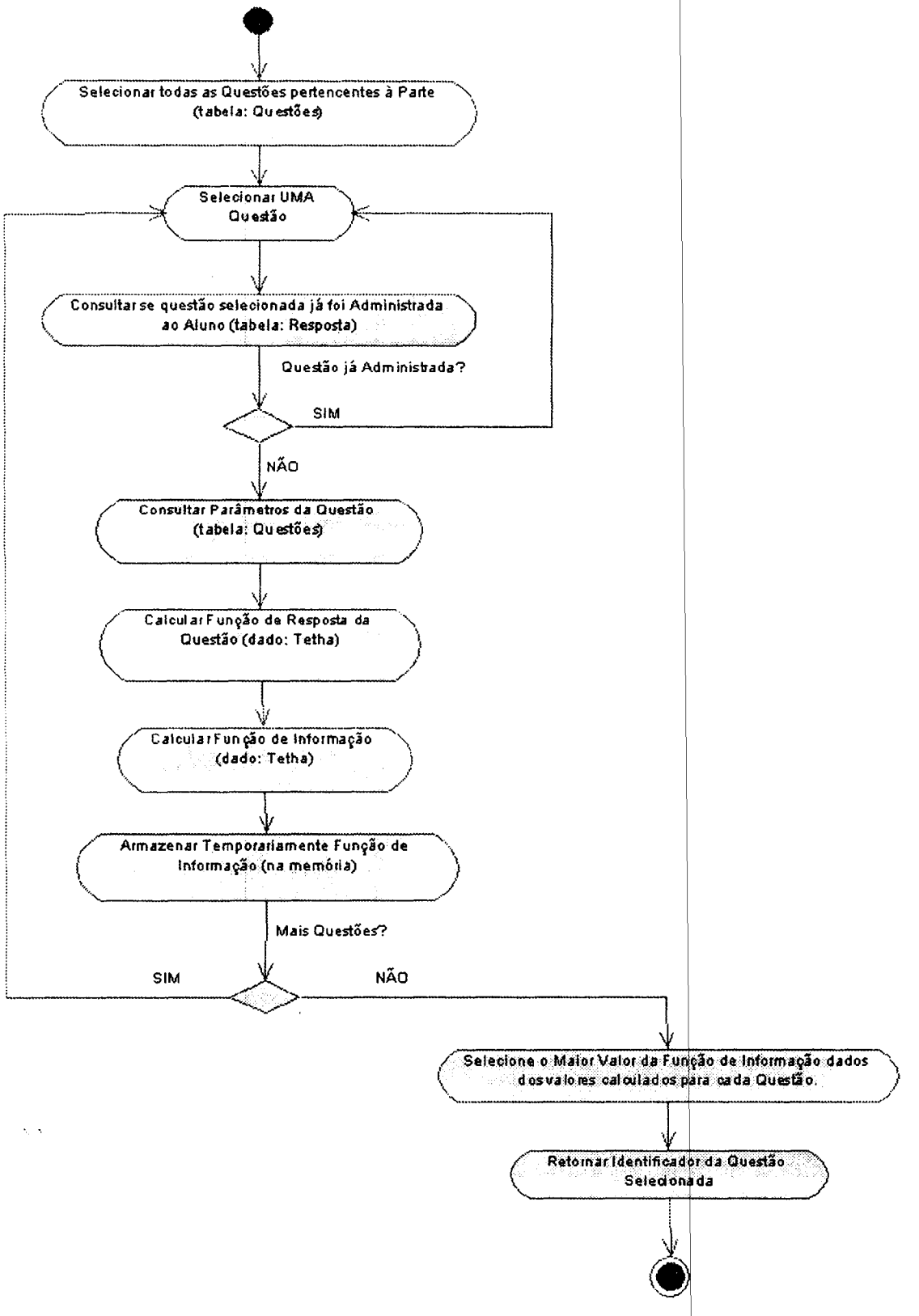

Figura C.19: Diagrama de Atividades – Selecionar Questão mais Adequada.

