IMPLEMENTAÇÃO DE MÉTODOS ESTATÍSTICOS PARA AVALIAÇÃO EDUCACIONAL NO SOFTWARE R

Conrad Elber Pinheiro

Dissertação apresentada ao
Instituto de Matemática e Estatística
da Universidade de São Paulo
para obtenção do grau de
Mestre em Estatística

Área de concentração: Estatística
Orientador: Prof. Dr. Dalton Francisco de Andrade

São Paulo, Fevereiro de 2006.

IMPLEMENTAÇÃO DE MÉTODOS ESTATÍSTICOS PARA AVALIAÇÃO EDUCACIONAL NO SOFTWARE R

Este exemplar corresponde à versão final devidamente corrigida e aprovada pela banca da dissertação de mestrado defendida por Conrad Elber Pinheiro.

São Paulo, Fevereiro de 2006.

Comissão julgadora:

Prof. Dr. Dalton Francisco de Andrade (orientador) – UFPA
Prof. Dr. Heliton Ribeiro Tavares – UFPA
Prof. Dr. Heraldo Vianna – FCC

À minha mãe Ivanise (em memória)
e meu pai Ronaldo

Agradecimentos

Agradeço a todos os professores de estatística do IME que me motivaram e me incentivaram a fazer o mestrado e, em especial, à Cláudia por ter sido orientadora de programa e ao Dalton por toda paciência que teve comigo ao longo dos trabalhos e pela sua disposição em estar sempre pronto a ensinar.

À banca examinadora composta pelos professores Heraldo e Héliton pelas críticas, elogios e sugestões dadas ao trabalho.

Aos meus colegas e amigos desde a época da graduação Andréia, Cíntia, Fernanda, Humberto, Iracema por todos incentivos que me deram.

À Rosangela, que compartilhou, não apenas o mesmo orientador e o mesmo assunto, mas também as alegrias, angustias, desesperos e descobertas ao longo dos estudos. Por toda a força que me deu ao longo de todos esses anos.

Ao Caio por toda paciência e disposição que teve em me ajudar sem a qual não teria sido possível a conclusão dos estudos.

Ao Gilvan por ter disponibilizado e compartilhado seus conhecimentos em informática me ajudando de maneira significativa na finalização das implementações.

Ao meu pai e minha mãe (em memória) que me deram todo apoio, força, incentivo e motivação ao longo de toda a graduação e de todo o mestrado, me permitindo chegar até aqui.

Enfim, obrigado a todas as pessoas que me ajudaram!

Resumo

Neste trabalho são apresentadas implementações computacionais, utilizando o software livre R, de métodos estatísticos para a análise de itens e a equalização de testes utilizando-se a Teoria Clássica e a Teoria da Resposta ao Item. Os métodos implementados são: análise clássica de itens, equalizações pelos métodos de Tucker, Levine, Equipercentil e Braun-Holland na Teoria Clássica, estimação dos parâmetros dos itens pelo método da máxima verossimilhança marginal e equalização a posteriori pelos métodos média-desvio e média-média na Teoria da Resposta ao Item.

Abstract

In this work we present computational implementations, based on the free software R, of statistical methods to item analyses and testing equating, via Classical and Item Response Theories. The following methods were implemented: classical item analysis, the Tucker, Levine, Equipercentile and Braun-Holland methods for equating in the Classical Theory and item parameter estimation via marginal maximum likelihood and posteriori equalization via mean-mean and mean-sigma methods in the Item Response Theory.

Índice

1. Introdução 10

2. Modelos de obtenção de dados 13
2.1. Propriedades
2.2. Principais modelos de obtenção de dados 15
2.2.1. Modelo de grupos aleatórios simples 16
2.2.2. Modelo de grupos aleatórios contrabalanceados 16
2.2.3. Modelo de grupos equivalentes 17
2.2.4. Modelos de grupos aleatórios com teste âncora 18
2.2.5. Modelo de grupos não equivalentes com teste âncora 19
2.3. Outros modelos 20

3. Teoria Clássica: Análise de Itens e Métodos de Equalização 21
3.1. Análise de itens 21
3.2. Elementos para a análise de itens 22
3.2.1. Dificuldade do item 22
3.2.2. Índice de discriminação 23
3.2.3. Coeficiente de correlação ponto bisserial 24
3.2.4. Coeficiente alfa de Cronbach 24
3.2.5. Erro padrão de medida 25
3.3. Equalização na TC 25
3.3.1. Equalização Linear 26
3.3.2. Equalização equipercentil 27
3.3.2.1. Procedimento gráfico 28
3.3.2.2. Procedimento analítico 31
3.3.2.3. Propriedades dos escores equalizados 34
3.3.3. Equalização Linear versus Eqüipercentil 35

4. Equalização para grupos não equivalentes com teste âncora através da TC 37
4.1. Equalização linear 37
4.1.1. Método de Tucker 38
4.1.2. Método do escore observado de Levine 41
4.2. Pesos nas populações sintéticas 46
4.3. Resumo dos métodos 46
4.4. Tucker X Levine 48
4.5. Equalização eqüipercentil 53
4.5.1. Método de estimação de freqüências 53
4.5.2. Exemplo 55
4.5.3. Método linerar de Braun-Holland 59

5. Teoria da Resposta ao Item: Análise de Itens, Estimação e Equalização .. 61
5.1 Modelo logístico com 3 parâmetros (ML3) 62
5.1.1. Interpretação e Curva Característica do Item (CCI) 63
5.1.2. Suposições do modelo 65
5.2. Análise de Itens 66
5.3. Estimação dos parâmetros dos itens (calibração) e das habilidades .. 67
5.3.1. Estimação por máxima verossimilhança conjunta 68
5.3.2. Estimação de máxima verossimilhança marginal 69
5.3.3. Estimação das habilidades quando os parâmetros dos itens são conhecidos 71
5.4. Métodos de Equalização 73
5.4.1. Equalização a *posteriori* 73
5.4.1.1. Método média-devio (*mean-sigma*) 75
5.4.1.2. Método média- média (*mean- mean*) 76
5.4.1.3. Comparação entre os dois métodos 76

6. Implementações utilizando o software R 78
6.1. Download dos softwares e implementações 78

6.2. Técnica de programação 79
6.3. Entradas e saídas dos programas 79
6.4. Carregamento dos programas 80
6.5. *EstatR.exe* 81
6.6. Conjunto de dados 81
6.7. Implementações desenvolvidas 82
6.7.1. Programas – Análise Clássica de Itens 82
6.7.2. Programa de equalização na TC – método de Tucker 85
6.7.3. Programa de equalização na TC – método de Levine 89
6.7.4. Programa de equalização na TC – método equipercentil 92
6.7.5. Programa de equalização na TC – método de Braun-Holland ... 94
6.7.6. Programa de estimação na TRI – estimação dos parâmetros dos itens 97
6.7.7. Programa de estimação na TRI – estimação das habilidades ... 101
6.7.8. Programa de equalização na TRI – 2 testes 102
6.7.9. Programa de equalização na TRI – escala pré-definida 106

7. Conclusões e sugestões 109

Apêndice – Manual de utilização dos programas 111
A.1. *EstatR.exe* 111
A.2. Configuração dos arquivos de entrada 112
A.2.1. Arquivos *entradax.txt* e *entraday.txt* 113
A.2.2. Arquivos *parametrosx.txt* e *parametrosy.txt* 116
A.2.3. Arquivos *habilidadesx.txt* e *habilidadesy.txt* 117
A.2.4. Arquivos *itenscomunsx.txt* e *itenscomunsy.txt* 118
A.3. Programa – Análise Clássica de Itens 118
A.4. Programa de equalização na TC – método de Tucker 122
A.5. Programa de equalização na TC – método de Levine 125
A.6. Programa de equalização na TC – método equipercentil 128
A.7. Programa de equalização na TC – método de Braun-Holland 130

1. Introdução

Em situações como admissões em colégios, avaliações e seleções de indivíduos, em que provas são aplicadas em diferentes pessoas e em momentos diferentes faz-se necessário a utilização de vários modelos de provas diferentes, por questões de segurança e de imparcialidade. Se as mesmas questões fossem utilizadas em cada uma das aplicações, elas se tornariam conhecidas e os indivíduos que fizessem a última prova teriam grande vantagem sobre aqueles que fizeram primeiro. Em outras situações, como programas de aperfeiçoamento, em que é preciso reavaliar os mesmos indivíduos, devemos assegurar que as provas aplicadas são diferentes para garantir que estamos medindo o nível de conhecimento atual do indivíduo e não a habilidade de ele relembrar respostas da prova feita anteriormente.

Entretanto, quando desejamos comparar os desempenhos dos grupos submetidos a testes aplicados, é necessário que eles sejam equivalentes em algum sentido. Por exemplo: suponhamos que alunos da 4ª série do Ensino Fundamental tenham realizado uma prova de Matemática e alunos da 8ª série também realizem uma outra prova da mesma disciplina. Podemos ter interesse em comparar os desempenhos desses alunos. Por isso, procedimentos estatísticos, chamados de métodos de equalização, tem sido desenvolvidos para solucionar esses problemas. Métodos de equalização são procedimentos empíricos para estabelecer uma relação entre dois testes aplicados entre dois grupos que realizaram tais testes. Quando a equalização é feita, podemos comparar o desenvolvimento de um mesmo indivíduo ao longo do tempo, de diferentes indivíduos (como por exemplo, o desempenho entre alunos de escolas diferentes) e comparar dados (itens) de testes diferentes com características diferentes. Basicamente, trabalhamos com dois tipos de grupos de indivíduos: os equivalentes, em que os alunos possuem habilidades e conhecimentos semelhantes e, nessa situação, pode-se aplicar dois testes e concluir qual deles é o mais fácil ou mais difícil; os não equivalentes, em que os alunos possuem

conhecimentos diferentes como, por exemplo, turma da manhã e turma da noite, ou ainda, alunos da 4ª e da 8ª séries. Neste caso, podemos realizar uma equalização que nos permita comparar os grupos que realizaram as provas.

Atualmente são conhecidos métodos de equalização utilizando-se a Teoria Clássica (TC), em que a comparação é feita a partir dos testes como um todo, e utilizando-se a Teoria da Resposta ao Item (TRI), em que os itens são os elementos principais. A equalização clássica é um processo que, por permitir o ajuste dos escores dos diferentes testes aplicados, tenta estabelecer uma equivalência entre os escores de dois testes. Ela ajusta as diferenças em dificuldade entre testes construídos para serem equivalentes tanto em dificuldade como em conteúdo. É importante deixar claro que a equalização ajusta os escores para diferenças em dificuldade, mas não para diferenças em conteúdo.

O principal objetivo deste trabalho é apresentar implementações de métodos estatísticos que permitam realizar a análise de itens e a equalização tanto na TC como na TRI. Embora já existam alguns softwares que façam tais procedimentos, são, em sua maioria, programas comerciais e, por isso, optou-se em trabalhar com o software livre **R**. Para isso, apresentaremos, inicialmente, conceitos básicos dos principais métodos de análise de itens e de equalização na Teoria Clássica e na Teoria da Resposta ao Item. Maiores detalhes desses e de outros métodos poderão ser encontrados em Senno (2006). Na realidade, este trabalho representa o início do desenvolvimento de um amplo ambiente de software para a implementação de vários métodos estatísticos para avaliação educacional.

Apresentaremos, no Capítulo 2, os principais modelos de obtenção de dados. No Capítulo 3, são mostradas a análise de itens na TC e os principais métodos de equalização. Escolhido um dos métodos apresentados no Capítulo 2 (grupos não equivalentes com teste âncora) desenvolvemos, no Capítulo 4, os métodos de equalização para este modelo na TC. No Capítulo 5, tratamos da análise de itens, calibração e equalização na TRI. Finalmente, no Capítulo 6, são descritos, com o

auxílio de um exemplo, os programas implementados. O manual de utilização desses programas podem ser vistos no Apêndice. No Capítulo 7 apresentamos a conclusão do trabalho.

2. Modelos de obtenção de dados

No processo de equalização, tanto na TC como na TRI, encontramos variáveis que não podem ser observadas diretamente nos indivíduos, os chamados traços (construtos) latentes, que são características do indivíduo, como o conhecimento, e são chamadas de **escores verdadeiros** na TC e **habilidades** ou **proficiências** na TRI. Assim, na TC, o escore observado de um indivíduo é composto pelo escore verdadeiro mais uma medida de erro. É assumido que se um indivíduo fosse submetido a vários testes, na média, o erro tenderia a zero. Como na prática o indivíduo não é submetido a re-testes, o escore verdadeiro não pode ser observado. Em testes de múltipla escolha, o **escore observado** corresponde ao número de itens corretos que um indivíduo respondeu corretamente.

2.1. Propriedades

Suponhamos que temos dois testes, um na forma $\mathbf{X}$ e outro na forma $\mathbf{Y}$ e que desejamos tornar comparáveis seus escores, ou seja, desejamos realizar uma equalização de testes. Para a realização dessa equalização, algumas propriedades têm sido propostas na literatura (Angoff, 1971; Lord, 1980; Petersen et al, 1989):

1) Mesma habilidade – os dois testes devem medir as mesmas características (traço latente, habilidade ou conhecimento).

2) Eqüidade – para todo grupo de indivíduos com habilidades idênticas, a distribuição de freqüência condicional dos escores no teste $\mathbf{Y}$, após a transformação, é a mesma que a distribuição de freqüência condicional dos escores no teste $\mathbf{X}$. Definimos:

x como sendo um determinado escore na forma $\mathbf{X}$;

y como sendo um determinado escore na forma $\mathbf{Y}$;

G como a função de distribuição acumulada dos escores na forma $\mathbf{Y}$;

eq_Y como sendo a função de equalização utilizada na conversão de escore na forma **X** para a escala da forma **Y**;
G^* como sendo a distribuição acumulada de eq_Y.
Segundo Lord (1980),

$$G^*[eq_Y(x)|\tau] = G(y|\tau), \text{ para todo } \tau,$$

onde τ representa uma habilidade de determinado indivíduo.
Isto significa que um indivíduo, que possui uma habilidade τ, com determinado escore terá média, desvio padrão e distribuição de escores observados idênticos para os escores convertidos na forma **Y** e escores na forma **X**. Porém, essa propriedade só é válida se os testes **X** e **Y** são idênticos, ou seja, se as duas formas forem estritamente **paralelas** (consideramos que dois testes são paralelos se, após a transformação para uma mesma escala, ou seja, após a equalização, suas médias, desvios-padrão e correlações são iguais). Entretanto, na prática, não conseguimos construir formas idênticas. Além disso, para Kolen & Brennan (1995), se formas idênticas de testes fossem construídas, não seria necessário fazer a equalização. Então, o uso da eqüidade proposta por Lord é impossível ou desnecessária. Kolen & Brennan (1995) sugere uma propriedade menos restritiva que utiliza a esperança:

$$E[eq_Y(X)|\ \tau] = E(Y|\tau) \text{ para todo } \tau.$$

Essa propriedade implica que os indivíduos receberão, na média, o mesmo escore na forma **Y** e na forma **X** equalizada.

3) Invariância populacional – a transformação é a mesma independente do grupo utilizado para conduzir a equalização.

4) Simetria – a transformação é inversível, ou seja, a função usada para transformar um escore na forma **X** para a forma **Y** deve ser a inversa da função usada para transformar um escore na forma **Y** para a forma **X**.

A invariância e a simetria são originárias do propósito da equalização: produzir uma equivalência entre os escores. Essas propriedades implicam que a conversão é única e que pode ser invertida. Portanto, a equalização não pode ser interpretada como um simples problema de regressão, pois, em geral, a regressão pelo método dos mínimos quadrados de x em y não coincide com a regressão de y em x. Por isso, os métodos de regressão não são bons candidatos para usarmos como procedimentos de equalização. Na prática, essas quatro propriedades nunca são plenamente satisfeitas. Por isso, em geral, aceita-se que para testes unidimensionais apenas as propriedades de invariância e simetria sejam satisfeitas para a realização da equalização.

2.2. Principais Modelos de obtenção de dados

Quando se aplicam testes, podemos utilizar vários modelos de aplicação que permitam fazer, posteriormente, a equalização. Em geral, a equalização só é possível quando temos um grupo em comum de indivíduos ou quando temos itens em comuns nas duas ou mais formas dos testes. Apresentamos cinco tipos básicos de maneiras de se obter dados, conforme veremos a seguir, que estão descritas em Petersen et al (1989), Kolen & Brennan (1995) e Angoff (1971). No contexto apresentado a seguir, chamaremos de **grupo** uma amostra de indivíduos obtida de certa população. **Subgrupo** é uma partição de um grupo. Os subgrupos realizam os testes simultaneamente, enquanto que os grupos podem realizar os testes em momentos distintos.

2.2.1. Modelo de grupos aleatórios simples

É o modelo mais simples a ser descrito. Nele, duas formas de testes (provas) a serem equalizadas, **X** e **Y**, são administradas a um mesmo conjunto de indivíduos, ou seja, a um mesmo grupo, num mesmo dia, uma após a outra (Figura 2.1). A vantagem deste método é que diferenças nos níveis de dificuldade não são confundidas com diferenças em habilidade, pois são os mesmos indivíduos que estão fazendo os testes.

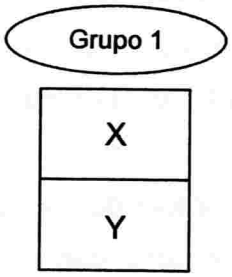


Figura 2.1: Modelo de grupos aleatórios simples

Entretanto, deve-se assumir que os escores obtidos na segunda prova não são afetados pela primeira prova feita. Ou seja, assumimos que fatores como aprendizado, fadiga e prática, que chamamos de **efeitos de ordem**, não afetam os escores da segunda forma aplicada. Como esses efeitos de ordem geralmente estão presentes e não havendo motivos para desprezá-los, este modelo raramente é utilizado na prática.

2.2.2. Modelo de grupos aleatórios contrabalanceados

Este modelo tenta corrigir os problemas de efeitos de ordem que ocorrem na situação descrita anteriormente. Aqui, um grupo é dividido em dois subgrupos aleatórios de modo que um subgrupo pegue um teste que contém primeiro a forma **X** seguida da forma **Y** e o outro subgrupo faça o teste que tem a forma **Y** impressa antes da **X** (Figura 2.2).

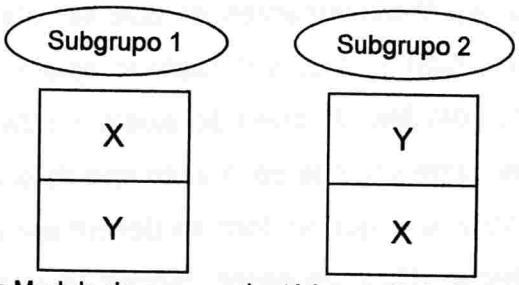


Figura 2.2: Modelo de grupos aleatórios contrabalanceados

As provas são distribuídas num processo espiral, ou seja, o primeiro indivíduo pega a prova com a forma **X** impressa primeiro; o segundo indivíduo pega a prova com a forma **Y** impressa em primeiro e assim por diante. O resultado desse processo é aleatório, desde que um número suficiente de indivíduos realize as provas e que eles não estejam sentados em uma seqüência alternada como, por exemplo, por sexo (homem – mulher – homem – ...). Além disso, é importante assegurar que efeitos de ordem como a fadiga não estejam presentes.

2.2.3. Modelo de grupos equivalentes

Na prática, é muito difícil haver tempo suficiente para que todos os indivíduos façam mais de uma forma de teste, como ocorre no modelo de grupos aleatórios contrabalanceados. Uma alternativa é o modelo de grupos equivalentes, que não requer que todos os indivíduos façam todas as formas de testes. Aqui, cada forma é administrada a um grupo aleatório, conforme Figura 2.3.

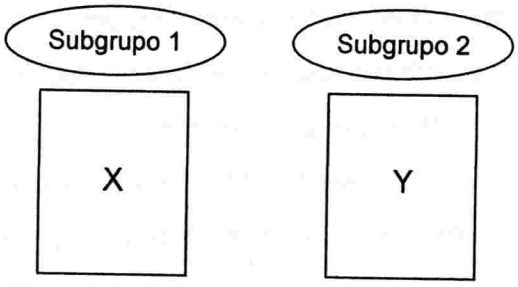


Figura 2.3: Modelo de grupos equivalentes

Se o tempo máximo para a realização de cada forma for o mesmo, então pode-se utilizar o método espiral para a criação dos subgrupos. Neste modelo, é preciso que os subgrupos sejam semelhantes no que se refere às habilidades que estão sendo mensuradas. Não é possível realizar ajustes para diferenças aleatórias entre os subgrupos pelo fato de que não existem indivíduos nem itens em comum entre os subgrupos. Uma solução para esse tipo de problema é o uso de amostras grandes. Outro problema é que as formas devem ser aplicadas ao mesmo tempo, o que pode dificultar a utilização deste método. A vantagem deste método é que não ocorrem efeitos de ordem como a fadiga e aprendizado.