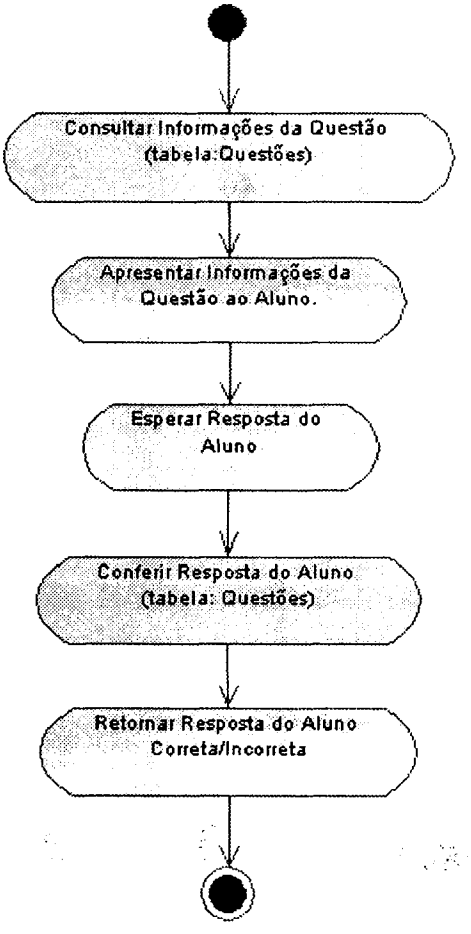

Figura C.20: Diagrama de Atividades – Administrar Questão.