2.2.4. Modelos de grupos aleatórios com teste âncora

Este modelo permite o ajuste para diferenças aleatórias entre os grupos. Aqui, uma forma **X** é administrada a um subgrupo, outra forma **Y** é administrada simultaneamente a outro subgrupo e uma terceira forma de teste em comum **V**, chamada de **teste âncora**, é administrada a ambos grupos conforme Figura 2.4.

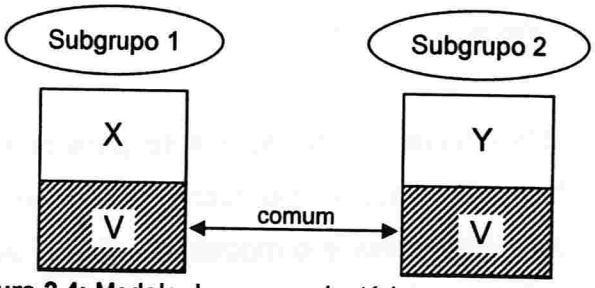


Figura 2.4: Modelo de grupos aleatórios com teste âncora

A forma **V** ou teste âncora, deve ser administrada na mesma ordem em ambos subgrupos, de modo que efeitos como aprendizagem e fadiga sejam afetados da mesma maneira. Um teste âncora pode ser interno ou externo. Ele é dito **teste âncora interno** quando os itens em comum fazem parte dos dois testes a serem equalizados. Neste caso, os escores obtidos a partir desses itens em comum são usados para computar o escore total dos testes. Chamamos de **teste âncora**

externo quando dois testes são aplicados em momentos distintos. Cada indivíduo realiza o teste âncora separadamente do teste **X** ou do teste **Y** e os escores obtidos nos testes âncoras não são computados no escore total.

2.2.5. Modelo de grupos não equivalentes com teste âncora

Este modelo é freqüentemente utilizado quando mais de uma forma por teste não pode ser administrada por questões de segurança ou por outros interesses práticos. Este modelo é idêntico ao anterior, com a diferença que as formas dos testes são aplicadas a grupos de indivíduos diferentes e, portanto, os testes podem ser aplicados em momentos distintos. Logo, os grupos não são equivalentes e, por isso, é esperado encontrar diferenças nas médias entre os dois grupos que pode ser tanto devido às diferenças nos grupos como devido a diferenças nas formas dos testes.

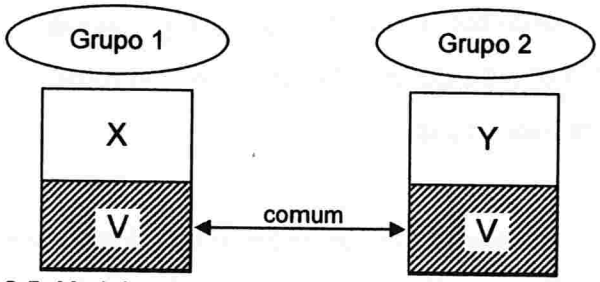


Figura 2.5: Modelo de grupos não equivalentes com teste âncora

Neste modelo é importante que o teste âncora (interno ou externo) seja similar em conteúdo e dificuldade aos testes que serão equalizados (**X** e **Y**). E, ainda, cada item deve ocupar a mesma posição (número do item) nas duas formas além de serem exatamente os mesmos (não deve haver, por exemplo, trocas de palavras e re-arranjo das alternativas).

Este modelo é largamente utilizado pelo fato de necessitar que apenas uma forma de teste seja administrada em cada momento da aplicação, ao contrário dos modelos anteriormente vistos, que necessitam mais de uma forma por aplicação

ou de grande amostras. A desvantagem deste método é que ele possui custos relativamente elevados.

Aqui, a única ligação entre os dois grupos são os itens comuns, que acabam por serem utilizados para detectar as diferenças entre os grupos. Segundo Kolen & Brennan (1995), embora haja uma variedade de métodos de equalização para este modelo, nenhum método tem sido encontrado de modo que dê ajustes estatísticos adequados quando os grupos são muito diferentes. Segundo Petersen et al (1989), os ajustes obtidos baseados em modelos que utilizam testes âncoras, que são similares aos testes que devem ser equalizados, são muito mais satisfatórios que aqueles baseados em testes não-paralelos.

2.3. Outros modelos

Petersen et al (1989) apresenta, ainda, um modelo de pré-equalização com uma e com duas seções variáveis e um modelo de pré-equalização do item, em que podemos equalizar escores de dois testes que não foram inteiramente aplicados a nenhum conjunto de indivíduos.

Dentre os métodos apresentados, os mais utilizados na prática são o modelo de grupos equivalentes e o modelo de grupos não equivalentes com teste âncora. Neste trabalho, trataremos apenas deste último pelo fato de ser um método bastante utilizado no dia-a-dia com relação à comparação de testes aplicados a grupos que se diferenciam em habilidade.

3. Teoria Clássica: Análise de Itens e Métodos de Equalização

Tradicionalmente, no ensino, costuma-se utilizar para expressar o resultado de uma prova apenas os escores brutos. Tais resultados dependem do conjunto de itens utilizados no teste e as interpretações realizadas dependem da prova como um todo. Dessa forma, só podemos comparar, utilizando-se métodos de equalização, indivíduos que foram submetidos a mesma prova ou a provas que possuem itens em comum. Para realizarmos procedimentos de equalização de testes, faz-se necessário uma análise prévia de itens, visto que a equalização com itens "ruins" fica comprometida, ou seja, a utilização de itens que tenham problemas, por exemplo, quanto à discriminação, conforme será discutido mais adiante. É possível, através da análise clássica de itens, identificar questões fáceis e difíceis, identificar possíveis problemas técnicos na elaboração do item que podem não permitir discriminar os indivíduos com bom rendimento daqueles com baixo rendimento, além de problemas de ensino e de aprendizagem no campo educacional. A partir desse tipo de análise é possível decidir se um item pode ser ou não empregado no futuro em um novo instrumento de medida.

3.1. Análise de itens

Vianna (1982) diz que as propostas da análise dos itens são basicamente três:
- fornecer informações sobre o desempenho dos examinados que permitam identificar possíveis problemas de aprendizagem;
- fornecer dados quantitativos que permitam identificar deficiências técnicas que comprometam o item como unidade de informação;
- desenvolver no professor (ou pesquisador) a capacidade de elaborar bons itens, exercitada através da reestruturação de itens cujas análises revelaram carência de requisitos básicos.

Ainda segundo o mesmo autor, a análise quantitativa dos itens visa:

- determinar itens considerados muito fáceis mesmo para alunos com baixo rendimento;
- determinar itens muito difíceis mesmo para os melhores examinados;
- encontrar itens que não discriminam os alunos com melhor aproveitamento dos que possuem baixo aproveitamento.

3.2. Elementos para a análise de itens

A seguir, apresentaremos algumas medidas que podem ser utilizadas para a análise clássica de itens de um teste.

3.2.1. Dificuldade do item

Uma primeira medida para análise do item é a **dificuldade do item**. Garrett (1958) afirma que essa dificuldade pode ser determinada de vários modos: através do julgamento de pessoas competentes que classifiquem o item em ordem de dificuldade; pela rapidez com a qual o item pode ser resolvido; pelo número de examinados que conseguem acertar o item. Os dois primeiros modos são apenas um primeiro passo na montagem de um teste. Porém, a proporção de acertos do item por determinado grupo é o método-padrão para se determinar a dificuldade. Alguns autores, como Nunnally (1964) e Vianna (1982). chamam de *índice de facilidade* à proporção de acerto do item e *de índice de dificuldade* à proporção de erros desse item. Outros, como Garrett (1958) já chamam de *índice de dificuldade* à proporção de acertos do item. Ao longo deste trabalho, adotaremos a definição apresentada por Garrett (1958). De modo geral, considera-se que itens tem dificuldade moderada quando a proporção de acertos está compreendida entre 0,40 e 0,60. A sugestão apresentada por alguns autores é de que, na construção de testes, devemos optar por itens que tenham uma média de dificuldade igual a 0,50. Mais ainda, que 25% dos itens tenham dificuldade inferior a 0,25, 25%

superior a 0,75 e os 50% restantes tenham dificuldade intermediária (entre 0,25 e 0,75).

3.2.2. Índice de discriminação

Uma característica de um bom item é que ele deve discriminar os examinados de desempenho superior daqueles com desempenho deficiente. Para se determinar esse índice podemos calcular a diferença entre a proporção de acerto do grupo superior (GS) e a proporção de acertos do grupo inferior (GI). Entende-se por GS os 27% dos respondentes com os escores mais altos e por GI os 27% dos respondentes com os escores mais baixos. Segundo Vianna (1982) e Garrett (1958) o uso de 27%, e não outro valor, permite formar grupos extremos tão grandes quanto o possível e, ao mesmo tempo, tão diferentes quanto o possível. Índices de discriminação maiores ou iguais a 0,40 são considerados satisfatórios. Uma discriminação negativa para um item pode indicar uma ambigüidade de construção desse item, pois indivíduos mais capazes identificam implicações que não foram consideradas pelo construtor da questão e são levados a registrar uma resposta diferente da que é considerada correta, enquanto que indivíduos deficientes, após um exame superficial da questão, assinalam a resposta correta. Vianna (1982) estabelece a seguinte classificação:

Tabela 3.1: Escala classificatória de índices de discriminação de um item

Índice de discriminação	Classificação
0,40 ou maior	Bom.
0,30 – 0,39	Bom, mas sujeito a aprimoramento.
0,20 – 0,29	Item marginal, sujeito à reelaboração.
0,19 ou menor	Item deficiente, que deve ser rejeitado.

Notemos que a Tabela 3.1 apresenta apenas um critério geral que depende do tipo de item utilizado, ou seja, podemos ter, em algumas situações, que um índice de discriminação igual a 0,30 seja adequado no caso de estarmos lidando com um item difícil.

3.2.3. Coeficiente de correlação ponto bisserial

Este coeficiente corresponde ao coeficiente de correlação de Pearson entre o escore total e a resposta ao item (1 indicando acerto do item e 0 indicando erro) de cada um dos respondentes. Valores superiores a 0,30 são considerados satisfatórios. Valores negativos indicam possíveis problemas com o item, visto que indicaria que o item está sendo acertado por alunos com baixo rendimento, enquanto que os melhores alunos estão errando o item. Valores muito baixos também indicam que o item possui algum tipo de deficiência, pois indicaria que não há uma correlação satisfatória entre o escore total do teste e os acertos do item. Uma outra medida similar, obtida através de uma transformação pela normal é o coeficiente bisserial. Porém, Garrett (1958) argumenta que o uso do ponto bisserial é preferível em relação ao bisserial pelo fato de ser uma estatística mais fidedigna (ou seja, é um instrumento mais digno de confiança) e também pelo fato de que ao se atribuir notas do tipo 0 ou 1, a hipótese de normalidade na distribuição das respostas certo-errado pode não ser sustentada, tornando o uso do ponto bisserial mais apropriado.

3.2.4. Coeficiente alfa de Cronbach

É um valor entre 0 e 1. Quanto mais próximo de 1, melhor é o teste como um todo. Admite-se que os valores de alfa aceitáveis devem ser superiores a 0,70 (Lucero & Meza, 2002). Esse valor é calculado através da fórmula (Garrett, 1958):

$$\text{ALFA} = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} p_i(1-p_i)}{s^2}\right], \qquad (3.1)$$

onde

k representa o número de itens do teste;

p_i é a proporção de acerto do item i;

s^2 é a variância dos escores.

A partir dessa fórmula, podemos notar que o valor de ALFA aumenta quando as correlações entre os itens e o escore total são altas. Além disso, ALFA é afetado pela heterogeneidade dos indivíduos: quanto mais heterogêneo for o grupo que realizou o teste, maior será a confiabilidade do instrumento. Dessa forma, este coeficiente não é uma propriedade exclusiva do instrumento em si, mas sim, é uma característica do instrumento para um determinado grupo que o realiza em uma situação específica (Lucero & Meza, 2002).

3.2.5. Erro padrão de medida

O erro padrão de medida do teste é calculado através da fórmula (Garrett, 1958):

$$EPM = s.\sqrt{1-\alpha}, \qquad (3.2)$$

onde α é o coeficiente alfa de fidedignidade do teste e s é o desvio-padrão dos escores. Pode-se notar que quanto mais próximo de 1 for o valor do coeficiente de fidedignidade, menor o erro padrão de medida. Para α=1, tem-se EPM = 0.

3.3. Equalização na TC

Como já vimos, muitas vezes precisamos recorrer, por diversas razões, a várias formas e aplicações de um mesmo teste. Porém, como raramente duas formas de testes são exatamente equivalentes em dificuldade, torna-se necessário equalizar as formas, ou seja, converter o sistema de unidades (ou métrica) de uma forma para a métrica da outra forma, de modo que, após a conversão, elas sejam equivalentes. Após a equalização, podemos, por exemplo, medir o crescimento (através de um treinamento), mesclar e comparar dados originários de formas diferentes de teste com diferenças em dificuldades e comparar diretamente desempenhos de dois indivíduos que realizaram testes em formas diferentes.

Para a realização da equalização, duas restrições são necessárias: a primeira é que os instrumentos (testes) devem medir as mesmas características. Em segundo lugar, a conversão deve ser única, exceto pelo erro aleatório e pelo método de transformação escolhido. O resultado deve ser independente dos dados utilizados para a transformação e deve ser aplicável a todas as situações (Angoff, 1971). Quando duas formas de testes são equivalentes, é razoável assumir que as formas de distribuição dos escores serão as mesmas e a conversão de escores da forma **X** para a forma **Y** pode ser realizada mudando-se simplesmente a origem e a unidade de medida, ou seja, fazendo ajustes apenas nos dois primeiros momentos. Na TC, isso pode ser feito através de dois métodos de equalização: o linear e o equipercentil. Neste capítulo, daremos uma visão geral de como funcionam os dois métodos. No capítulo seguinte, trataremos mais detalhadamente esses métodos em um dos modelos de obtenção de dados apresentados no Capítulo 2.

3.3.1. Equalização Linear

A equalização linear permite que diferenças nas dificuldades entre os dois testes (**X** e **Y**) variem ao longo da escala de escores. Por exemplo, neste tipo de equalização, pode-se ter a forma **X** mais difícil que a forma **Y** para escores baixos e mais fácil para escores altos. Na equalização linear, uma transformação é escolhida de modo que os escores nos dois testes são considerados equalizados quando eles correspondem a um mesmo número de desvios-padrão acima ou abaixo da média de algum grupo de respondentes (Angoff, 1971). Definimos $\mu(X)$ e $\mu(Y)$ como sendo as médias dos escores obtidos nas formas **X** e **Y**, respectivamente; $\sigma(X)$ e $\sigma(Y)$, os desvios-padrão dos escores nas formas **X** e **Y**, respectivamente. A conversão linear é definida pela padronização dos escores (z-escores) nas duas formas:

$$\frac{x-\mu(X)}{\sigma(X)} = \frac{y-\mu(Y)}{\sigma(Y)} \qquad (3.3)$$

Resolvendo a equação para y, obtemos:

$$l_Y(x) = y = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right] \qquad (3.4)$$

Notemos que a equação (3.4) tem a forma $l_Y(x) = A.x + B$ com $A = \frac{\sigma(Y)}{\sigma(X)}$ e $B = \mu(Y) - A.\mu(X)$. Se (3.3) fosse resolvida para x, obteríamos o mesmo resultado ao fazer a transformação de escala. Em outras palavras, a relação obtida é simétrica, visto que a equação utilizada para converter escores da forma **X** na forma **Y** é a inversa da relação para se obter escores da forma **Y** transformados para a forma **X**. A equação (3.4) se assemelha a uma regressão linear simples, porém, na regressão, os termos $\frac{\sigma(Y)}{\sigma(X)}$ são multiplicados pela correlação entre X e Y. A regressão linear não pode ser considerada uma função de equalização pelo fato de a regressão de X em Y ser diferente da regressão de Y em X, a menos que o coeficiente de correlação seja 1, ou seja, a regressão não satisfaz à propriedade simétrica. Por isso, em geral, a regressão não deve ser usada como função de equalização (Kolen & Brennan, 1995 e Lord, 1980).

3.3.2. Equalização equipercentil

Na equalização equipercentil, uma curva é utilizada para descrever as diferenças em dificuldades, o que torna este método mais geral que a equalização linear. Assim, por exemplo, a forma **X** pode ser mais difícil que a forma **Y** para altos e baixos escores, porém mais fácil para escores intermediários.

Dizemos que a função equipercentil é uma função de equalização equipercentil se a distribuição de escores na forma **X** convertidas para a escala da forma **Y** é igual a distribuição populacional dos escores na forma **Y**. A função equipercentil é desenvolvida através da identificação de escores na forma **X** que tem o mesmo

percentil de escores na forma **Y**. Em resumo, na equalização equipercentil uma transformação é feita de modo que os escores de dois testes são equivalentes se eles têm um mesmo percentil (Angoff, 1971). É importante ressaltar que a equalização equipercentil não faz suposições a respeito dos testes que serão equalizados. Este modelo de equalização simplesmente transforma e ajusta as unidades de escore de um teste de modo que coincida com a distribuição de escores de outro teste.

A equalização equipercentil pode ser obtida através de dois passos: primeiro, devemos construir uma tabela ou gráfico da freqüência relativa acumulada para as duas formas que serão equalizadas. Depois, a partir das freqüências relativas de uma forma, obtemos os escores (equalizados) correspondentes à mesma freqüência na outra forma. Na prática, um problema muito comum ocorre quando se trabalha com variáveis discretas. Por exemplo, num teste, os escores obtidos podem assumir apenas valores inteiros. Neste caso, freqüentemente, não podemos encontrar um escore na forma **Y** que tenha o mesmo percentil que determinado escore na forma **X**. Por isso, devemos tornar a distribuição contínua. Kolen & Brennan (1995) apresenta um processo de continuidade utilizando a função de percentil. Dada uma variável aleatória inteira X e uma variável aleatória U que é uniformemente distribuída em $[-{}^{1}/_{2} ; {}^{1}/_{2}]$, definimos uma nova variável aleatória $X^{*}=X+U$. Essa nova variável é contínua e a função de distribuição acumulada corresponde a **função de *rank* percentil**. A inversa da distribuição acumulada desta nova variável existe e é a **função percentil**.

3.3.2.1. Procedimento gráfico

Para ilustramos o procedimento gráfico, vamos supor um teste hipotético composto de 10 itens aplicado nas formas **X** e **Y**, conforme tabela 3.2. Definimos:

x é um determinado escore obtido na forma **X**, ou **escore bruto** na forma **X**;
y é um determinado escore obtido na forma **Y**, ou **escore bruto** na forma **Y**;

f(x) a proporção de indivíduos que obteve um determinado escore x;
g(y) a proporção de indivíduos que obteve um determinado escore y;
F(x) a função de distribuição dos escores na forma **X**;
G(x) a função de distribuição dos escores na forma **Y**;
P(x) *rank* percentil na forma **X**;
Q(y) *rank* percentil na forma **Y**.

Os *ranks* percentis são calculados utilizando o conceito de continuidade apresentado anteriormente:

$$P(x) = 100.\ [F(x-1) + f(x) / 2] \quad e$$

$$Q(y) = 100.\ [G(y-1) + g(y) / 2].$$

Tabela 3.2: distribuição de escores das formas **X** e **Y** num teste hipotético

x	f(x)	F(x)	P(x) (%)	y	g(y)	G(y)	Q(y) (%)
0	0,00	0,00	0,0	0	0,00	0,00	0,0
1	0,01	0,01	0,5	1	0,02	0,02	1,0
2	0,03	0,04	2,5	2	0,05	0,07	4,5
3	0,04	0,08	6,0	3	0,09	0,16	11,5
4	0,05	0,13	10,5	4	0,12	0,28	22,0
5	0,10	0,23	18,0	5	0,20	0,48	38,0
6	0,19	0,42	32,5	6	0,25	0,73	60,5
7	0,24	0,66	54,0	7	0,20	0,93	83,0
8	0,18	0,84	75,0	8	0,04	0,97	95,0
9	0,12	0,96	90,0	9	0,02	0,99	98,0
10	0,04	1,00	98,0	10	0,01	1,00	99,5

Para ser consistente com as definições de *rank* percentil, a função *rank* percentil é construída de modo que cada ponto do gráfico representa a freqüência relativa acumulada no limite inferior de cada intervalo de escore inteiro, ou seja, cada ponto assume o valor i–0,5 para i=0,1,2,...,n onde n é o número de itens. Em seguida, os pontos dos gráfico são unidos através de segmentos de reta. Esse

método de construção de gráfico é chamado de **interpolação linear da distribuição de freqüências relativas acumuladas**. Os segmentos que unem os pontos dos gráficos não precisam ser lineares. Métodos de interpolação curvilinear, como *cubic splines*, podem ser utilizados. Uma explicação detalhada deste processo pode ser visto em Kolen & Brennan (1995) e Senno (2006).

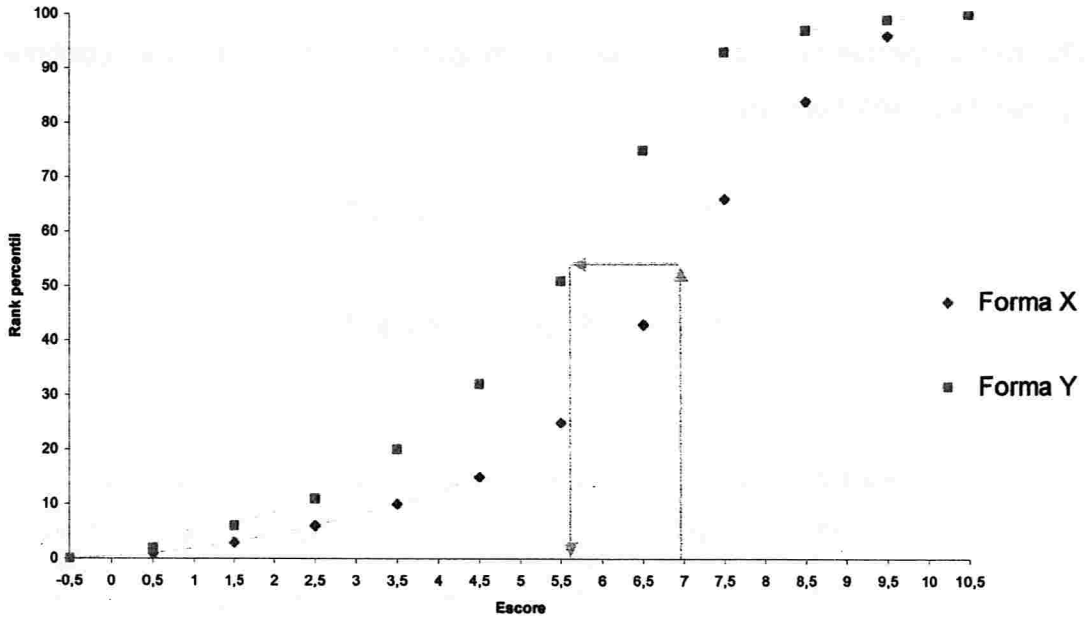


Figura 3.1: processo gráfico de equalização equipercentil

Na Figura 3.1, os *ranks* percentis de escores entre -0,5 e 0 são maiores que zero. Isso é resultado da definição que escores iguais a zero assumem distribuição uniforme no intervalo [-0,5 ; 0,5]. De maneira análoga, um escore igual a 10 assume distribuição uniforme em [9,5 ; 10,5] e, por isso, escores acima de 10 tem *rank* percentil menor que 100. Utilizando o procedimento gráfico, a partir de um determinado escore x, encontramos um escore y que possui o mesmo percentil (isso está indicado através das setas). Assim, por exemplo, se desejamos achar o escore equivalente a x=7, usando o método gráfico, obtemos aproximadamente y=5,7.

Uma desvantagem observada nesse tipo de procedimento é que ele é impreciso, pois não temos condições de determinar o valor exato do escore equivalente. Esse procedimento é bom para se ter uma idéia do escore equivalente, mas um valor mais exato pode ser obtido a partir de um método analítico, como veremos a seguir. Um problema particular às vezes ocorre nesse procedimento: pode acontecer de, em alguma forma do teste, nenhum indivíduo receber determinado escore. Nesse caso, o gráfico apresentado anteriormente será, para esses escores, horizontal e pode-se atribuir para esse percentil um valor arbitrário, embora geralmente se considere o valor correspondente ao ponto médio dos escores correspondentes a essa reta horizontal.

3.3.2.2. Procedimento analítico

O procedimento gráfico pode ser transformado em um método analítico, através do uso de fórmulas, que é mais preciso. Para isso, vamos definir K_X como sendo o número de itens na forma **X**. Seja X uma variável aleatória que representa os escores do teste na forma **X** e que pode assumir valores inteiros de 0 a K_X. Seja f(x) a função densidade definida por:

$$\begin{cases} f(x) \geq 0 \text{ para } x=0,1,...,K_X \\ f(x) = 0 \text{ caso contrário.} \end{cases}$$

Seja F(x) a função de distribuição acumulada definida por:

$$\begin{cases} 0 \leq F(x) \leq 1 \text{ para } x=0,1,...,K_X \\ F(x) = 0 \text{ para } x<0 \\ F(x)=1 \text{ para } x>K_X. \end{cases}$$

Consideremos os possíveis valores de x não inteiros. Seja x^* o inteiro mais próximo de x de modo que $x^* - 0,5 \leq x < x^* + 0,5$. A função *rank* percentil para a forma **X** é dada por:

$$P(x) = 100.\{F(x^* - 1) + [x - (x^* - 0,5)].[F(x^*) - F(x^* - 1)]\}, \text{ para } -0,5 \leq x < K_X + 0,5$$
$$= 0, \text{ para } x < -0,5$$
$$= 100, \text{ para } x \geq K_X + 0,5. \qquad (3.5)$$

Para ilustrar, consideremos, no exemplo hipotético apresentado na seção anterior, o cálculo do percentil para x=7. Usando a equação (3.5) temos:

$$P(7) = 100.\{F(6) + [7 - (7 - 0,5)].[F(7) - F(6)]\} =$$
$$= 100.\{0,42 + [0,5].[0,66 - 0,42]\} =$$
$$= 54\%,$$

valor que coincide com o apresentado na Tabela 3.2.

A função *rank* percentil inversa é representada por P^{-1} e pode ser obtida resolvendo-se a equação (3.5) para x:

$$x = P^{-1}(P(x)) = \frac{P(x)/100 - F(x^* - 1)}{[F(x^*) - F(x^* - 1)]} + (x^* - 0,5), \; 0 \leq P < 100 \qquad (3.6)$$
$$= K_X + 0,5 \; , \; P^* = 100,$$

onde x* é o menor escore que possui distribuição acumulada relativa maior que P.

Exemplo: suponhamos que precisemos calcular o escore correspondente a um percentil de 30% na forma **X**. Observando a Tabela 3.2 temos que F(5) = 0,23 e F(6) = 0,42 Neste caso, o menor escore que possui distribuição acumulada relativa maior que 30% é $x^* = 6$. Logo, utilizando a fórmula (3.6) temos:

$$x = P^{-1}(30) = \frac{30/100 - F(5)}{[F(5) - F(4)]} + (6 - 0,5)$$
$$= \frac{0,30 - 0,23}{0,42 - 0,23} + (5,5) = 5,9.$$

Na equalização equipercentil, nosso interesse é encontrar um escore na forma **Y** que tenha o mesmo *rank* percentil de um escore na forma **X**, ou seja, o equipercentil da forma **Y** equivalente a um escore x na forma **X** é dado por

$$e_Y(x) = y = Q^{-1}[P(x)],\ -0,5 \le x \le K_X + 0,5\,. \tag{3.7}$$

Ou seja, para encontrarmos o escore equivalente a x na forma **Y**, primeiro calculamos o *rank* percentil de x na forma **X** e, então, achamos o escore na forma **Y** que possui o mesmo *rank* percentil na forma **Y**. Logo:

$$\begin{aligned} e_Y(x) = Q^{-1}[P(x)] &= \frac{P(x)/100 - G(y^* - 1)}{G(y^*) - G(y^* - 1)} + (y^* - 0,5)\ ,\ 0 \le P(x) < 100 \\ &= K_Y + 0,5\ ,\ P(x)=100, \end{aligned} \tag{3.8}$$

onde K_Y é o número de itens na forma Y e y^* é o menor escore que possui distribuição acumulada relativa maior que P.

Exemplo: encontrar o escore equivalente na forma Y a um escore x=7. Primeiro, verificamos, a partir da Tabela 3.2, que P(7) = 54,0%. O menor escore y que possui distribuição acumulada maior que 54,0% é y^*=6. Logo, utilizando a equação (3.8) obtemos:

$$\begin{aligned} e_Y(x) = Q^{-1}[P(x)] &= \frac{54/100 - G(5)}{G(6) - G(5)} + (6 - 0,5) = \\ &= \frac{0,54 - 0,48}{0,73 - 0,48} + (5,5) = 5,74\ , \end{aligned}$$

valor que é muito próximo, porém mais exato, que aquele encontrado através do procedimento gráfico (5,7).

Observação: quando algum dos f(x) [ou g(x)] são iguais a zero, devemos obter o escore correspondente a certo percentil a partir de

$$x = \frac{P^{-1} + P_2^{-1}}{2}, \tag{3.9}$$

onde P^{-1} é obtido a partir da equação (3.6), e P_2^{-1} é calculado da forma a seguir:

$$x = P_2^{-1}(P(x)) = \frac{P(x)/100 - F(x^{**})}{[F(x^{**}+1) - F(x^{**})]} + (x^{**} + 0{,}5),\ 0 < P \leq 100 \qquad (3.10)$$
$$= -0{,}5\ ,\ P = 0,$$

onde x^{**} é o maior escore com distribuição acumulada relativa menor que P(x).

A equação (3.10) gera os mesmos resultados que (3.6) para valores de f(x) diferentes de zero.

3.3.2.3. Propriedades dos escores equalizados

Ao se fazer uma equalização, o ideal é que os escores equalizados na forma **X** tenham a mesma distribuição dos escores na forma **Y**. Quando os escores são contínuos, essa distribuição será a mesma (Kolen & Brennan, 1995). Porém, ao se fazer o processo de continuidade de uma distribuição discreta, algumas diferenças poderão aparecer e essas distribuições podem ser diferentes. A tabela seguinte apresenta os escores da forma **X** equalizados para a mesma escala da forma **Y**:

Tabela 3.3: Forma **Y** equivalente à forma **X** num teste hipotético

x	f(x)	P(x) (%)	G(y)	y*	$e_y(x)$
0	0,01	0,0	0,020	0	-0,50
1	0,02	2,0	0,060	1	0,50
2	0,03	4,5	0,110	1	1,20
3	0,04	8,0	0,200	2	1,90
4	0,05	12,5	0,320	3	2,67
5	0,1	20,0	0,510	4	3,50
6	0,18	34,0	0,750	5	4,61
7	0,23	54,5	0,930	6	5,65
8	0,18	75,0	0,970	7	6,50
9	0,12	90,0	0,990	7	7,33
10	0,04	98,0	1,000	9	9,00

A partir dessa tabela, calculamos os quatro primeiros momentos para a forma **X**, forma **Y** e para a equalização de **X** para a escala **Y**:

Tabela 3.4: Momentos da equalização de **X** e da forma **Y**

escore	μ	σ	assimetria	curtose
y	5,1400	1,9850	-0,4625	3,0684
x	6,5100	2,1000	-0,8174	3,5400
$e_Y(x)$	5,1378	1,9559	-0,5531	3,1353

Na Tabela 3.4, calculamos a assimetria através da fórmula

$$\text{ass} = \frac{E[X - \mu(X)]^3}{[\sigma(X)]^3},$$

onde um valor positivo indica uma distribuição com uma cauda assimétrica que se estende em direção a valores mais positivos. Um valor negativo indica uma distribuição com uma cauda assimétrica que se estende em direção a valores mais negativos. Caso a assimetria seja igual a zero, isso indica que a distribuição é normal. Já a curtose é calculada através de:

$$\text{curt} = \frac{E[X - \mu(X)]^4}{[\sigma(X)]^4},$$

em que um valor igual a 3 representa a normal (mesocúrtica); valores maiores que 3 indicam uma distribuição leptocúrtica; valores menores que 3 indicam uma platicúrtica. Para obter os momentos de $e_Y(x)$, utilizamos os escores equalizados para a forma **Y** e a freqüência de distribuição f(x). Teoricamente, os momentos de $e_Y(x)$ deveriam ser idênticos aos de y. Porém, como podemos notar na Tabela 3.4, ocorrem discrepâncias. Na prática, não é possível encontrarmos distribuições exatamente iguais sendo considerada uma condição ideal de caráter mais teórico que prático (Kolen & Brennan, 1995).

3.3.3. Equalização Linear versus Equipercentil

No processo de equalização, formas diferentes de um teste são construídas de modo a serem as mais semelhantes possíveis. Assim, é razoável assumirmos que, para uma dada população, a forma da distribuição dos escores brutos seja

linear pode ser utilizada. Quando existem diferenças muito significativas nas distribuições é preferível a equalização equipercentil. Uma discussão mais detalhada a respeito desses dois métodos pode ser encontrada em Petersen et al (1989). O equipercentil pode alterar a forma da distribuição, sendo mais flexível quanto a dificuldade do teste, ou seja, considera diferenças na dificuldade entre os testes ao longo da escala de escores. Já a linear, não altera a forma da distribuição dos escores. Por exemplo: se a forma $\mathbf{Y}$ é mais difícil ou discrimina mais os estudantes bem preparados que a forma $\mathbf{X}$, o uso do equipercentil comprime os escores dos estudantes mais preparados na forma $\mathbf{Y}$ de modo que eles recebam escores similares. O uso da equalização linear, entretanto, preserva essas diferenças (Kolen & Brennan, 1995).

Os métodos aqui apresentados podem ser utilizados tanto para o modelo de grupos equivalentes quanto para grupos não equivalentes. Para a realização das implementações optou-se por trabalhar apenas com grupos não equivalentes, que é a situação mais encontrada na prática. No capítulo seguinte vamos tratar da equalização na presença de teste âncora.

4. Equalização para grupos não equivalentes com teste âncora através da TC

Como já vimos no Capítulo 2, no modelo de obtenção de dados para grupos não equivalentes com teste âncora temos duas populações que fazem testes diferentes. Esses testes possuem itens em comum (vide Figura 2.5). Como também já foi apresentado, os testes âncora podem ser ditos internos ou externos. Realizaremos aqui um estudo detalhado deste modelo, que é o que mais se assemelha aos utilizados na TRI, possibilitando uma comparação entre as implementações.

4.1. Equalização linear

Suponhamos um teste, que chamaremos de "novo", na forma **X**, um teste aplicado em momento anterior na forma **Y** e um teste comum a ambos que chamaremos de **V**. Sejam X, Y e V variáveis aleatórias que representam os escores obtidos em cada uma das formas. A forma **V** pode ser interna ou externa. Se **V** for um teste interno, então X e Y incluem os escores de **V**. Se **V** for externo, X e Y não incluem os escores de **V**. Os métodos de equalização que apresentaremos distinguem-se pelos pressupostos básicos. Os modelos de equalização envolvem duas populações que devem ser combinadas através da equalização para definir uma única população. Nesse sentido, surge o conceito de **população sintética** ou **população ponderada**: pesos são atribuídos às populações, ou seja, atribuímos peso w_1 para a população 1 e peso w_2 para a população 2 de modo que $w_1 + w_2 = 1$ com $w_1, w_2 \geq 0$. Como já vimos, para escores observados, a equalização linear dos escores na forma **X** transformados para a escala da forma **Y** é dada por:

$$l_{Y_s}(x) = y = \frac{\sigma_s(Y)}{\sigma_s(X)}x + \left[\mu_s(Y) - \frac{\sigma_s(Y)}{\sigma_s(X)}\mu_s(X)\right] \qquad (4.1)$$

onde o índice **s** representa a população sintética.

Definimos:

$$X_s = w_1X_1 + w_2X_2 \text{ e}$$
$$Y_s = w_1Y_1 + w_2Y_2 \qquad (4.2)$$

onde

X_s, Y_s representam os escores da população sintética nas formas $\mathbf{X}$ e $\mathbf{Y}$;
X_1, Y_1 representam os escores da população 1 nas formas $\mathbf{X}$ e $\mathbf{Y}$;
X_2, Y_2 representam os escores da população 2 nas formas $\mathbf{X}$ e $\mathbf{Y}$.

Os parâmetros em (4.1) são dados por:

$$\mu_s(X) = w_1\mu_1(X) + w_2\mu_2(X); \qquad (4.3)$$

$$\mu_s(Y) = w_1\mu_1(Y) + w_2\mu_2(Y); \qquad (4.4)$$

$$\sigma_s^2(X) = w_1\sigma_1^2(X) + w_2\sigma_2^2(X) + w_1w_2\left[\mu_1(X) - \mu_2(X)\right]^2; \qquad (4.5)$$

$$\sigma_s^2(Y) = w_1\sigma_1^2(Y) + w_2\sigma_2^2(Y) + w_1w_2\left[\mu_1(Y) - \mu_2(Y)\right]^2, \qquad (4.6)$$

onde $\mu_i(K)$, i=1,2 com K=$\mathbf{X},\mathbf{Y}$ representa o escore observado médio do grupo i realizando o teste na forma K. $\sigma_i^2(K)$ é a variância do grupo i realizando o teste na forma K. Admitindo que a população 1 tenha realizado a forma $\mathbf{X}$ e a população 2 a forma $\mathbf{Y}$, então notamos que $\mu_2(X), \sigma_2^2(X), \mu_1(Y)$ e $\sigma_1^2(Y)$ não podem ser observados diretamente. Para solucionar esse problema de estimação, apresentaremos os métodos de Tucker e Levine. As deduções e demonstrações das equações e fórmulas que apresentaremos neste capítulo podem ser encontradas com detalhes em Kolen & Brennan (1995) e Senno (2006).

4.1.1. Método de Tucker

As suposições deste método são:

1) REGRESSÃO LINEAR: é assumido que a regressão de X sobre V é a mesma para os grupos 1 e 2. De maneira análoga, supõe-se que a regressão de Y sobre V é a mesma para esses grupos. Seja a regressão na forma $x = \alpha v + \beta$ com:

$$\alpha_1(X \mid V) = \frac{\sigma_1(X,V)}{\sigma_1^2(V)}; \tag{4.7}$$

$$\beta_1(X \mid V) = \mu_1(X) - \alpha_1(X \mid V)\mu_1(V); \tag{4.8}$$

$$\alpha_2(X \mid V) = \frac{\sigma_2(X,V)}{\sigma_2^2(V)}; \tag{4.9}$$

$$\beta_2(X \mid V) = \mu_2(X) - \alpha_2(X \mid V)\mu_2(V). \tag{4.10}$$

Pela suposição da regressão linear, temos que:

$$\alpha_1(X \mid V) = \alpha_2(X \mid V); \tag{4.11}$$

$$\beta_1(X \mid V) = \beta_2(X \mid V). \tag{4.12}$$

2) VARIÂNCIA CONDICIONAL: assume-se que as variâncias condicionais de X dado V e de Y dado V são as mesmas para as populações 1 e 2:

$$\sigma_1^2(X)\left[1-\rho_1^2(X,V)\right] = \sigma_2^2(X)\left[1-\rho_2^2(X,V)\right]; \tag{4.13}$$

$$\sigma_1^2(Y)\left[1-\rho_1^2(Y,V)\right] = \sigma_2^2(Y)\left[1-\rho_2^2(Y,V)\right],$$

onde ρ é a correlação, ou seja,

$$\rho_1(X,V) = \frac{\sigma_1(X,V)}{\sigma_1(X)\sigma_1(V)}, \tag{4.14}$$

onde $\sigma_1(X,V)$ é a covariância.

Como a regressão de X em V é linear, então:

$$\mu_2(X) = \mu_1(X) - \alpha_1(X \mid V)\left[\mu_1(V) - \mu_2(V)\right]. \tag{4.15}$$

De maneira análoga, obtemos:

$$\mu_2(Y) = \mu_2(Y) + \alpha_2(Y \mid V)\left[\mu_1(V) - \mu_2(V)\right]. \tag{4.16}$$

As variâncias são dadas por:

$$\sigma_2^2(X) = \sigma_1^2(X) - \alpha_1^2(X \mid V)\left[\sigma_1^2(V) + \sigma_2^2(V)\right] \tag{4.17}$$

$$\sigma_1^2(Y) = \sigma_2^2(Y) - \alpha_2^2(Y \mid V)\left[\sigma_1^2(V) + \sigma_2^2(V)\right]. \tag{4.18}$$

Lembrando que $\mu_1(X)$ é estimado diretamente e substituindo o resultado (4.15) em (4.3), obtemos:

$$\mu_s(X) = \mu_1(X)[w_1 + w_2] - w_2\alpha_1(X \mid V)[\mu_1(V) - \mu_2(V)].$$

Sendo $w_1 + w_2 = 1$ e chamando $\gamma_1 = \alpha_1(X \mid V)$ temos:

$$\mu_s(X) = \mu_1(X) - w_2\gamma_1[\mu_1(V) - \mu_2(V)]. \qquad (4.19)$$

Analogamente:

$$\mu_s(Y) = \mu_2(Y) + w_1\gamma_2[\mu_1(V) - \mu_2(V)], \qquad (4.20)$$

com $\gamma_2 = \alpha_2(Y \mid V)$.