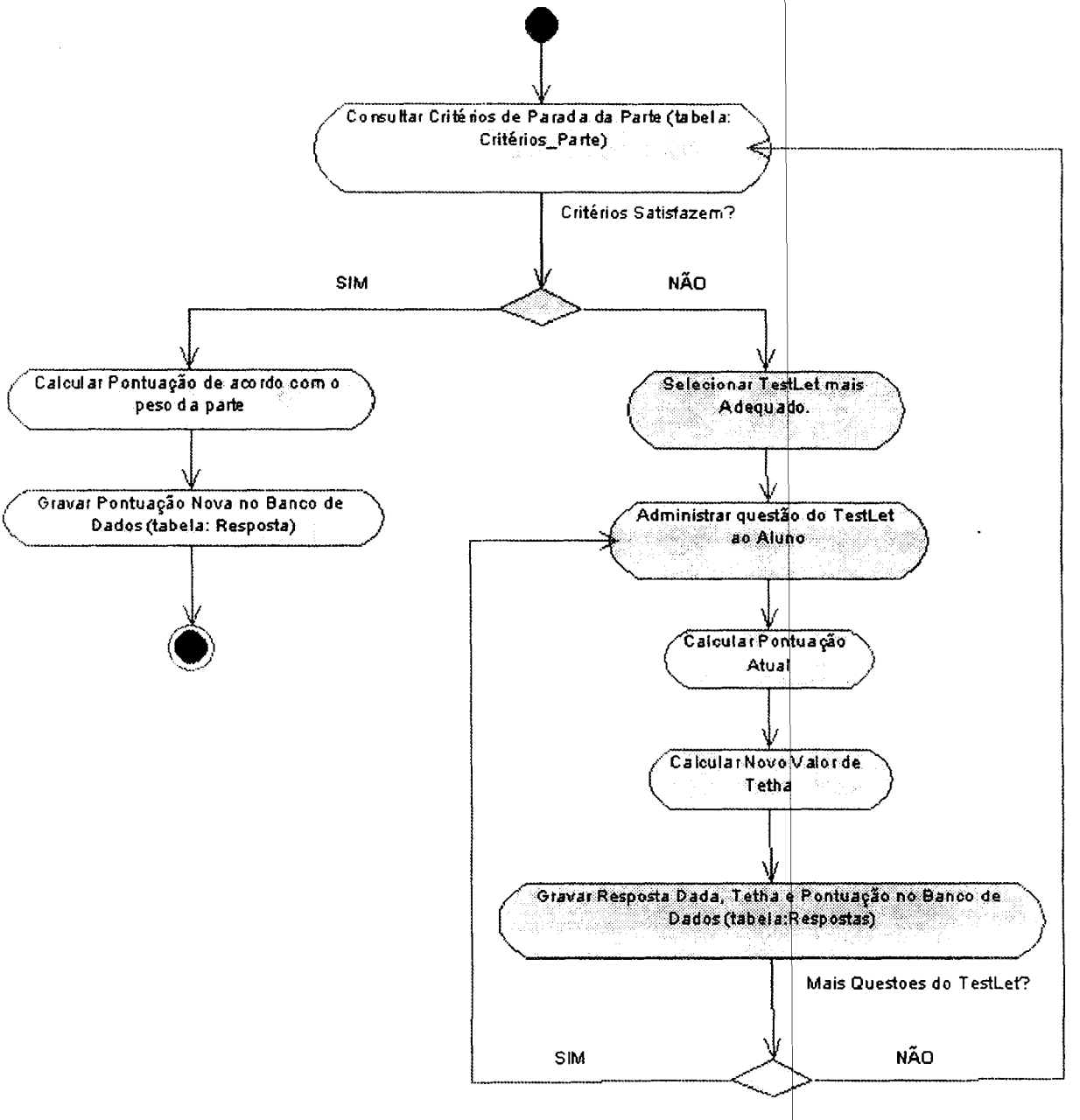

Figura C.21: Diagrama de Atividades – Aplicar Parte com TestLets.

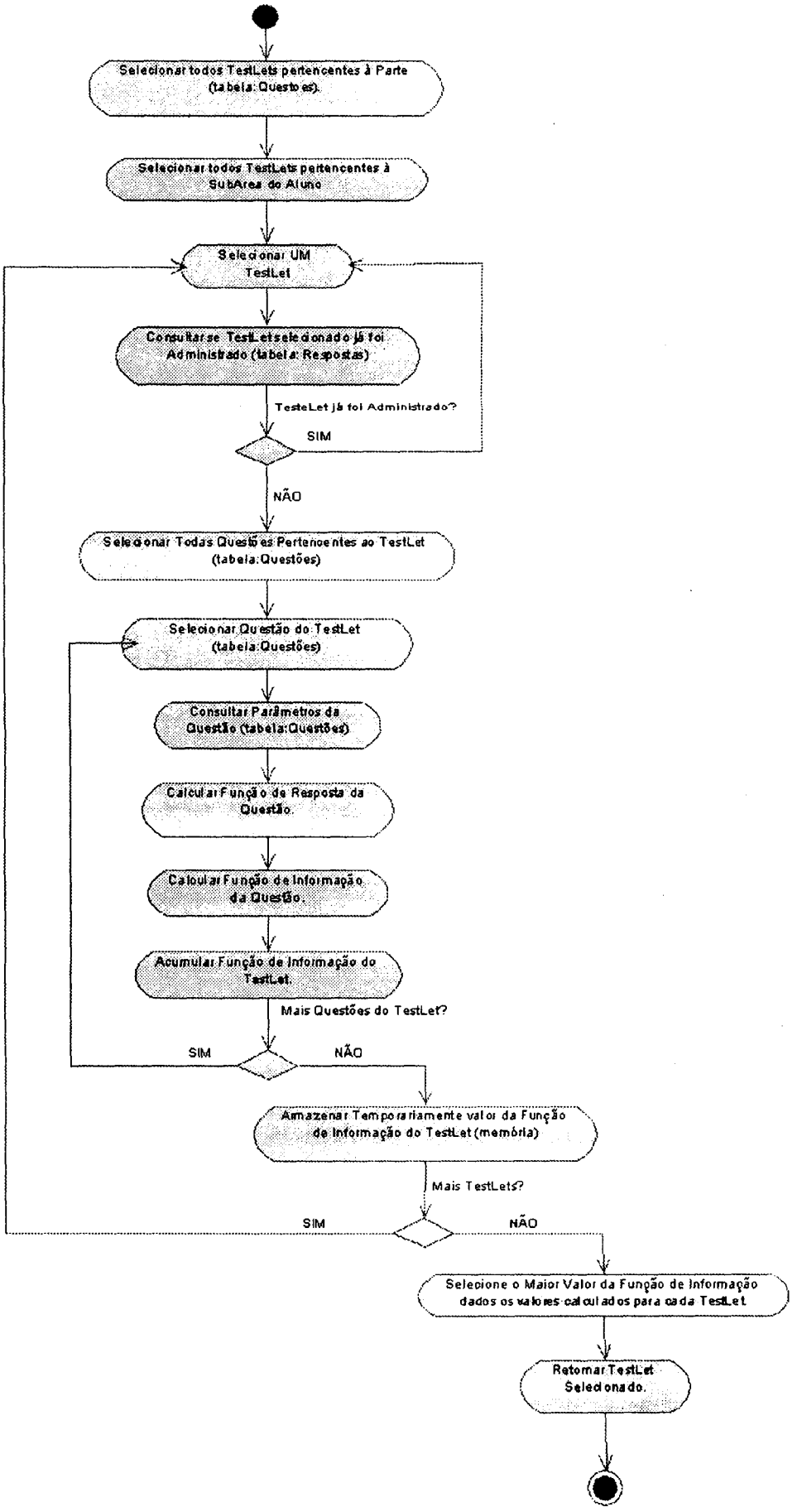

Figura C.22: Diagrama de Atividades – Selecionar TestLet mais Adequado

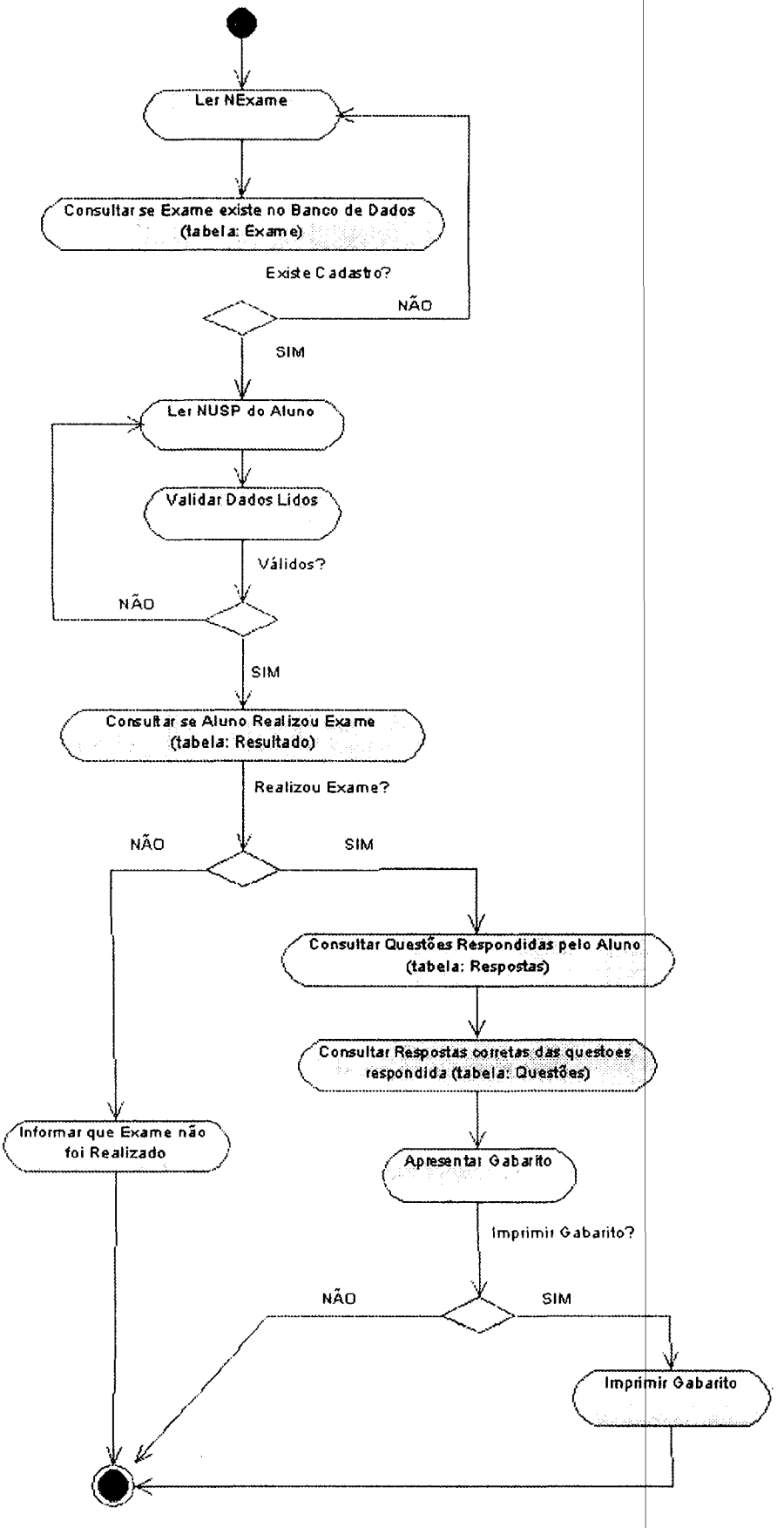

Figura C.23: Diagrama de Atividades – Consultar Gabarito.