A variância da população sintética é:

$$\sigma_s^2(X) = \sigma_1^2(X) - w_2\gamma_1^2\left[\sigma_1^2(V) - \sigma_2^2(V)\right] + w_1w_2\gamma_1^2\left[\mu_1(V) - \mu_2(V)\right]^2, \qquad (4.21)$$

com $\gamma_1 = \alpha_1(X \mid V) = \dfrac{\sigma_1(X,V)}{\sigma_1^2(V)}$.

Analogamente:

$$\sigma_s^2(Y) = \sigma_2^2(Y) + w_1\gamma_2^2\left[\sigma_1^2(V) - \sigma_2^2(V)\right] + w_1w_2\gamma_2^2\left[\mu_1(V) - \mu_2(V)\right]^2, \qquad (4.22)$$

com $\gamma_2 = \alpha_2(Y \mid V) = \dfrac{\sigma_2(Y,V)}{\sigma_2^2(V)}$.

Note que as equações obtidas independem se os itens âncoras são internos ou externos.

4.1.2. Método do escore observado de Levine

Aqui, temos um método de equalização dos escores observados que utiliza a equação (4.1) para transformar os escores de **X** em escores observados em **Y**. Porém, as suposições aqui feitas se referem aos **escores verdadeiros** (*true scores*) ou **escores ajustados** T_X, T_Y e T_V de modo que os escores observados X, Y e V podem ser escritos como:

$$X = T_X + E_X, \tag{4.23}$$

$$Y = T_Y + E_Y, \tag{4.24}$$

$$V = T_V + E_V, \tag{4.25}$$

com E_X, E_Y e E_V sendo os erros que têm média zero e não são correlacionados com os escores ajustados.

As suposições feitas para o método de Levine são:

1) CORRELAÇÃO: assume-se que **X**, **Y** e **V** medem a mesma coisa e que, considerando-se as populações 1 e 2:

$$\rho_1(T_X, T_V) = \rho_2(T_X, T_V) = 1 \tag{4.26}$$

$$\rho_1(T_Y, T_V) = \rho_2(T_Y, T_V) = 1 \tag{4.27}$$

2) REGRESSÃO LINEAR: assume-se que as regressões de T_X sobre T_V e T_Y sobre T_V são as mesmas nas duas populações, ou seja, o coeficiente angular é dado por:

$$\alpha_1(T_X \mid T_V) = \frac{\rho_1(T_X, T_V)\sigma_1(T_X)}{\sigma_1(T_V)}.$$

Como $\rho_1(T_X, T_V) = 1$,

$$\alpha_1(T_X \mid T_V) = \frac{\sigma_1(T_X)}{\sigma_1(T_V)}, \tag{4.28}$$

$$\alpha_2(T_X \mid T_V) = \frac{\sigma_2(T_X)}{\sigma_2(T_V)}. \tag{4.29}$$

Mas, pela suposição da regressão linear, temos:

$$\alpha_1(T_X \mid T_V) = \alpha_2(T_X \mid T_V), \tag{4.30}$$

$$\beta_1(T_X \mid T_V) = \beta_2(T_X \mid T_V), \tag{4.31}$$

obtendo

$$\mu_1(X) - \frac{\sigma_1(T_X)}{\sigma_1(T_V)}\mu_1(V) = \mu_2(X) - \frac{\sigma_2(T_X)}{\sigma_2(T_V)}\mu_2(V), \tag{4.32}$$

$$\mu_1(Y) - \frac{\sigma_1(T_Y)}{\sigma_1(T_V)}\mu_1(V) = \mu_2(Y) - \frac{\sigma_2(T_Y)}{\sigma_2(T_V)}\mu_2(V). \tag{4.33}$$

3) VARIÂNCIA DO ERRO: assume-se que a variância do erro para X é a mesma nas populações 1 e 2, assim como para Y e V. Como os escores verdadeiros e os erros não são correlacionados, temos que a variância do erro é igual à diferença entre a variância do escore observado e do escore verdadeiro, ou seja:

$$\sigma_1^2(X) - \sigma_1^2(T_X) = \sigma_2^2(X) - \sigma_2^2(T_X), \tag{4.34}$$

$$\sigma_1^2(Y) - \sigma_1^2(T_Y) = \sigma_2^2(Y) - \sigma_2^2(T_Y), \tag{4.35}$$

$$\sigma_1^2(V) - \sigma_1^2(T_V) = \sigma_2^2(V) - \sigma_2^2(T_V). \tag{4.36}$$

Lembrando que $\mu_2(X), \sigma_2^2(X), \mu_1(Y)$ e $\sigma_1^2(Y)$ não são observados, precisamos estimar esses valores para obtermos as estimativas para a população sintética. A partir de (4.32) temos:

$$\mu_2(X) = \mu_1(X) - \frac{\sigma_2(T_X)}{\sigma_2(T_V)}\left[\mu_1(V) - \mu_2(V)\right]. \tag{4.37}$$

Analogamente:

$$\mu_1(Y) = \mu_2(Y) - \frac{\sigma_2(T_Y)}{\sigma_2(T_V)}\left[\mu_1(V) - \mu_2(V)\right].$$

As variâncias são dadas por:

$$\sigma_2^2(X) = \sigma_1^2(X) - \frac{\sigma_1^2(T_X)}{\sigma_1^2(T_V)}\left[\sigma_1^2(V) - \sigma_2^2(V)\right]. \quad (4.38)$$

Analogamente:

$$\sigma_1^2(Y) = \sigma_2(Y) - \frac{\sigma_2^2(T_Y)}{\sigma_2^2(T_V)}\left[\sigma_1^2(V) - \sigma_2^2(V)\right]. \quad (4.39)$$

Utiliza-se, geralmente, no método de Levine, equações obtidas a partir de um modelo chamado de **congenérico**. Nesse modelo, além de E_X e T_X, bem como E_V e T_V não serem correlacionados, temos uma suposição adicional de que T_X e T_V são linearmente e perfeitamente correlacionados:

$$T_x = \lambda_x T + \delta_x, \quad (4.40)$$

e

$$T_v = \lambda_v T + \delta_v. \quad (4.41)$$

Isolando T em (4.41) e substituindo em (4.40) temos:

$$T_x = \frac{\lambda_x}{\lambda_v} T_v + \left(\delta_x - \frac{\lambda_x}{\lambda_v}\delta_v\right). \quad (4.42)$$

Sob o modelo congenérico, as equações (4.23) e (4.25) ficam:

$$X = T_x + E_x = \lambda_x T + \delta_x + E_x, \quad (4.43)$$

$$V = T_v + E_v = \lambda_v T + \delta_v + E_v. \quad (4.44)$$

O modelo congenérico ainda assume que:

$$\sigma^2(E_x) = \lambda_x \sigma^2(E), \quad (4.45)$$

$$\sigma^2(E_v) = \lambda_v \sigma^2(E), \quad (4.46)$$

ou seja, as variâncias dos erros são proporcionais a λ_x e λ_v. Essas duas equações implicam em:

$$\frac{\sigma^2(E_X)}{\sigma^2(E_V)} = \frac{\lambda_X}{\lambda_V}. \tag{4.47}$$

A partir de (4.43) e aplicando as propriedades de variância e também (4.45) temos:

$$\sigma^2(X) = \lambda_X^2\sigma^2(T) + \lambda_X\sigma^2(E). \tag{4.48}$$

Analogamente:

$$\sigma^2(V) = \lambda_V^2\sigma^2(T) + \lambda_V\sigma^2(E). \tag{4.49}$$

A correlação entre X e V é dada por:

$$\sigma^2(X,V) = \lambda_X\lambda_V\sigma^2(T) + \sigma(E_X,E_V). \tag{4.50}$$

A partir de (4.43) e (4.44) obtemos:

$$\gamma = \frac{\sigma(T_X)}{\sigma(T_V)} = \frac{\lambda_X\sigma(T)}{\lambda_V\sigma(T)} = \frac{\lambda_X}{\lambda_V}, \tag{4.51}$$

em que γ é chamado de razão efetiva do comprimento do teste para $\mathbf{X}$ e $\mathbf{V}$, respectivamente. Por isso, devemos considerar duas situações:

1) $\mathbf{V}$ é interno, ou seja, está incluído em $\mathbf{X}$. Por isso, o comprimento total do teste é corresponde ao número de itens de $\mathbf{X}$. Seja A uma variável aleatória de um teste $\mathbf{A}$ que corresponde à parte não comum de $\mathbf{X}$ tal que X = A + V. Sob o modelo congenérico, a covariância entre os erros para A e V é nula, pois $\mathbf{A}$ e $\mathbf{V}$ são compostos por itens diferentes. Assim:

$$\sigma(E_X,E_V) = \sigma(E_{A+V},E_V) = \sigma(E_V,E_V) = \sigma^2(E_V) = \lambda_V\sigma^2(E). \tag{4.52}$$

Substituindo (4.52) em (4.50) temos:

$$\sigma(X,V) = \lambda_V\left[\lambda_X\sigma^2(T) + \sigma^2(E)\right]. \tag{4.53}$$

De (4.48) e (4.53):

$$\gamma = \frac{\lambda_X}{\lambda_V} = \frac{\sigma^2(X)}{\sigma(X,V)} = \frac{1}{\alpha(V \mid X)}. \tag{4.54}$$

Logo, o modelo de Levine sob o modelo congenérico para o caso de um teste âncora interno usa:

$$\gamma_1 = \frac{1}{\alpha_1(V \mid X)} = \frac{\sigma_1^2(X)}{\sigma_1(X,V)}, \tag{4.55}$$

$$\gamma_2 = \frac{1}{\alpha_2(V \mid Y)} = \frac{\sigma_2^2(Y)}{\sigma_2(Y,V)}. \tag{4.56}$$

Esses resultados são aplicados às equações (4.19) a (4.22).

2) V é externo, ou seja, $\mathbf{X}$ e $\mathbf{V}$ não tem itens em comum. Sob o modelo congenérico:

$$\sigma(E_X, E_V) = 0. \tag{4.57}$$

A razão efetiva entre os comprimentos dos testes $\mathbf{X}$ e $\mathbf{V}$ é:

$$\gamma = \frac{\lambda_X}{\lambda_V} = \frac{\sigma^2(X) + \sigma(X,V)}{\sigma^2(V) + \sigma(X,V)}. \tag{4.58}$$

Portanto, o modelo de Levine sob o modelo congenérico para o caso de um teste âncora externo usa, nas equações (4.19) a (4.22):

$$\gamma_1 = \frac{\sigma_1^2(X) + \sigma_1(X,V)}{\sigma_1^2(V) + \sigma_1(X,V)}, \tag{4.59}$$

$$\gamma_2 = \frac{\sigma_2^2(Y) + \sigma_2(Y,V)}{\sigma_2^2(V) + \sigma_2(Y,V)}. \tag{4.60}$$

Além do método de Levine para escores observados, que acabamos de apresentar, existem o método de Levine para escores verdadeiros, dos quais não trataremos aqui. Senno (2006) e Kolen & Brennan (1995) mostram que os

resultados obtidos para ambos métodos são muito semelhantes. Por isso, o método para escores observados é mais utilizado visto que é mais simples.

4.2. Pesos nas populações sintéticas

Com respeito aos pesos w_1 e w_2, com $w_1+w_2=1$ e $w_1, w_2 \geq 0$, temos alguns casos a analisar:

1) $w_1=1$ e $w_2=0$: a população sintética é 1 (a que realiza a prova **X**).

2) $w_1 = \frac{N_1}{N_1+N_2}$ e $w_2 = \frac{N_2}{N_1+N_2}$ onde N_1 e N_2 são amostras das populações 1 e 2, respectivamente. Aqui, os pesos são proporcionais aos tamanhos das amostras.

3) $w_1=0,5$ e $w_2=0,5$: ambas populações têm igual importância para a população sintética.

Kolen & Brennan (1995) afirma que, na prática, os pesos raramente interferem de forma significativa nos parâmetros das equações de equalização tanto para os modelos de Tucker quanto de Levine. Em função disso, muitas vezes se utilizam os pesos $w_1=1$ e $w_2=0$, o que simplifica bastante algumas equações. Além disso, fazer $w_1=1$ significa que o grupo sintético é a "nova população", que é, geralmente, a única população que realiza a "nova forma" do teste no modelo de grupos não equivalentes.

4.3. Resumo dos métodos

Apresentaremos, a seguir, um quadro-resumo com as equações utilizadas nos métodos de Tucker e de Levine descritos na Seção 4.1.

Quadro-resumo 1: método de Tucker para equalização linear em modelos de grupos não equivalentes

$$l_{Y_s}(x) = y = \frac{\sigma_s(Y)}{\sigma_s(X)}x + \left[\mu_s(Y) - \frac{\sigma_s(Y)}{\sigma_s(X)}\mu_s(X)\right] \quad (4.1)$$

$$\mu_s(X) = \mu_1(X) - w_2\gamma_1\left[\mu_1(V) - \mu_2(V)\right] \quad (4.19)$$

$$\mu_s(Y) = \mu_2(Y) + w_1\gamma_2\left[\mu_1(V) - \mu_2(V)\right] \quad (4.20)$$

$$\sigma_s^2(X) = \sigma_1^2(X) - w_2\gamma_1^2\left[\sigma_1^2(V) - \sigma_2^2(V)\right] + w_1w_2\gamma_1^2\left[\mu_1(V) - \mu_2(V)\right]^2 \quad (4.21)$$

$$\sigma_s^2(Y) = \sigma_2^2(Y) + w_1\gamma_2^2\left[\sigma_1^2(V) - \sigma_2^2(V)\right] + w_1w_2\gamma_2^2\left[\mu_1(V) - \mu_2(V)\right]^2 \quad (4.22)$$

com

Testes âncora internos e externos

$$\gamma_1 = \alpha_1(X \mid V) = \frac{\sigma_1(X,V)}{\sigma_1^2(V)} \quad (4.61)$$

$$\gamma_2 = \alpha_2(Y \mid V) = \frac{\sigma_2(Y,V)}{\sigma_2^2(V)} \quad (4.62)$$

Obs.: a região pontilhada se anula quando usamos $w_1=1$ e $w_2=0$.

Quadro-resumo 2: método de Levine para escores observados sob o modelo congenérico na equalização linear em modelos de grupos não equivalentes

$$l_{Y_s}(x) = y = \frac{\sigma_s(Y)}{\sigma_s(X)}x + \left[\mu_s(Y) - \frac{\sigma_s(Y)}{\sigma_s(X)}\mu_s(X)\right] \quad (4.1)$$

$$\mu_s(X) = \mu_1(X) - w_2\gamma_1\left[\mu_1(V) - \mu_2(V)\right] \quad (4.19)$$

$$\mu_s(Y) = \mu_2(Y) + w_1\gamma_2\left[\mu_1(V) - \mu_2(V)\right] \quad (4.20)$$

$$\sigma_s^2(X) = \sigma_1^2(X) - w_2\gamma_1^2\left[\sigma_1^2(V) - \sigma_2^2(V)\right] + w_1w_2\gamma_1^2\left[\mu_1(V) - \mu_2(V)\right]^2 \quad (4.21)$$

$$\sigma_s^2(Y) = \sigma_2^2(Y) + w_1\gamma_2^2\left[\sigma_1^2(V) - \sigma_2^2(V)\right] + w_1w_2\gamma_2^2\left[\mu_1(V) - \mu_2(V)\right]^2 \quad (4.22)$$

com

Testes âncora internos

$$\gamma_1 = \frac{1}{\alpha_1(V \mid X)} = \frac{\sigma_1^2(X)}{\sigma_1(X,V)}, \quad (4.55)$$

$$\gamma_2 = \frac{1}{\alpha_2(V \mid Y)} = \frac{\sigma_2^2(Y)}{\sigma_2(Y,V)}. \quad (4.56)$$

Testes âncora externos

$$\gamma_1 = \frac{\sigma_1^2(X) + \sigma_1(X,V)}{\sigma_1^2(V) + \sigma_1(X,V)} \quad (4.59)$$

$$\gamma_2 = \frac{\sigma_2^2(Y) + \sigma_2(Y,V)}{\sigma_2^2(V) + \sigma_2(Y,V)} \quad (4.60)$$

Obs.: a região pontilhada se anula quando usamos $w_1=1$ e $w_2=0$.

4.4. Tucker X Levine

Kolen & Brennan (1995) afirma que o método de Levine é mais apropriado que o método de Tucker quando os grupos se diferem muito. Porém, se há suspeitas de

que as formas sejam muito diferentes, o método de Tucker é preferível. O conceito de "diferir" é um tanto complexo. Mas é certo que se as formas dos testes não têm conteúdo em comum, podemos dizer que as formas são, de fato, "muito diferentes". Tanto Tucker como Levine fazem suposições sobre linearidade que podem ser verificadas de modo direto. Por exemplo: podemos fazer a regressão de X sobre V na população 1. Se não for linear, as suposições não são satisfeitas. Uma alternativa é utilizar métodos eqüipercentis, conforme veremos adiante.

Exemplo: suponhamos dois testes hipotéticos **X** e **Y**, ambos com 40 itens dos quais 10 são comuns (teste **V**), ou seja, estamos falando de um teste âncora interno para grupos não equivalentes. Obtemos uma amostra da população 1 contendo N_1=1210 respondentes e da população 2 contendo N_2=1250 respondentes. Os resultados obtidos para os testes estão representados na tabela a seguir:

Tabela 1: Resumo estatístico para os testes hipotéticos **X**, **Y** e **V**

Grupo	Prova	$\hat{\mu}$	$\hat{\sigma}$	covariância	Correlação
1	X	17,4285	6,3445	9,9841	0,8779
1	V	3,9251	1,7924		
2	Y	20,1023	7,2521	11,1812	0,8474
2	V	4,5370	1,8193		

Vamos calcular as equações de equalização para os métodos de Tucker e Levine considerando-se, para efeitos de comparação, três situações para os pesos w_1 e w_2:

Situação 1: w_1=1 e w_2=0.

Situação 2: w_1=0,5 e w_2=0,5.

Situação 3: $w_1 = \frac{N_1}{N_1 + N_2} = \frac{1210}{1210 + 1250} = 0,4919$ e w_2 = 1 - 0,4919 = 0,5081.

As fórmulas utilizadas no método de Tucker estão apresentadas no Quadro-resumo 1. Assim, por (4.19) e lembrando que na situação 1 temos $w_1=1$ e $w_2=0$ obtemos:

$$\hat{\mu}_s(Y) = \hat{\mu}_1(X) = 17,4285,$$

$$\hat{\gamma}_2 = \frac{\hat{\sigma}_2(Y,V)}{\hat{\sigma}_2^2(V)} = \frac{11,1812}{1,8193^2} = 3,3782.$$

Por (4.20):

$$\hat{\mu}_s(Y) = 20,1023 + 3,3782\,[3,9251 - 4,5370]$$

$$\hat{\mu}_s(Y) = 18,0352.$$

Usando (4.21) e (4.22) temos os desvios-padrão:

$$\hat{\sigma}_s^2(X) = \hat{\sigma}_1^2(X) \Rightarrow \hat{\sigma}_s(X) = \hat{\sigma}_1(X) = 6,3445,$$

$$\hat{\sigma}_s(Y) = \sqrt{7,2521^2 + 3,3782^2\left[1,7924^2 - 1,8193^2\right]} = 7,1753.$$

Logo, a equação de equalização, usando (4.1), é:

$$\hat{l}_{Y_s}(x) = \frac{7,1753}{6,3445}\left[x - 17,4285\right] + 18,0352$$

$$\hat{l}_{Y_s} = 1,1309x - 1,6754. \qquad (4.63)$$

Usando o método de Levine para escores observados sob o modelo congenérico temos:

$$\hat{\mu}_s(X) = 17,4285.$$

Por (4.56):

$$\hat{\gamma}_2 = \frac{\sigma_2^2(Y)}{\sigma_2(Y,V)} = \frac{7,2521^2}{11,1812} = 4,7037.$$

Por (4.20) temos:

$$\hat{\mu}_s(Y) = 20,1023 + 4,7037.\left[3,9251 - 4,5370\right] = 17,2241.$$

De (4.21):

$$\hat{\sigma}_s(X) = 6,3445.$$

De (4.22):

$$\hat{\sigma}_s(Y) = \sqrt{7,2521^2 + 4,7037^2\left[1,7924^2 - 1,8193^2\right]} = 7,1024.$$

Logo, a equação de equalização fica:

$$\hat{l}_{Y_s}(x) = \frac{7,1024}{6,3445}\left[x - 17,4285\right] + 17,2241$$

$$\hat{l}_{Y_s}(x) = 1,1195x - 2,2862. \qquad (4.64)$$

Na tabela 4.2, a seguir, apresentamos os resultados obtidos para os métodos de Tucker e Levine considerando as três situações mencionadas anteriormente. Devido ao fato de os valores das situações 2 e 3 serem muito próximos, vamos acrescentar uma quarta situação, apenas para efeitos de comparação.

Observando-se as equações ajustadas, notamos que o método de Tucker produz resultados similares independente dos pesos w_1 e w_2 escolhidos. O mesmo comentário vale para o método de Levine. Porém, notamos diferenças entre os métodos de Tucker e Levine. Devido ao fato de as diferenças entre os resultados obtidos, considerando os diferentes pesos, serem pequenas, é comum, conforme dissemos em 4.2, utilizar w_1=1 e w_2=0. Enfatizamos, nesse argumento, a necessidade de se ter implementado esse processo, visto que exige uma grande quantidade de cálculos. Tal implementação será apresentada no Capítulo 6.

Tabela 4.2: resultados das equalizações para provas hipotéticas **X**, **Y** e **V**.

Método	w_1	w_2	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\mu}_s(X)$	$\hat{\mu}_s(Y)$	$\hat{\sigma}_s(X)$	$\hat{\sigma}_s(Y)$	$\hat{l}_{Y_s}(x)$
Tucker	1	0	——	3,3782	17,4285	18,0352	6,3445	7,1753	1,1309x–1,6754
Tucker	0,5	0,5	3,1077	3,3782	18,3793	19,0688	6,4518	7,2874	1,1295x–1,6910
Tucker	0,4919*	0,5081*	3,1077	3,3782	18,3947	19,0855	6,4524	7,2880	1,1295x–1,6915
Tucker	0,2	0,8	3,1077	3,3782	18,9498	19,6889	6,4484	7,2839	1,1296x–1,7161
Levine	1	0	——	4,7037	17,4285	17,2241	6,3445	7,1024	1,1195x–2,2862
Levine	0,5	0,5	4,0317	4,7037	18,6820	18,6865	6,5250	7,3216	1,1220x–2,2761
Levine	0,4919*	0,5081*	4,0317	4,7037	18,6620	18,6632	6,5241	7,3205	1,1221x–2,2768
Levine	0,2	0,8	4,0317	4,7037	19,4021	19,5267	6,5184	7,3136	1,1220x–2,2422

* N_1=1210 e N_2=1250. Logo, $w_1 = \frac{N_1}{N_1 + N_2} = 0,4919$ e $w_2 = 1 - 0,4919 = 0,5081$.

4.5. Equalização eqüipercentil

Os métodos eqüipercentis para grupos não equivalentes utilizam a distribuição do escore total e escore dos itens comuns. Considera, também, a população sintética. Analisaremos neste trabalho dois métodos: o de estimação de freqüências e o de Braun-Holland.

4.5.1. Método de estimação de freqüências

Este método estima a distribuição acumulada de escores na forma $\mathbf{X}$ e forma $\mathbf{Y}$ para uma população sintética obtida através do modelo de grupos não equivalentes. Consideremos duas formas $\mathbf{X}$ e $\mathbf{Y}$ e um teste $\mathbf{V}$ comum a ambas formas. Seja, X, Y e V variáveis aleatórias que representam o número de acertos x, y e v em cada teste, respectivamente. O método de estimação de freqüências utiliza distribuições condicionais de escores:

$$f(x \mid v) = \frac{f(x,v)}{h(v)}, \tag{4.65}$$

que nos fornece

$$f(x,v) = f(x \mid v).h(v), \tag{4.66}$$

onde

f(x,v) é a distribuição conjunta do escore total e do escore dos itens comuns. Em outras palavras, representa a probabilidade de X=x e V=v.

f(x|v) é a distribuição condicional de escores na forma $\mathbf{X}$ de indivíduos que obtiveram um escore particular v.

h(v) é a distribuição marginal de escores dos itens comuns.

Consideremos as distribuições para a população sintética para as duas populações 1 e 2 que realizam, respectivamente, as provas $\mathbf{X}$ e $\mathbf{Y}$:

$$f_s(x) = w_1 f_1(x) + w_2 f_2(x), \tag{4.67}$$

$$g_s(y) = w_1 g_1(y) + w_2 g_2(y), \tag{4.68}$$

onde
f é a distribuição para a forma **X**;
g é a distribuição para a forma **Y**;
$w_1+w_2=1$.

Notemos que $f_1(x)$ e $g_2(y)$ podem ser estimadas diretamente, mas $f_2(x)$ e $g_1(y)$ não podem. Para obtermos tais estimativas, faz-se a suposição de que a distribuição condicional de X e de Y para um dado escore comum v é a mesma em ambas populações. Ou seja:

$$f_1(x \mid v) = f_2(x \mid v), \tag{4.69}$$

$$g_1(y \mid v) = g_2(y \mid v). \tag{4.70}$$

A partir de (4.66) e utilizando (4.69) e (4.70) obtemos:

$$f_2(x,v) = f_2(x \mid v) h_2(v),$$

$$f_2(x,v) = f_1(x \mid v) h_2(v), \tag{4.71}$$

e

$$g_1(y,v) = g_1(y \mid v) h_1(v),$$

$$g_1(y,v) = g_2(y \mid v) h_1(v). \tag{4.72}$$

Como $f_1(x \mid v)$, $h_2(v)$, $g_2(y \mid v)$ e $h_1(v)$ podem ser estimados diretamente das populações, é possível obter $f_2(x,v)$ e $g_1(y,v)$. As distribuições marginais são tais que:

$$f_2(x) = \sum_v f_2(x,v) = \sum_v f_1(x \mid v) h_2(v), \tag{4.73}$$

$$g_1(y) = \sum_v g_1(y,v) = \sum_v g_2(y \mid v) h_1(v). \tag{4.74}$$

Substituindo (4.73) e (4.74) em (4.67) e (4.68), respectivamente, temos:

$$f_s(x) = w_1 f_1(x) + w_2 \sum_v f_1(x \mid v) h_2(v) \quad (4.75)$$

$$g_s(y) = w_1 \sum_v g_2(y \mid v) h_1(v) + w_2 g_2(y). \quad (4.76)$$

Para a população sintética, podemos obter, a partir de $f_s(x)$, a distribuição acumulada $F_s(x)$. Analogamente, obtemos $G_s(y)$ de $g_s(y)$. Seja P_s a função *rank* percentil da forma $\mathbf{X}$ e Q_s a da forma $\mathbf{Y}$, conforme já explicado no Capítulo 2. A função equipercentil é:

$$e_{Y_s} = Q_s^{-1}\left[P_s(x)\right]. \quad (4.77)$$

A suposição de estimação de freqüências feitas para as equações (4.69) e (4.70) não podem ser verificadas diretamente pois, para isso, a população 1 deveria realizar o teste na forma $\mathbf{Y}$ e a população 2, na forma $\mathbf{X}$. Porém, na prática, isso não ocorre. Essa suposição é obtida quando as populações 1 e 2 são razoavelmente idênticas. Logo, quanto mais similar for a população 1 da população 2, mais próximos da suposição nós chegaremos. Logo, a equalização pelo método de estimação de freqüências só poderá ser feita quando as duas populações forem razoavelmente idênticas entre si. O quão razoavelmente idênticas elas são depende do contexto da equalização e do grau de similaridade requerido (Kolen & Brennan, 1995). Quando as populações são diferentes, Kolen & Brennan (1995) sugere a utilização da TRI.

4.5.2. Exemplo

Para ilustrar, consideremos dois testes hipotéticos $\mathbf{X}$ e $\mathbf{Y}$ compostos por 6 itens cada e com 4 itens comuns (totalizando provas com 10 itens). As distribuições conjuntas de $\mathbf{X}$ e $\mathbf{V}$ e de $\mathbf{Y}$ e $\mathbf{V}$ estão mostradas nas tabelas a seguir.

Tabela 4.3: Distribuição de probabilidades para a população 1 da forma X e dos itens comuns

	v						
x	0	1	2	3	4	$f_1(x)$	$F_1(x)$
0	0,035	0,031	0,027	0,020	0,019	0,132	0,132
1	0,033	0,032	0,028	0,022	0,023	0,138	0,270
2	0,031	0,032	0,030	0,029	0,027	0,149	0,419
3	0,028	0,034	**0,035**	0,031	0,029	0,157	0,576
4	0,027	0,030	0,033	0,032	0,030	0,152	0,728
5	0,024	0,023	0,030	0,030	0,033	0,140	0,868
6	0,018	0,022	0,028	0,029	0,035	0,132	1,000
$h_1(v)$	0,196	0,204	0,211	0,193	0,196	1,000	

Tabela 4.4: Distribuição de probabilidades para a população 2 da forma Y e dos itens comuns

	v						
y	0	1	2	3	4	$g_2(y)$	$G_2(y)$
0	0,036	0,034	0,028	0,021	0,020	0,139	0,139
1	0,033	0,031	0,028	0,021	0,022	0,135	0,274
2	0,031	0,032	0,031	0,028	0,026	0,148	0,422
3	0,027	0,035	0,034	0,030	0,027	0,153	0,575
4	0,027	0,031	0,034	0,032	0,031	0,155	0,730
5	0,021	0,022	0,032	0,030	0,033	0,138	0,868
6	0,019	0,021	0,029	0,028	0,035	0,132	1,000
$h_2(v)$	0,194	0,206	0,216	0,190	0,194	1,000	

Os valores na parte inferior das tabelas 4.3 e 4.4 representam as distribuições marginais dos itens comuns para a população 1 ($h_1(v)$) e população 2 ($h_2(v)$).

Para exemplificar, temos que, na Tabela 4.3, a probabilidade de um indivíduo obter escore 3 na prova X e escore 2 nos itens comuns é 0,035. A probabilidade de um indivíduo na população 1 obter escore 2 nos itens comuns é 0,211. A probabilidade de um indivíduo da população 1 obter escore x=3 é 0,157.

Suponhamos $w_1=1$ e $w_2=0$. Assim, as equações (4.75) e (4.76) ficam:

$$f_s(x) = f_1(x), \tag{4.78}$$

$$g_s(y) = \sum_v g_2(y \mid v) h_1(v). \tag{4.79}$$

Logo, a distribuição dos escores da população sintética, por (4.78), é a mesma que a da população 1, que é o grupo sintético pois w_1=1. Conseqüentemente, $F_s(x)$=$F_1(x)$, que já foi calculado na Tabela 4.3. Para calcular $g_s(y)$ devemos obter, antes, $g_2(y|v)$ que é dado por:

$$g_2(y \mid v) = \frac{g_2(y,v)}{h_2(v)}. \tag{4.80}$$

A partir da Tabela 4.4 e utilizando (4.80) obtemos os seguintes resultados apresentados na Tabela 4.5:

Tabela 4.5: Distribuição condicional da forma Y dados os escores dos itens comuns para a população 2.

			v		
y	**0**	**1**	**2**	**3**	**4**
0	0,186	0,165	0,130	0,111	0,103
1	0,170	0,150	0,130	0,111	0,113
2	0,160	0,155	0,144	0,147	0,135
3	0,139	0,170	0,157	0,158	0,139
4	0,139	0,150	0,157	0,168	0,160
5	0,108	0,107	0,148	0,158	0,170
6	0,098	0,103	0,134	0,147	0,180
soma	1,000	1,000	1,000	1,000	1,000

Para encontrarmos os valores para serem substituídos na equação (4.79) devemos multiplicar os resultados obtidos na Tabela 4.5 pelos valores de $h_1(v)$ apresentados na Tabela 4.3. Os resultados são mostrados na Tabela 4.6.

Tabela 4.6: Distribuição da forma Y e dos escores dos itens comuns para a população 1 usando as suposições de estimação de freqüência

	v						
y	0	1	2	3	4	$g_1(y)$	$G_1(y)$
0	0,036	0,034	0,027	0,021	0,020	0,139	0,139
1	0,033	0,031	0,027	0,021	0,022	0,135	0,274
2	0,031	0,032	0,030	0,028	0,026	0,148	0,422
3	0,027	0,035	0,033	0,030	0,027	0,153	0,575
4	0,027	0,031	0,033	0,033	0,031	0,155	0,730
5	0,021	0,022	0,031	0,030	0,033	0,138	0,868
6	0,019	0,021	0,028	0,028	0,035	0,132	1,000
h1(v)	0,196	0,204	0,211	0,193	0,196	1,000	

Assim, conseguimos estimar a distribuição acumulada $G_1(y)$, que não pode ser diretamente observada pelo fato da população 1 ter realizado apenas o teste na forma Y. A partir disso, podemos calcular o equipercentil equivalente na forma Y ($e_Y(x)$) conforme já foi explicado no capítulo anterior. Utilizando (3.6):

$$e_Y(x) = Q^{-1}[P(x)] = \frac{P(x)/100 - G(y^* - 1)}{G(y^*) - G(y^* - 1)} + (y^* - 0{,}5)\ ,\ 0 \le P(x) < 100$$

$$= K_Y + 0{,}5\ ,\ P(x)=100$$

obtemos a Tabela 4.7.

Tabela 4.7: Distribuições acumuladas e equipercentil equivalente admitindo um teste âncora externo

x	$F_1(x)$	$P_1(x)$ (%)	y	$G_1(y)$	$Q_1(y)$ (%)	x	$e_Y(x)$	y*
0	0,132	6,60	0	0,139	6,95	0	-0,03	0
1	0,270	20,10	1	0,274	20,64	1	0,96	1
2	0,419	34,45	2	0,422	34,79	2	1,98	2
3	0,576	49,75	3	0,575	49,83	3	2,99	3
4	0,728	65,20	4	0,730	65,23	4	4,00	4
5	0,868	79,80	5	0,868	79,88	5	4,99	5
6	1,000	93,40	6	1,000	93,39	6	6,00	6

Observação: um possível problema de estimação ocorre quando nenhum indivíduo recebe um particular escore dos itens comuns em um grupo, mas existem indivíduos do outro grupo que receberam tal escore. Pela suposição feita, temos que $g_1(y|v)=g_2(y|v)$ para todo v. Mas, se ninguém teve determinado escore v na amostra, não é possível estimar $g_1(y|v)$. Jarjoura & Kolen (1985) sugere para a

distribuição condicional, o uso de um escore próximo a v (como v+1), o que acaba por gerar um erro muito pequeno na prática.

4.5.3. Método linear de Braun-Holland

Este método utiliza a média e o desvio-padrão sob as mesmas suposições da estimação de freqüências para realizar uma equalização linear. Para a população sintética temos:

$$\mu_s(X) = \sum_x x.f_s(x), \tag{4.81}$$

$$\sigma_s^2(X) = \sum_x [x - \mu_s(X)]^2 .f_s(x), \tag{4.82}$$

onde $f_s(x)$ é dado por (4.75). Os resultados obtidos nessas equações podem ser substituídos em (4.1). Considerando o exemplo apresentado anteriormente, das tabelas 4.3 a 4.6, e usando (4.81) e (4.82) temos:

Tabela 4.8: Média e desvio-padrão para o método de Braun-Holland nos testes **X** e **Y**

x	$f_1(x)$	y	$g_1(y)$
0	0,132	0	0,139
1	0,138	1	0,135
2	0,149	2	0,148
3	0,157	3	0,153
4	0,152	4	0,155
5	0,140	5	0,138
6	0,132	6	0,132
$\mu_1(X)$	**3,007**	**$\mu_1(Y)$**	**2,993**
$\sigma_1(X)$	**3,745**	**$\sigma_1(Y)$**	**3,746**

Logo, a equação de equalização é dada por:

$$l_{Y_s}(x) = \frac{3,746}{3,745}[x - 3,007] + 2,993$$

$$l_{Y_s}(x) = 1,000x - 0,015 \tag{4.83}$$

Usando (4.83) calculamos os valores equalizados, conforme tabela 4.9:

Tabela 4.9: resultados da equalização pelo método de de Braun-Holland

x	$l_{Ys}(x)$
0	-0,015
1	0,985
2	1,985
3	2,985
4	3,985
5	4,985
6	5,985

Segundo Kolen & Brennan (1995), o método de Braun-Holland é idêntico ao método de Tucker quando:

1) a regressão de X sobre V e de Y sobre V são lineares;

2) a regressão de X sobre V e de Y sobre V são homocedásticas, ou seja, a variância de X dado v é a mesma para todo v e a variância de Y dado v é a mesma para todo v.

Por isso, Braun-Holland pode ser visto como uma generalização de Tucker, embora seja computacionalmente mais complicado. Braun-Holland é preferível em relação a Tucker quando a regressão não é linear (Kolen & Brennan, 1995).

5. Teoria da Resposta ao Item: Análise de Itens, Estimação e Equalização

Ao se trabalhar com a TC vários problemas podem ser apontados. Inicialmente, temos que as características dos indivíduos e dos testes não podem ser separados: cada um é medido em função do outro. A dificuldade do item é dada pela proporção de acertos do item, o que dificulta a comparação de indivíduos que realizaram testes diferentes, e dificulta, também, a comparação de itens cujas características foram obtidas usando grupos diferentes de examinados. A TC tem dificuldades de quantificar, por exemplo, o quanto diferente em habilidade são dois indivíduos que realizaram dois testes diferentes. Dois outros pontos de insatisfação com a TC são referentes à confiabilidade e ao erro padrão. A confiabilidade é definida como a correlação entre os escores de testes paralelos os quais, na prática, são difíceis de serem encontrados. Quanto ao erro padrão, assume-se que é o mesmo para todos os indivíduos, o que não é uma suposição plausível. Finalmente, a TC não consegue quantificar a probabilidade de um indivíduo responder corretamente a um item, ou seja, a TC não fornece um modelo que tenha uma medida de precisão para cada habilidade. A TRI foi criada para solucionar os problemas apresentados pela TC.

A TRI propõe a utilização de modelos de variáveis latentes para representar a relação entre a probabilidade de um aluno apresentar uma certa resposta a um item e seus traços latentes ou habilidades na área do conhecimento avaliada, os quais não são observados diretamente. Ao contrário da TC, a TRI tem como elemento central o item e não a prova como um todo, permitindo, assim, a comparação de indivíduos de populações distintas que foram submetidos a provas diferentes, mas que possuíam alguns itens comuns; também permite comparar indivíduos de uma mesma população submetidos a provas totalmente diferentes, com ou sem itens comuns. Outras aplicações da TRI sugeridas por Kolen & Brennan (1995) são: desenvolvimento de testes, criação de bancos de itens e comparações entre eles, adaptação de testes e equalização de testes. Os vários

modelos propostos na literatura dependem da natureza do item (dicotômicos ou não), do número de populações envolvidas (uma ou mais) e da quantidade de traços latentes que está sendo medida (Andrade et al 2000). Na literatura encontramos vários modelos, como, por exemplo, modelos logísticos de 1, 2 ou 3 parâmetros, de resposta nominal, de resposta gradual, de escala gradual, dentre outros. Uma descrição desses modelos pode ser encontrada em Valle (1999) e Andrade et al (2000).

5.1 Modelo logístico com 3 parâmetros (ML3)

Um dos modelos mais utilizados, e que será abordado neste trabalho, é o modelo logístico unidimensional (avalia apenas um traço latente ou habilidade) de 3 parâmetros para itens de múltipla escolha dicotômicos ou dicotomizados (do tipo certo ou errado) para um único grupo e é dado por:

$$P(U_{ij}=1|\theta_j)=c_i+(1-c_i)\frac{1}{1+e^{-Da_i(\theta_j-b_i)}}, \qquad (5.1)$$

com i=1,2,...,I e j=1,2,...,n, onde:
U_{ij} é uma variável dicotômica que assume valores 1 (quando o indivíduo j responde corretamente o item i) ou 0 (quando o indivíduo j responde incorretamente o item i).

θ_j representa a habilidade (traço latente) do j-ésimo indivíduo.

$P(U_{ij}=1 \mid \theta_j)$ é a probabilidade de um indivíduo j com habilidade θ_j responder corretamente ao item i.

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

a_i é o parâmetro de discriminação (ou de inclinação) do item i, com valor proporcional à inclinação da Curva Característica do Item (CCI) no ponto b_i, conforme será mostrado a seguir.

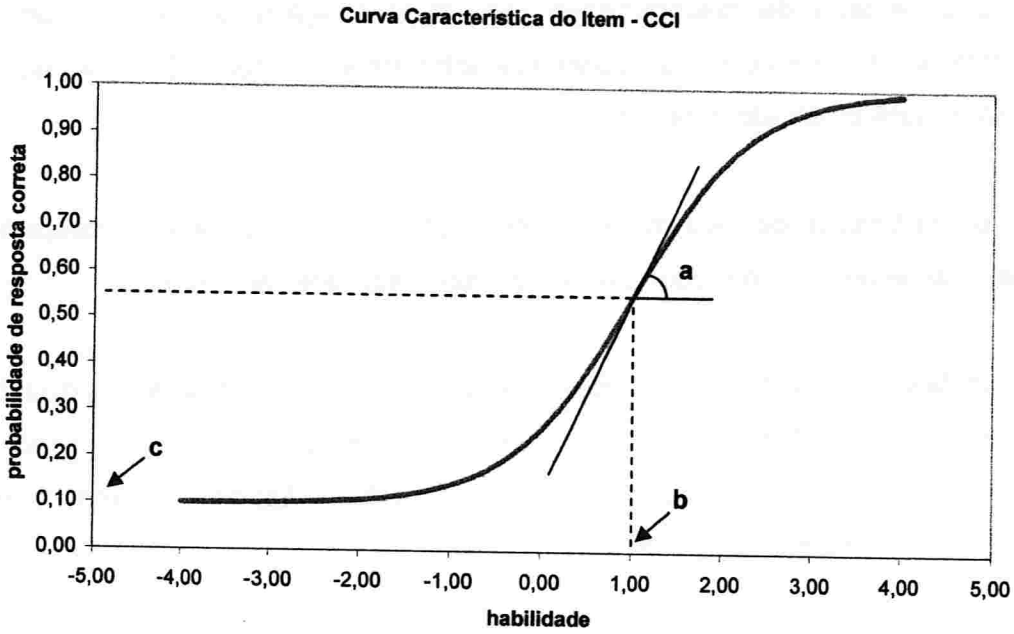
c_i é o parâmetro de acerto ao acaso do item i, ou seja, corresponde à probabilidade de um indivíduo com baixa habilidade acertar o item i.

D é um fator de escala constante e conhecido. Ele será igual a 1,7 quando se deseja que a função logística forneça resultados semelhantes ao da função ogiva normal. Caso contrário, utiliza-se D=1, que é o valor utilizado no desenvolvimento das implementações realizadas.

5.1.1. Interpretação e Curva Característica do Item (CCI)

Enxergando $P(U_{ij}=1|\theta)$ como sendo a proporção de respostas corretas apresentadas por indivíduos com habilidade θ ao responder o item i, podemos construir o gráfico de probabilidade de acerto em função da habilidade para um item i. Esse gráfico é chamado de Curva Característica do Item - CCI, conforme mostrado na Figura 5.1.

Figura 5.1: Exemplo de uma Curva Característica do Item



Podemos notar no gráfico que o modelo proposto (ML3) baseia-se no fato de que indivíduos com maior habilidade possuem maior probabilidade de acertar o item e que esta relação não é linear. A escala de habilidades é arbitrária: não importa necessariamente os seus valores, mas sim as relações de ordem existentes entre os pontos dessa escala. O parâmetro **c** representa a probabilidade de acerto ao acaso, ou seja, representa a probabilidade de um indivíduo com baixa habilidade acertar o item. Na CCI, esse valor corresponde ao valor do eixo das ordenadas de uma das assíntotas horizontais (a outra assíntota tem valor de ordenada igual a 1). Assim, o valor de **c** independe da escala adotada. Quando não é permitido "chutar", **c** é igual a zero. O parâmetro **b** é medido na mesma escala da habilidade e representa a habilidade necessária para uma probabilidade de acerto igual a (1+**c**)/2. Assim, quanto maior é o valor de **b**, maior é a habilidade exigida por ele.

O parâmetro **a** corresponde ao valor do coeficiente angular da reta tangente à curva no ponto de inflexão. Apesar de ter o mesmo nome na TC, na TRI o parâmetro de dificuldade do item não é medido por uma proporção (valor entre 0 e 1) e o parâmetro de discriminação não corresponde à uma correlação (valor entre -1 e 1). Teoricamente, na TRI, esses dois parâmetros podem assumir qualquer

valor real compreendido entre $-\infty$ e $+\infty$. Veremos, na próxima seção, os valores admissíveis para os parâmetros. Na prática, as habilidades e os parâmetros dos itens são estimados a partir das respostas de um grupo de indivíduos submetidos a esses itens. Uma vez estabelecida a escala de habilidades, os valores dos parâmetros não se alteram, ou seja, tais valores são característicos do item e passam a não depender mais dos indivíduos (desde que estes indivíduos tenham suas habilidades medidas na mesma escala), ao contrário do que ocorre na TC, em que os valores dos parâmetros dependem dos indivíduos que realizam a prova.

5.1.2. Suposições do modelo

O modelo apresentado (ML3) tem duas suposições que devem ser satisfeitas para que seja possível a sua utilização:

1) Unidimensionalidade – deve haver uma homogeneidade do conjunto de itens que supostamente deve medir um único traço latente (habilidade). Em outras palavras, deve haver apenas uma habilidade responsável pela realização de todos os itens da prova. É razoável supor que mais de um traço latente seja medido na realização de um teste, porém, para satisfazer o postulado da unidimensionalidade, basta supormos que haja uma habilidade dominante responsável pelo conjunto de itens (Valle, 1999).

2) Independência local – também chamada de independência condicional. Assume que para uma dada habilidade as respostas aos diferentes itens da prova são independentes. Esta suposição é fundamental para o processo de estimação dos parâmetros do modelo. Na realidade, como unidimensionalidade implica independência local, tem-se somente uma e não duas suposições a serem verificadas. Assim, itens devem ser elaborados de modo a satisfazer a suposição de unidimensionalidade (Valle, 1999). Sob independência local, a probabilidade de que um indivíduo com habilidade θ responda corretamente aos itens 1 e 2 é igual

ao produto da probabilidade de se responder corretamente o item 1 pela probabilidade de se responder corretamente ao item 2 (Kolen & Brennan, 1995).

5.2. Análise de Itens

Como já foi dito na seção 5.1, o valor do parâmetro **c** (acerto casual) corresponde a uma probabilidade. Temos, para um item i, $0 \leq \mathbf{c_i} \leq 1$. Com relação ao parâmetro de dificuldade (**b**), temos, para um item i, $-\infty < \mathbf{b_i} < +\infty$. Porém, quando a escala utilizada é (0,1), ou seja, tem média 0 e desvio-padrão igual a 1, os valores de **b** variam tipicamente entre -2 e 2 quando observamos cada grupo separadamente nessa escala. Valores muito elevados indicam que o item é muito difícil; valores muito pequenos indicam que o item é muito fácil. Em ambas situações esse item pode não ser capaz de identificar os alunos com melhor desempenho daqueles que possuem baixo rendimento, visto que um item muito fácil é acertado pela maioria dos alunos, enquanto que um item muito difícil é errado pela maior parte dos alunos. O parâmetro de discriminação (**a**) também pode assumir qualquer valor real, como já foi dito anteriormente, no intervalo de $-\infty$ a $+\infty$. Porém, não devemos ter itens com valores negativos de **a**, visto que isso indicaria que o item é deficiente, pois ele teria uma alta proporção de acertos para os alunos com baixa habilidade e poucos acertos no grupo dos que apresentam uma habilidade maior. Ou seja, é fundamental que um item tenha $\mathbf{a}>0$. Tipicamente, deseja-se que os valores de a oscilem entre 1,0 e 2,0 (itens com $a<0,75$ não apresentam boa discriminação). Ao analisarmos os valores das estimativas das habilidades dos indivíduos, também desejamos que, na escala (0,1) as habilidades oscilem (tipicamente) entre -2 e 2. Outro fator importante na hora de se decidir se um item é "bom" ou não, é observar os valores dos erros padrões das estimativas dos parâmetros. Valores muito altos de erros padrões, principalmente com relação aos parâmetros **a** e **b**, podem indicar que o item é deficiente. Maiores detalhes sobre tais parâmetros podem ser encontrados em Kolen & Brennan (1995), Valle (1999) e Andrade (2001).

5.3. Estimação dos parâmetros dos itens (calibração) e das habilidades

Uma etapa importante da TRI está no processo de estimação dos parâmetros dos itens, que é conhecido por calibração, e das habilidades. Basicamente, temos três situações possíveis:

1) quando se conhece os parâmetros dos itens e se deseja estimar as habilidades dos indivíduos: nesta situação geralmente utiliza-se um banco de itens já calibrados e deseja-se estimar as habilidades dos indivíduos de modo a poder classificá-los, como por exemplo em exames vestibulares;

2) quando se conhecem as habilidades dos indivíduos e deseja-se estimar os parâmetros dos itens: esta situação é, na prática, muito difícil de ocorrer, visto que, geralmente temos conhecimento a respeito dos itens e não sobre os indivíduos;

3) quando desconhecemos os parâmetros dos itens e as habilidades: é uma situação muito comum na prática. Porém, aqui, surgem problemas computacionais dependendo do método de estimação escolhido, como por exemplo, o de se inverter matrizes de grandes dimensões quando se utiliza o processo de estimação por Máxima Verossimilhança Conjunta (Baker, 1992).

Em todas as situações, assume-se o modelo proposto como verdadeiro e, a partir das respostas dadas por um grupo de indivíduos a um conjunto de itens, os parâmetros dos itens e/ou habilidades são estimados a partir do método de máxima verossimilhança ou métodos bayesianos. Ambos métodos exigem procedimentos iterativos que envolvem cálculos bastante complexos. Nos três casos citados, tanto os parâmetros dos itens como as habilidades são estimados numa mesma escala, geralmente a (0,1).

Vários autores têm sugerido que cada respondente seja submetido a pelo menos 30 itens e que cada item seja submetido a pelo menos 300 respondentes, para que se obtenham estimativas com erros padrões pequenos (Andrade, 2001).

Neste trabalho, as implementações feitas referentes ao processo de estimação se basearam em métodos de máxima verossimilhança para os casos em que os parâmetros devem ser estimados quando as habilidades são desconhecidas e no caso em que se deseja estimar as habilidades tendo os parâmetros conhecidos.

5.3.1. Estimação por máxima verossimilhança conjunta

Sejam X_{ij} a variável aleatória dicotômica em que 1 indica acerto e 0 indica erro do i-ésimo item respondido pelo j-ésimo indivíduo, com i=1,2,...,I e j=1,2,...,n. Seja o vetor aleatório (Ix1) $X_j=(X_{1j},...,X_{Ij})^t$ que representa as respostas do j-ésimo indivíduo a todos os itens, $\psi_i=(a_i, b_i, c_i)^t$ o vetor dos parâmetros do i-ésimo item e $\psi=(\psi_1^t, ..., \psi_I^t)^t$ o vetor dos parâmetros de todos os itens e $\theta=(\theta_1,..., \theta_n)^t$ o vetor (nx1) das habilidades dos n indivíduos. Seja também $P_{ij} = P(X_{ij}=1|\theta_j)$ o modelo logístico unidimensional de 3 parâmetros com $Q_{ij}=1-P_{ij}$. Sob a suposição de independência local, a probabilidade do vetor de resposta x_j do indivíduo j condicionado na sua habilidade e nos parâmetros dos itens é dada por

$$P_j(x_j \mid \theta_j, \psi) = \prod_{i=1}^{I} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}},$$

e a função de verossimilhança, baseada nas respostas de uma amostra aleatória de n indivíduos é dada por

$$L(x_1,...,x_n \mid \theta, \psi) = \prod_{j=1}^{n} P_j(x_j \mid \theta_j, \psi).$$

Tomando o logaritmo natural da função de verossimilhança, temos:

$$\ln L(x_1,...,x_n \mid \theta, \psi) = \sum_{j=1}^{n} \sum_{i=1}^{I} \left[x_{ij} \ln P_{ij} + (1-x_{ij}) \ln Q_{ij} \right].$$

Para se obter os estimadores por máxima verossimilhança, é necessário derivar, igualar a zero e resolver 3I+n equações simultaneamente. Trata-se de um problema computacional bastante complexo. Além disso, há um problema de indeterminação associada ao modelo, em que valores que maximizam a função não podem ser determinados de modo único. Esse problema não ocorre quando se conhecem as habilidades e se deseja estimar os parâmetros ou quando se conhecem os parâmetros e se deseja estimar as habilidades. Para solucionar o problema de indeterminação, define-se uma escala arbitrária para os valores das habilidades ou para os valores dos parâmetros de dificuldade **b**, visto que ambos são medidos na mesma escala. Usualmente, estabelece-se os valores 0 para a média e 1 para o desvio padrão dos valores das habilidades. (Andrade, 2001 e Valle, 1999). O método de máxima verossimilhança marginal, apresentado a seguir, procura uma solução mais simples para este problema.

5.3.2. Estimação de máxima verossimilhança marginal

Para contornar os problemas descritos em 5.3.1., uma saída é utilizar um procedimento dividido em duas etapas: inicialmente os parâmetros dos itens são estimados, assumindo-se uma distribuição para as habilidades dos indivíduos. Depois, as habilidades são estimadas supondo-se que os parâmetros dos itens estimados na etapa anterior são os verdadeiros valores dos parâmetros.

Ao contrário do procedimento anterior, aqui faz-se uma suposição sobre a distribuição da habilidade: é assumido que os respondentes representam uma amostra aleatória de uma população na qual a habilidade é distribuída segundo uma determinada função densidade $g(\theta|\tau)$, onde τ é o vetor dos parâmetros desta distribuição.

Estimativas de máxima verossimilhança para os parâmetros dos itens são obtidas a partir da maximização da função de verossimilhança marginal:

$$L(x_1,\ldots,x_n \mid \psi) = \prod_{j=1}^{n} \int P_j(x_j \mid \theta_j, \psi).g(\theta_j \mid \tau)d\theta,$$

que depende das habilidades somente através da distribuição a priori $g(\theta)$. Note que a mesma distribuição a priori é assumida para todos os θ's. O problema de indeterminação do modelo é resolvido ao estabelecer-se a distribuição a priori, isto é, no final do processo de estimação tem-se as estimativas dos parâmetros dos itens em uma métrica definida pelos parâmetros de locação e de escala da priori. Em geral, utiliza-se como priori a distribuição normal com média 0 e desvio padrão 1 (Andrade, 2001 e Valle, 1999). Vale ressaltar que não se está aplicando nenhum argumento bayesiano. A densidade $g(\theta)$ pode ser considerada no sentido de se realizar um experimento de retirar um indivíduo dessa população e observar seu traço latente θ (Azevedo, 2003). A abordagem de estimação que citamos é conhecida como abordagem de Bock & Lieberman, a qual conduz a necessidade de estimação conjunta dos parâmetros de todos os itens e, embora produza estimadores com propriedades assintóticas desejáveis, é inaplicável para testes com mais de 12 itens (Azevedo, 2003). Uma alternativa eficaz é a utilização da abordagem de Bock & Aitkin que, basicamente, propuseram uma reformulação conveniente nas equações de estimação dos parâmetros dos itens e o uso de uma adaptação do algoritmo EM. Esta abordagem parte da suposição de que os itens são independentes de forma que

$$\frac{\partial^2 \ln L(\psi,\eta)}{\partial\psi_i \partial\psi_l} = 0, \text{ para } i \neq l,$$

com $\eta = (\mu, \sigma^2)$, onde μ é a média e σ^2 a variância das habilidades dos indivíduos. A proposta de Bock & Aitkin foi adotar a independência entre os itens de forma a possibilitar, através de algum processo iterativo, que os itens sejam estimados individualmente. Vale notar que as suposições de independência local e a suposição de independência dos itens são diferentes: a primeira está relacionada às respostas dos indivíduos, enquanto que a segunda se refere apenas aos itens (Andrade, et al 2000). Essa abordagem resulta em estimativas consistentes para os parâmetros dos itens e é computacionalmente muito mais simples. Com as estimativas dos parâmetros dos itens considerados como sendo os verdadeiros

valores dos parâmetros, estimam-se as habilidades dos respondentes através de métodos de máxima verossimilhança (ou bayesianos) na mesma métrica dos parâmetros dos itens, conforme veremos na seção a seguir. Em algumas situações, estes procedimentos podem não fornecer resultados satisfatórios. Isso ocorre principalmente na estimação do parâmetro c devido à própria natureza do parâmetro, que está associado à probabilidade de acerto de indivíduos com habilidade muito pequena, que em geral não são muitos. Um problema similar ocorre com a estimação do parâmetro b de itens muito fáceis ou muito difíceis. Para o processo de estimação ser bem sucedido, é importante ter-se respondentes com habilidades cobrindo todo o espectro do conhecimento a ser avaliado. Nessas situações problemáticas, sugere-se que procedimentos bayesianos sejam utilizados a partir da incorporação de distribuições a priori também para os parâmetros dos itens. Os procedimentos bayesianos fornecem estimativas para todos os itens e habilidades, mesmo para os indivíduos que acertaram ou erraram todos os itens ou para itens respondidos corretamente ou erroneamente por todos os indivíduos (Valle, 1999). Uma solução para a utilização de métodos de máxima verossimilhança quando nos deparamos com escores nulos ou perfeitos, ou seja, com indivíduos que erraram ou acertaram todos os itens, é atribuir, ao indivíduo que obteve escore nulo, 0,5 ao item mais fácil e, ao indivíduo que obteve escore perfeito, atribuir 0,5 ao item mais difícil (Andrade et al 2000).

5.3.3. Estimação das habilidades quando os parâmetros dos itens são conhecidos

Temos, agora, a situação em que os parâmetros dos itens são conhecidos e desejamos estimar as habilidades dos indivíduos que responderam o teste. Na prática, esta situação é bastante freqüente e ocorre quando submetemos indivíduos a um teste cujos itens foram obtidos a partir de um banco de itens ou como conseqüência da estimação por verossimilhança marginal conforme apresentamos na seção 5.3.2. Como a calibração dos itens deve ser feita com um

número grande de indivíduos, de modo a cobrir todo o intervalo de variação dos valores do parâmetro **b** dos itens, a estimação das habilidades de um grupo pequeno de indivíduos é mais confiável se forem utilizados itens já calibrados (Andrade et al 2000). Sob a suposição de independência local, a probabilidade do vetor de respostas X_j do indivíduo j, condicionado a sua habilidade θ_j, é dada por:

$$P_j(X_j \mid \theta_j) = \prod_{i=1}^{I} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} .$$

Logo, o logaritmo natural da função verossimilhança é dado por:

$$\ln L(x_j \mid \theta_j) = \sum_{i=1}^{I} \left[x_{ij} \ln P_{ij} + (1 - x_{ij}) \ln Q_{ij} \right].$$

Assim, o estimador de máxima verossimilhança para a habilidade do j-ésimo indivíduo é o valor de θ que maximiza a função anterior. Ou seja, devemos resolver a equação:

$$\frac{\partial \ln L(x_j \mid \theta_j)}{\partial \theta_j} = 0 .$$

Como esta equação não possui solução explícita, faz-se necessário a utilização de processos iterativos como Newton-Raphson ou Scoring de Fisher. Como todo estimador de máxima verossimilhança, $\hat{\theta}_j$ (estimador de θ_j) é normalmente distribuído com média θ_j e variância dada pela inversa da matriz de informação

$$I(\theta_j) = -E\left(\frac{\partial^2 \ln L(\theta)}{\partial \theta_j^2} \right).$$

O erro padrão de $\hat{\theta}_j$ é igual à raiz quadrada de $I(\theta_j)$. Maiores detalhes dos métodos de estimação via máxima verossimilhança podem ser encontrados em Azevedo (2003) e Andrade, et al (2000), que são ótimas referências em português, e em Baker (1992).

5.4. Métodos de Equalização

Como já dissemos, equalizar significa equiparar, tornar comparável, o que no caso da TRI significa colocar parâmetros de itens provenientes de testes distintos ou habilidades de respondentes de diferentes grupos numa mesma métrica, isto é, numa escala comum, tornando os itens e/ou as habilidades comparáveis. Existem dois tipos de equalização:

1) equalização via população: quando um único grupo de respondentes é submetido a provas distintas, basta que todos os itens sejam calibrados conjuntamente para termos a garantia de que todos estarão na mesma métrica.

2) equalização via itens comuns: neste tipo de equalização há necessidade de termos testes com itens comuns, que servem de ligação entre duas populações que realizaram as provas.

Abordaremos neste trabalho apenas a segunda situação em que temos dois grupos fazendo dois tipos de prova com alguns itens em comum. Outros casos são descritos em Valle (1999). A equalização pode ser realizada durante o processo de calibração, através de modelos de grupos múltiplos, ou pode ser feita a posteriori, que é a situação correspondente às implementações feitas. Maiores detalhes sobre os modelos de grupos múltiplos podem ser obtidos em Andrade (2001).

5.4.1. Equalização a posteriori

A idéia deste método é que os itens e/ou habilidades provenientes de duas populações sejam calibrados separadamente e, após isso, se realize o processo de equalização. Obviamente, é necessário que haja itens comuns entre os dois testes, de modo que seja possível compará-los. Para cada um dos dois conjuntos temos, para os itens comuns, estimativas dos parâmetros dos itens e das

habilidades, cada uma na métrica das respectivas populações. Então, a partir de uma relação linear, colocamos os parâmetros de um dos conjuntos na métrica do outro conjunto. Com todos os itens na mesma métrica podemos estimar as habilidades de todos os respondentes. Suponhamos que temos dois grupos de indivíduos que realizam dois testes, **X** e **Y**, com itens comuns. Desejamos equalizar os parâmetros dos itens da prova **X** e das habilidades dos indivíduos que responderam essa prova para a mesma escala do teste **Y**. Pela propriedade de invariância, em que qualquer que seja a escala de medida, a probabilidade de um indivíduo responder corretamente a um item é sempre a mesma (Senno, 2006), dado que o modelo é adequado aos dados, os parâmetros **a** e **b** de um certo item apresentado a dois grupos de indivíduos devem satisfazer, a menos de flutuações amostrais, para o ML3, as seguintes relações lineares:

$$a_{Yi} = \frac{1}{\alpha} a_{Xi}, \tag{5.2}$$

$$b_{Yi} = \alpha . b_{Xi} + \beta, \tag{5.3}$$

$$c_{Yi} = c_{Xi}, \tag{5.4}$$

e

$$\theta_{Yj} = \alpha . \theta_{Xj} + \beta, \tag{5.5}$$

onde

a_{Xi}, b_{Xi} e c_{Xi} são os valores dos parâmetros de discriminação, dificuldade e acerto casual correspondente ao item i do teste **X** (i=1,...,I)

a_{Yi}, b_{Yi} e c_{Yi} são os valores dos parâmetros de discriminação, dificuldade e acerto casual correspondente ao item i do teste **Y** (i=1,...,I);

θ_{Xj} corresponde à habilidade de um indivíduo j (j=1,...,n) que fez a prova **X**;

θ_{Yj} corresponde à habilidade de um indivíduo j (j=1,...,n) que fez a prova **Y**.

Vale a pena ressaltar que no processo de equalização, os parâmetros de acerto casual (**c**) permanecem invariantes ao se mudar os demais parâmetros da métrica de um teste para a métrica de outro. Após a determinação de α e β, podemos facilmente transformar os parâmetros dos itens do teste **X** para a escala do teste **Y**, bem como as habilidades dos indivíduos que fizeram a prova **X** para a métrica

do teste **Y**, permitindo, assim, uma comparação entre os respondentes provenientes das duas populações bastando utilizar adaptações das equações (5.2) a (5.5):

$$a^*_{Yi} = \frac{1}{\alpha} a_{Xi}, \tag{5.6}$$

$$b^*_{Yi} = \alpha . b_{Xi} + \beta, \tag{5.7}$$

$$c^*_{Yi} = c_{Xi}, \tag{5.8}$$

e

$$\theta^*_{Yj} = \alpha . \theta_{Xj} + \beta, \tag{5.9}$$

onde

a^*_{Yi} corresponde ao valor do parâmetro de discriminação do item i da prova **X** (a_{Xi}) transformado para a escala do teste **Y**. De maneira análoga, b^*_{Yi} e c^*_{Yi} correspondem aos parâmetros de dificuldade e de acerto casual do item i da prova **X** (b_{Xi} e c_{Xi}) transformados para a escala do teste **Y**. Da mesma forma, θ^*_{Yj} corresponde a habilidade do indivíduo j que realizou a prova **X** (θ_{Xj}) transformado para a métrica do teste **Y**. Para determinarmos os coeficientes α e β, apresentaremos dois métodos: média-desvio (*mean-sigma*) e média-média (*mean-mean*). Estes métodos e outros estão detalhados em Kolen & Brennan (1995).

5.4.1.1. Método média-devio (*mean-sigma*)

Este método, ao contrário de uma regressão linear, é invariante (simétrico) em relação às variáveis utilizadas. Isto significa que podemos utilizar as equações de equalização tanto para transformar a métrica de **X** para **Y** como de **Y** para **X** que o resultado será o mesmo. Os coeficientes pelo método média-desvio para a transformação dos parâmetros da métrica do teste **X** para a escala do teste **Y** são:

$$\alpha = \frac{S_Y}{S_X},$$

e

$$\beta = \mu_Y - \alpha.\mu_X ,$$

onde S_X e S_Y são os desvios-padrão amostrais e μ_X e μ_Y são as médias amostrais dos parâmetros de dificuldade (**b**) dos itens comuns nos grupos que realizaram os testes **X** e **Y** respectivamente.

5.4.1.2. Método média- média (*mean- mean*)

Este método também é invariante (simétrico) e os coeficientes para a transformação dos parâmetros da métrica do teste **X** para a escala do teste **Y** são:

$$\alpha = \frac{\mu_Y^*}{\mu_X^*},$$

e

$$\beta = \mu_Y - \alpha.\mu_X ,$$

onde μ_X^* e μ_Y^* são as médias amostrais dos parâmetros de discriminação (a) e μ_X e μ_Y são as médias amostrais dos parâmetros de dificuldade (b) dos itens comuns nos grupos que realizaram os testes **X** e **Y** respectivamente.

5.4.1.3. Comparação entre os dois métodos

Os resultados produzidos pela equalização através do método média-desvio e média-média geralmente são diferentes. Kolen & Brennan (1995) argumenta que alguns autores dão preferência ao método média-desvio pelo fato de as estimativas do parâmetro **b** serem mais estáveis que as estimativas do parâmetro **a**, enquanto que outros autores dizem que o método média-média é preferível por que a média é, tipicamente, mais estável que o desvio-padrão, e o método média-média só utiliza médias na determinação de α e β. Maiores detalhes a respeito

destes métodos podem ser vistos em Kolen & Brennan (1995), Valle (1999) e Senno (2006).

6. Implementações utilizando o software R

Nos Capítulos anteriores foram apresentados métodos de análise de itens, de equalização na TC e na TRI, além de processos de calibração na TRI. Como vimos, para a realização de tais procedimentos há muitos cálculos, os quais são muito difíceis e trabalhosos de serem feitos à mão. Daí a necessidade de se implementar programas que realizem esses procedimentos de modo a torná-los acessíveis a um número maior de pessoas. Pensando nisso, foi utilizado o software **R** para se fazer tais implementações. O **R** é um pacote estatístico que é muito similar ao S-PLUS, porém é um software de código aberto e totalmente gratuito, permitindo ser instalado livremente em qualquer computador. O **R** pode ser baixado gratuitamente a partir do endereço **http://cran.r-project.org** e é compatível com o Windows, Mac e Linux. Neste Capítulo faremos uma apresentação das implementações desenvolvidas dizendo o que cada uma delas faz, bem como alguns aspectos importantes no que se refere aos algoritmos utilizados fazendo, para isso, uso de um conjunto de dados real.

6.1. Download dos softwares e implementações

Todas as implementações realizadas neste trabalho bem como os softwares aqui mencionados e manuais de utilização encontram-se disponíveis para download na página **http://www.inf.ufsc.br/~dandrade/Avaliacao_Educacional**. Nessa página há possibilidade de se baixar separadamente cada arquivo da implementação, bem como arquivos de exemplos, baixar um arquivo compactado (*.zip*) contendo todas as implementações ou, ainda, baixar o pacote *EstatR.exe* que contém um software que permite rodar todas as implementações (já incluídas nesse pacote). Esse software será apresentado mais adiante.

6.2. Técnica de programação

Para todas as implementações desenvolvidas, procurou-se criar programas, utilizando a linguagem do **R**, de modo que satisfizessem algumas idéias:

- tratar os programas da maneira mais simples possível, em termos de linguagem de programação, de modo a permitir que os programas possam ser facilmente adaptados;
- trabalhar com o menor número de "pacotes" adicionais para evitar problemas na hora de rodar as implementações;
- criar programas com o menor número de entradas possíveis facilitando o seu uso;
- tornar as implementações versáteis, permitindo, por exemplo, o usuário excluir ou incluir itens na sua análise de modo fácil e evitando ter que mexer no "corpo" dos programas;
- produzir saídas fáceis de se entender e "amigáveis", ou seja, saídas nas formas de gráficos e tabelas para facilitar a análise.

Em termos de pacotes adicionais, utilizamos a biblioteca *MASS*, que já acompanha o **R**, e é utilizada basicamente para a criação dos arquivos de saída de cada análise, e a biblioteca *tcltk* que produz as janelas de aviso como as emitidas ao final de cada programa.

6.3. Entradas e saídas dos programas

Para que os programas rodem corretamente, alguns parâmetros iniciais presentes no início de cada programa devem ser alterados. Assim, é necessário ter-se um certo conhecimento estatístico dos métodos a serem utilizados, visto que uma configuração incorreta dos parâmetros poderá gerar resultados que não farão sentido na prática. Todos programas possuem as explicações e mensagens escritas em língua portuguesa. Não há necessidade de alterar nenhum parâmetro

ao longo do programa. Isso é uma grande vantagem, pois não é preciso entender de programação para conseguir executar as implementações e isso também evita que alguma parte importante do programa seja alterada indevidamente e produza algum tipo de erro. Além disso, os usuários deverão criar arquivos específicos para as análises, contendo, por exemplo, os gabaritos, as respostas dos indivíduos, habilidades, parâmetros, conforme explicado no Apêndice. Tais arquivos devem ter nomes específicos e algumas configurações dentro deles devem ser realizadas. Esses arquivos são todos criados no formato texto (*.txt*), o que pode ser feito com algum tipo de editor de textos, como *Bloco de Notas* ou *Word*. Vale ressaltar que todos os programas criados realizam análises apenas para itens dicotômicos, ou seja, do tipo certo ou errado como, por exemplo, em provas do tipo teste onde apenas uma alternativa é correta. Porém, uma das versatilidades de todos os programas desenvolvidos é que eles aceitam a entrada de dados tanto na forma corrigida, ou seja, do tipo 0 ou 1 (0 para resposta incorreta e 1 para correta), como na forma do gabarito original (utilizando-se letras A,B,C,D...). Nas saídas dos programas, além do fato de estarem escritas em língua portuguesa, procurou-se gerar os resultados e estatísticas mais relevantes para cada tipo de análise realizada. Todos os resultados são sempre alocados em um arquivo de saída que é gerado no formato *.txt*, ou seja, um arquivo de texto que pode ser aberto com qualquer editor de textos simples. Em alguns programas também é gerado um arquivo de imagem (*.jpg*) contendo gráficos obtidos na análise realizada.

6.4. Carregamento dos programas

Há duas possibilidades para se carregar os programas:

1) Usando apenas o **R** – o usuário poderá abrir o **R** e abrir o arquivo contendo a implementação que deseja rodar. Após fazer as configurações iniciais, ele pode copiar o código e "colar" dentro do **R**. Esta é a maneira tradicional com que rodamos programas no ambiente do **R**.

2) Utilizando o ambiente para implementação de métodos estatísticos para avaliação educacional, em desenvolvimento no programa de Mestrado em Ciência da Computação no INE-UFSC. Esse ambiente é um software chamado *EstatR.exe* que permite a integração entre vários programas criados e o ambiente **R**.

6.5. *EstatR.exe*

O software *EstatR.exe* possui um único ambiente que permite a integração, por exemplo, de todos os programas criados e citados nesta dissertação. O programa *EstatR.exe* chama um programa escrito em **R** e roda-o utilizando o **R**. Por isso, deve-se ter uma versão do **R** instalada no computador. Caso não a tenha, a versão mais atualizada do **R** pode ser encontrada no endereço **http://cran.r-project.org** . Detalhes sobre a instalação e utilização deste software podem ser encontrados no Apêndice.

6.6. Conjunto de dados

Para darmos um exemplo do funcionamento das implementações, utilizamos dois conjuntos de dados correspondentes a provas de Língua Portuguesa aplicadas pela Secretaria de Estado da Educação de São Paulo – SEE/SP que implantou, em 1996, o Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo – SARESP. Uma das provas, composta por 30 itens, foi aplicada em abril de 1997 a alunos da 4ª série do Ensino Fundamental (prova $\mathbf{Y}$). A outra, composta de 32 itens, foi aplicada em novembro de 1997 a alunos de 3ª. série (prova $\mathbf{X}$). As duas provas tinham 21 itens comuns. Todos os 41 itens utilizados eram itens de 3ª série. Maiores detalhes sobre essas provas e métodos de aplicação da mesma podem ser vistos em Andrade et al (2000).

6.7. Implementações desenvolvidas

Para a criação dos programas, considerou-se que o usuário tenha em mãos um ou dois testes que deseja analisar. Chamemos esses testes de **X** e **Y**. Dependendo do tipo de análise a ser realizada, o usuário poderá ter em mãos apenas um dos testes. No caso de uma equalização entre testes, é necessário ter-se os dois. Em todas as equalizações desse tipo, consideramos que queremos transformar a métrica do teste **X** para a métrica do teste **Y**. Ao todo, foram criados 13 programas, sendo alguns deles similares, visto que se distinguem apenas pelo fato de carregarem um arquivo contendo um teste **X**, que se deve chamar *entradax.txt*, ou um arquivo contendo um teste **Y**, chamado *entraday.txt*. As configurações dos arquivos de entrada necessários para cada um dos programas encontram-se descritas no Apêndice.

6.7.1. Programas – Análise Clássica de Itens

Foram desenvolvidos dois programas que realizam uma análise clássica dos itens para cada um dos testes **X** e **Y**, respectivamente. Considera-se que os itens provêm de um único grupo para cada prova realizada. Essa análise prévia é muito importante no sentido de detectar itens que não são bons, conforme discutimos no Capítulo 3. O uso de itens ruins em uma das análises posteriores, como calibração ou equalização (na TC e na TRI) pode acarretar em resultados distorcidos por tais itens. A idéia é que, a partir da análise clássica dos itens, o usuário tenha condições de identificar itens com problemas e tenha a opção de excluí-los das demais análises. Cada um destes programas gera, ao final, dois arquivos, um deles no formato texto (*.txt*) e outro no formato imagem (*.jpg*). Como exemplo, vamos analisar as saídas obtidas no arquivo *saída-fase1-x.txt* a partir da análise do arquivo *entradax.txt*, conforme ilustra a Figura 6.1.

An. de itens Fase 1 Teste X

ITEM:	29					
Prop. corr.:	0.28	Alter.	Prop.	GI	GS	Pto.biss.	Gabarito
Ind. discrim.:	0.26	A	0.35	0.36	0.32	-0.03	
Pto. biss.:	0.28	B	0.12	0.19	0.04	-0.19	
		C	0.28	0.17	0.43	0.28	*
		D	0.20	0.18	0.19	-0.02	
		Outros	0.05	0.10	0.02	-0.20	
ITEM:	23					
Prop. corr.:	0.70	Alter.	Prop.	GI	GS	Pto.biss.	Gabarito
Ind. discrim.:	0.41	A	0.08	0.13	0.02	-0.16	
Pto. biss.:	0.36	B	0.15	0.20	0.09	-0.13	
		C	0.70	0.48	0.89	0.36	*
		D	0.03	0.09	0.01	-0.20	
		Outros	0.04	0.10	0.01	-0.20	
ITEM:	31					
Prop. corr.:	0.87	Alter.	Prop.	GI	GS	Pto.biss.	Gabarito
Ind. discrim.:	0.27	A	0.01	0.03	0.00	-0.12	
Pto. biss.:	0.35	B	0.11	0.26	0.03	-0.32	
		C	0.87	0.70	0.97	0.35	*
		D	0.00	0.00	0.00	NA	
		Outros	0.01	0.02	0.01	-0.04	

Figura 6.1: Exemplo do um arquivo *saída-fase1-x.txt*.

Assim, por exemplo, o item 29 teve uma proporção de acertos de 0,28. O índice de discriminação desse item foi 0,26 e o valor da correlação ponto bisserial do item, correspondente a alternativa correta, foi de 0,28. Observamos também que a alternativa que "recebeu" mais respostas foi a A (0,35), enquanto que a alternativa correta C (indicada com um asterisco na coluna Gabarito) teve uma proporção de acerto igual a 0,28. A alternativa correta (C) teve um acerto de 17% no grupo dos 27% alunos com escores mais baixos (GI – Grupo Inferior) e de 43% no grupo dos alunos com melhor desempenho (GS – Grupo Superior). Podemos notar, também que as alternativas incorretas possuem correlação ponto bisserial negativa, enquanto que a alternativa correta possui uma correlação positiva, indicando que o item apresentou um bom desempenho. Observe que na coluna Pto.biss. do item 31, aparece NA na alternativa D. Isso indica que essa alternativa não foi respondida por nenhum dos indivíduos (note que a proporção de acertos foi zero).

Ao final do arquivo são apresentadas algumas estatísticas do teste, conforme ilustra a Figura 6.2:

```
Estatisticas
  do teste
============
N.itens         32
N.descartados   0
N.individuos    640
escore medio    18.161
desvio-padrao   5.593
maximo          32.000
minimo          3.000
mediana         18.000
ALFA            0.821
EPM             2.367
num. indiv. GI  186
escore max. GI  14
num. indiv. GS  200
escore min. GS  22
```

Figura 6.2: Estatísticas do teste X em *saída-fase1-x.txt*.

Assim, observamos que a prova possui 32 itens dos quais nenhum foi descartado na análise. A prova foi respondida por 640 alunos que obtiveram um escore médio de 18,161 acertos com desvio padrão de 5,593 pontos. O maior escore obtido foi 32 e o menor foi 3. A mediana dos escores ficou em 18. O coeficiente alfa de Cronbach do teste foi 0,821, indicando que o teste é bom com um erro padrão igual a 2,367. No Grupo Inferior, tivemos 186 indivíduos (29%) e o maior escore nesse grupo foi de 14. Já no Grupo Superior, tivemos 200 (31%) indivíduos e o menor escore nesse grupo foi igual a 22. Na saída também é gerado um arquivo de imagem chamado *saída-fase1-hist-x.jpg* (*saída-fase1-hist-y.jpg*), onde se encontra um histograma de freqüências dos escores totais do teste analisado, conforme ilustra a Figura 6.3. Nela, podemos observar uma concentração grande de escores intermediários, enquanto há poucos alunos com escores muito baixos ou muito altos, se assemelhando com uma distribuição normal dos escores.

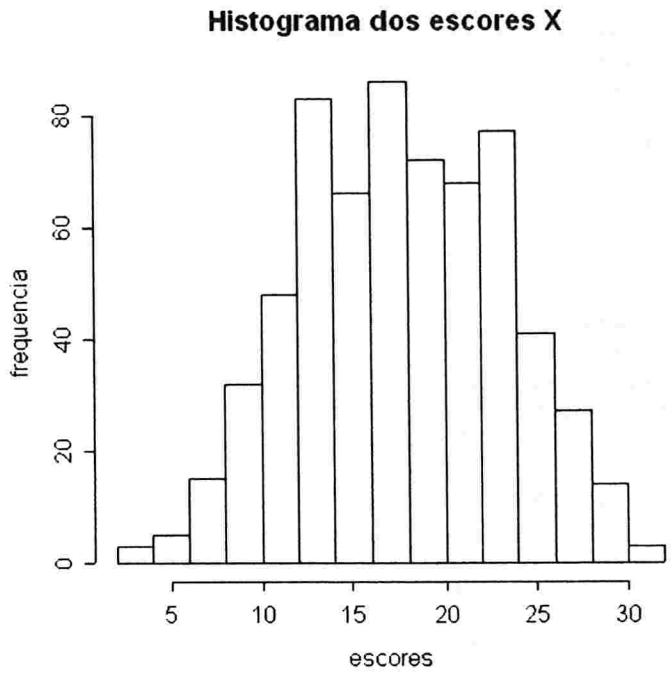


Figura 6.3: Exemplo de um arquivo *saída-fase1-hist-x.jpg*

6.7.2. Programa de equalização na TC – método de Tucker

Este programa realiza uma equalização linear para grupos não equivalentes com teste âncora através do método de Tucker. O algoritmo deste programa consiste basicamente na aplicação das fórmulas apresentadas no Quadro-resumo 1 do Capítulo 4. Notemos que, para este método, o resultado da equalização será o mesmo independente se o teste âncora é interno ou externo. Na saída, são gerados dois arquivos: um de texto chamado *saída-fase2-Tucker.txt* e outro de imagem chamado *saída-fase2-graf-Tucker.jpg*. Podemos observar o início do primeiro arquivo obtido a partir da análise dos conjuntos de dados dos testes $\mathbf{X}$ e $\mathbf{Y}$ (dados do SARESP) representado na Figura 6.4.

```
Metodo de Tucker
================
Pop.1 = teste X
Pop.2 = teste Y
Pesos pop. sintet.
------------------
w1:                  1
w2:                  0

Numero de itens
---------------
Teste X (nao comum): 11
Teste Y (nao comum): 9
Teste V (comum):     21

Numero respondentes
-------------------
Teste X:             640
Teste Y:             1989

Grupo + Teste        Media               Desvio-padrao
-------------        -----               -------------
G1 - teste X         8.111               1.955
G1 - teste V         10.050              4.178

G2 - teste Y         6.253               1.861
G2 - teste V         11.058              4.163

Estatisticas
------------
gama1:               0.287
gama2:               0.255
media pop. sint. X:  8.111
media pop. sint. Y:  5.995
desvio-padrao X:     1.955
desvio-padrao Y:     1.863
```

Figura 6.4: Exemplo de um arquivo *saída-fase2-Tucker.txt.*

Podemos observar que, na análise feita, foram atribuídos os seguintes pesos para a população sintética: w_1=1 e w_2=0. O número de itens não comuns do teste **X** é igual a 11; o número de itens não comuns do teste **Y** é 9; o número de itens comuns em ambos testes (**V**) é igual a 21.

O número de respondentes do teste **X** foi de 640, enquanto que o teste **Y** foi feito por 1989 indivíduos.

Para o grupo que realizou o teste **X** – itens não comuns, a média de acertos foi de 8,111 com desvio de 1,955. Para o grupo que realizou o teste **X** – itens comuns, a média de acertos foi de 10,050 com desvio de 4,178. Para o grupo que realizou o teste **Y** – itens não comuns, a média de acertos foi de 6,253 com desvio de 1,861. Para o grupo que realizou o teste **Y** – itens comuns, a média de acertos foi de 11,058 com desvio de 4,163. Conforme vimos no Capítulo 4, algumas estatísticas necessárias para a realização da equalização são apresentadas: γ_1=0,287, γ_2=0,255, a média da população sintética que realizou o teste **X** foi 8,111 e do teste **Y** foi de 5,995 com desvio-padrão de 1,955 e 1,863 respectivamente. Ainda no mesmo arquivo, temos a continuação da saída gerada que pode ser vista na Figura 6.5.

```
Modelo de equaliz.:  ly(x) = Ax + B
-------------------
A (inclinacao):      0.953
B (intercepto):      -1.737

Escores equalizados  - Teste Interno
-------------------
escore x             lY(x)
0                     0.000
1                     0.000
2                     0.170
3                     1.123
4                     2.076
5                     3.030
6                     3.983
7                     4.936
.                        .
.                        .
.                        .
29                   25.910
30                   26.863
31                   27.816
32                   28.770
```

Figura 6.5: Continuação dos resultados de um exemplo de *saída-fase2-Tucker.txt.*

Como podemos observar, é apresentada a equação de equalização que transforma os escores **x** do teste **X** na escala de escores do teste **Y** dada por

$$l_Y(x) = 0{,}953.x - 1{,}737.$$

A tabela apresentada na Figura 6.5 corresponde aos valores dos possíveis escores **x** do teste **X** transformados para a escala do teste **Y** utilizando a equação apresentada. Esses resultados são sempre considerando a utilização de um teste âncora interno. Os valores foram arredondados para 3 casas decimais. Um possível problema é que o método de Tucker realiza uma equalização linear e, dependendo dos valores de A e B encontrados podemos obter valores de escores que seriam impossíveis de serem obtidos na prática. No exemplo apresentado, o número máximo de itens no teste Y que um aluno pode acertar é 30 (9+21) e o número mínimo é 0. Porém, ao utilizar a equação obtida, podemos obter valores inferiores a 0 ou superiores a 30. Para solucionar esse tipo de problema, uma sugestão apresentada por Kolen & Brennan (1995) é truncar esses escores transformando-os em valores possíveis. Assim, observamos que os valores de $l_Y(x)$ para x igual a 1, 2 e 3 foram iguais a 0. Na verdade, os valores obtidos com a equalização eram negativos, porém, como isso é impossível de ocorrer na prática, transforma-se esses valores para 0, que é o menor escore possível de se obter. Algo similar ocorreria se fossem calculados, através da equação de equalização, escores maiores que 30: tais escores seriam truncados em 30. Em algumas simulações, foram verificados problemas nos cálculos das variâncias das populações sintéticas. Nesses casos, no arquivo de saída aparecem, ao invés de valores numéricos, as letras **NA**. Tais problemas podem ser ocasionados devido a:

- pouco número de itens / indivíduos;
- o conjunto de dados não satisfaz às suposições do método;
- nenhum indivíduo ter assinalado determinada alternativa.

Nesse caso, recomenda-se o uso de outro método para se tentar obter a equalização. Para ajudar o usuário a verificar a suposição da existência de uma regressão linear de X em V e de Y em V, é gerado um arquivo chamado *saída-fase2-graf-Tucker.jpg* que contém 4 gráficos: dois de regressão através do método dos mínimos quadrados e dois com os respectivos resíduos, conforme ilustra a Figura 6.6.

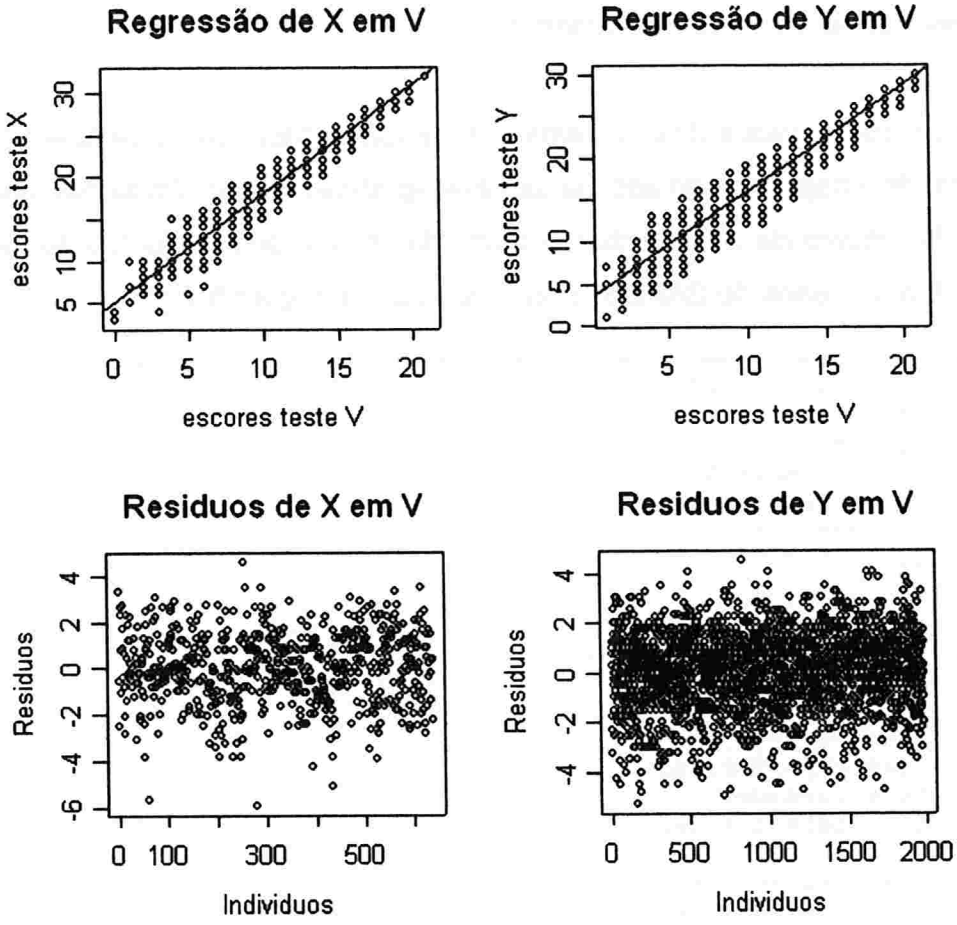


Figura 6.6: Exemplo de um arquivo *saída-fase2-graf.Tucker.txt.*

Cabe ao usuário dizer se as suposições estão adequadas para se utilizar este método de equalização. Pelos gráficos, podemos notar que há evidências de que os dados possam ser ajustados através de uma regressão linear, visto que os valores dos escores se aproximam de uma reta crescente e a maior parte dos resíduos está compreendida no intervalo de -2 a 2.

6.7.3. Programa de equalização na TC – método de Levine

Este programa realiza uma equalização linear para grupos não equivalentes com teste âncora através do método de Levine. O algoritmo deste programa consiste

basicamente na aplicação das fórmulas apresentadas no Quadro-resumo 2 do Capítulo 4. Notemos que, para este método, o resultado da equalização depende se o teste âncora é interno ou externo.

Na saída, são gerados dois arquivos: um de texto chamado *saída-fase2-Levine.txt* e outro de imagem chamado *saída-fase2-graf-Levine.jpg*. Podemos observar o início do arquivo de saída obtido a partir da análise dos conjuntos de dados dos testes **X** e **Y** (dados do SARESP) representado na Figura 6.7.

```
Metodo de Levine
================
Pop.1 = teste X
Pop.2 = teste Y
Pesos pop. sintet.
------------------
w1:                  1
w2:                  0

Teste ancora:        Interno

Numero de itens
---------------
Itens consider. X:   32
Itens consider. Y:   30
Num. itens comuns:   21

Numero respondentes
-------------------
Teste X:             640
Teste Y:             1989

Estatisticas
------------
gama1:               0.764
gama2:               0.783
media pop. sint. X:  8.111
media pop. sint. Y:  5.464
desvio-padrao X:     1.955
desvio-padrao Y:     1.882
```

Figura 6.7: Exemplo do arquivo *saída-fase2-Levine.txt*.

Podemos observar que no início desse arquivo é apresentado o nome do método utilizado (Levine) bem como uma indicação de que a população 1 se refere aos indivíduos que realizaram o teste **X** e a população 2 se refere aos que fizeram o

teste **Y**. Mais abaixo, são apresentados os valores dos pesos w_1 e w_2 da população sintética e o tipo de teste âncora, que, em nosso exemplo, é interno. São informados o número de itens totais e itens considerados do teste **X** e **Y**, que no caso, são 32 e 30, respectivamente, além do número de itens comuns aos dois testes (21). Conforme vimos no Capítulo 4, algumas estatísticas necessárias para a realização da equalização são apresentadas: γ_1=0,764, γ_2=0,783, a média da população sintética que realizou o teste **X** foi 8,111 e do teste **Y** foi de 5,464 com desvio-padrão de 1,955 e 1,882 respectivamente. Ainda no mesmo arquivo, temos a continuação da saída gerada que pode ser observada na Figura 6.8.

```
Modelo de equaliz.:  ly(x) = Ax + B
-----------------
A (inclinacao):      0.963
B (intercepto):      -2.346

Escores equalizados
-------------------
escore x             lY(x)
0                     0.000
1                     0.000
2                     0.000
3                     0.543
4                     1.506
5                     2.468
6                     3.431
.                       .
.                       .
.                       .
29                   25.575
30                   26.538
31                   27.501
32                   28.464
```

Figura 6.8: Continuação dos resultados de um exemplo de *saída-fase2-Levine.txt.*

Como podemos observar, é apresentada a equação de equalização que transforma os escores **x** do teste **X** na escala de escores do teste **Y** dada por

$$l_Y(x) = 0{,}963.x - 2{,}346.$$

A última tabela corresponde aos valores dos possíveis escores **x** do teste **X** transformados para a escala do teste **Y** utilizando a equação apresentada.

Possíveis problemas nas extremidades foram solucionados conforme descritos em 6.7.2. Assim, por exemplo, temos que os escores equalizados para x igual a 0, 1, 2 e 3 foram transformados em 0 pelo fato de serem negativos, e escores de x iguais a 31 e 32 foram transformados para 30 pelo fato de seus valores, utilizando a equação de equalização, serem superiores a 30 (número de itens do teste Y).

Em algumas simulações, foram verificados problemas nos cálculos das variâncias das populações sintéticas. Nesses casos, no arquivo de saída aparecem, ao invés de valores numéricos as letras **NA**, conforme já foi mencionado em 6.7.2 Nesse caso, recomenda-se o uso de outro método, como o equipercentil ou TRI para se tentar obter a equalização. É gerado um arquivo chamado *saída-fase2-graf-Levine.jpg* que contém os mesmos gráficos mostrados na Figura 6.6 pelo fato de estarmos utilizando os mesmos conjuntos de dados em ambos exemplos.

6.7.4. Programa de equalização na TC – método equipercentil

Este programa realiza uma equalização equipercentil para grupos não equivalentes com teste âncora através do método de estimação de freqüências apresentado no Capítulo 4. O algoritmo utilizado neste programa para gerar a equalização é uma generalização do exemplo apresentado na Seção 4.5.2. Os resultados são apresentados no arquivo de saída chamado *saída-fase2-Equipercentil.txt*, e parte dele é apresentado na Figura 6.9.

```
Metodo Equipercentil
====================

Teste X
=======
num.individuos       640
num.itens comuns     21
num.itens nao comuns 11

escore (x)           f(x)                F(x)                P(x) (%)
0                    0.00000             0.00000              0.0000
1                    0.00000             0.00000              0.0000
2                    0.00000             0.00000              0.0000
3                    0.00156             0.00156              0.0781
4                    0.00313             0.00469              0.3125
5                    0.00156             0.00625              0.5469
.                       .                   .                   .
.                       .                   .                   .
.                       .                   .                   .
30                   0.00781             0.99531             99.1406
31                   0.00156             0.99688             99.6094
32                   0.00313             1.00000             99.8438
```

Figura 6.9: Exemplo de parte do arquivo *saída-fase2-Equipercentil.txt*.

Podemos observar que, na saída, são mostradas, inicialmente, informações a respeito do teste **X**: 640 indivíduos realizaram a prova que continha 21 itens comuns e 11 itens não comuns, totalizando 32 itens. Mais abaixo, podemos ver uma tabela onde são apresentados os escores da prova **X** e, ao lado, a proporção de indivíduos que obteve cada escore (f(x)), a distribuição acumulada desses escores (F(x)) e o *rank* percentil correspondente a cada um dos escores (P(x)) dado em porcentagem. Assim, por exemplo, o escore x=5 foi obtido por 0,00156 dos alunos; obtiveram escore menor ou igual a 5 uma proporção de 0,00625 alunos e o *rank* percentil desse escore vale 0,5469. Uma tabela similar a essa é apresentada contendo as informações referentes ao teste **Y**. Mais no final do arquivo, são mostradas informações se o teste âncora é interno ou externo e são apresentados os escores **x** e seus respectivos valores equalizados para a escala do teste **Y**, conforme podemos ver na Figura 6.10.

```
Equalizacao
===========
Teste ancora:        Interno

escore (x)           ey(x)
0                     0.000
1                     0.000
2                     0.000
3                     2.054
4                     4.304
5                     4.931
6                     5.676
.                       .
.                       .
.                       .
30                   28.409
31                   29.103
32                   29.491

Momentos
========
escore                y                    x                   ey(x)
media                17.313               18.161              17.311
desvio-padrao         5.443                5.589               5.435
assimetria           -0.103               -0.012              -0.103
curtose               2.344                2.376               2.338
```

Figura 6.10: Exemplo do final do arquivo *saída-fase2-Equipercentil.txt.*

Mais abaixo, pode-se ver os resultados dos quatro primeiros momentos (média, desvio-padrão, assimetria e curtose) para os escores do teste **X**, do teste **Y** e da equalização obtida. Espera-se que os momentos de $e_Y(x)$ sejam próximos aos momentos de y, fato que pode ser constatado no exemplo da Figura 6.10.

6.7.5. Programa de equalização na TC – método de Braun-Holland

Este programa realiza uma equalização para grupos não equivalentes com teste âncora através do método linear de Braun-Holland apresentado no Capítulo 4.

O algoritmo deste programa consiste numa generalização do exemplo apresentado em 4.6. Notemos que, para este método, o resultado da equalização depende se o teste âncora é interno ou externo. Na saída, são gerados dois arquivos: um de texto chamado *saída-fase2-BraunHol.txt* e outro de imagem chamado *saída-fase2-graf-BraunHol.jpg*. Podemos observar o início do arquivo de

saída obtido a partir da análise dos conjuntos de dados dos testes **X** e **Y** representado na Figura 6.11.

```
Metodo de
Braun-Holland
=============

Teste ancora:          Interno

Pesos pop. sintet.
w1:                    1
w2:                    0
Itens consider. X:     32
Itens descart. X:      0
Itens consider. Y:     30
Itens descart. Y:      0
Num. itens comuns:     21

Estatisticas
------------
media X:               18.161
media Y:               17.311
desvio-padrao X:       5.589
desvio-padrao Y:       5.443
```

Figura 6.11: Exemplo do arquivo *saída-fase2-BraunHol.txt.*

Após a descrição do método utilizado (Braun-Holland) temos a informação se o teste âncora foi considerado interno ou externo e, a seguir, os pesos atribuídos à população sintética. Depois aparecem informações sobre as quantidades de itens considerados e descartados de cada teste. Mais abaixo, vemos algumas estatísticas do teste: o escore médio do teste X foi 18,161 com desvio padrão de 5,589. Já o teste Y teve média de escores igual a 17,311 com desvio padrão de 5,443. O final desse arquivo corresponde às informações de equalização do teste, conforme vemos na Figura 6.12.

```
Modelo de equaliz.:
-------------------
                      ly(x) = Ax + B
A (inclinacao):       0.974
B (intercepto):       -0.377

Escores equalizados
-------------------
escore x              lY(x)
0                      0.000
1                      0.597
2                      1.571
3                      2.545
4                      3.519
5                      4.493
6                      5.467
7                      6.441
.                         .
.                         .
.                         .
29                    27.868
30                    28.842
31                    29.816
32                    30.000
```

Figura 6.12: Continuação dos resultados de um exemplo de *saída-fase2-BraunHol.txt*.

Como podemos observar, é apresentada a equação de equalização que transforma os escores **x** do teste **X** na escala de escores do teste **Y** dada por

$$l_Y(x) = 0{,}974.x - 0{,}377.$$

A última tabela corresponde aos valores dos possíveis escores **x** do teste **X** transformados para a escala do teste **Y** utilizando a equação apresentada. Possíveis problemas nas extremidades foram solucionados conforme descritos em 6.7.2. Assim, por exemplo, temos que o escore de x igual 32 foi transformado para 30 pelo fato de seus valores, utilizando a equação de equalização, serem superiores a 30 (número de itens do teste Y).

É gerado, também, um arquivo chamado *saída-fase2-graf-BraunHol.jpg* que contém os mesmos gráficos mostrados na Figura 6.6 pelo fato de estarmos utilizando os mesmos conjuntos de dados em ambos exemplos.

6.7.6. Programa de estimação na TRI – estimação dos parâmetros dos itens

Este programa realiza a estimação dos parâmetros dos itens de um teste quando as habilidades são desconhecidas. É utilizado o modelo logístico com 3 parâmetros da TRI e as estimações são feitas através do método de máxima verossimilhança marginal com abordagem de Bock & Aitkin. Este programa utiliza como algoritmo no seu desenvolvimento as idéias de estimação apresentadas em 5.3.2., ou seja, fazem parte do seu desenvolvimento, a utilização do algoritmo EM e, dentro dele, o uso de Newton-Raphson para se obter a maximização dos parâmetros da função de verossimilhança. Ao longo das estimações realizadas utilizando-se vários conjuntos de dados, percebeu-se que, com certa freqüência, ocorrem problemas de estimação, como a não convergência de valores e, em estudos de simulação, problemas com os valores dos parâmetros obtidos. Um dos problemas que podemos apontar é devido ao próprio método, ou seja, ao uso da máxima verossimilhança marginal que não fornece boas estimativas quando temos poucos respondentes ou poucos itens. Algumas desvantagens apontadas em Valle (1999) para este método são:

- é trabalhoso computacionalmente;
- não está definido para itens com acerto total ou erro total;
- necessita de uma distribuição a priori para θ (habilidades dos respondentes);
- apresenta problemas na estimação do parâmetro c em alguns casos; deveria ser usado somente com um número suficientemente grande de respondentes.

Tendo em vista esses problemas, solucionamos o caso de não estar definido para escores perfeitos (um aluno acerta todos os itens) ou escore nulo (um aluno erra todos os itens) da forma apresentada em 5.3.2., ou seja, atribuindo 0,5 na questão mais fácil ao indivíduo que obteve escore nulo ou 0,5 na questão mais difícil ao indivíduo que obteve escore perfeito. A distribuição a priori para θ é obtida através do processo de geração de pontos de quadratura que estão bem descritos em Azevedo (2003) e Andrade et al (2000). De fato observamos problemas nas

estimativas do parâmetro c, ocorrendo algumas vezes, mesmo para um número grande de indivíduos (em torno de 2000). Um outro fato observado foi que as estimativas obtidas podem variar bastante dependendo dos valores dos "chutes iniciais" ou estimativas iniciais obtidas para os parâmetros. As estimativas usuais estão descritas em Andrade et al (2000) e se resumem as seguintes fórmulas:

- parâmetro de discriminação (**a**):

$$\widehat{a_i} = \sqrt{\frac{\rho_i^2}{1-\rho_i^2}}, \tag{6.1}$$

onde ρ_i corresponde ao coeficiente de correlação ponto bisserial do item **i** utilizado na TC.

- parâmetro de dificuldade (**b**):

$$\hat{b}_i = -\frac{\Phi^{-1}(\pi_i)}{\rho_i}, \tag{6.2}$$

com

$$\pi_i = \Phi(-b_i\rho_i),$$

onde:

ρ_i corresponde ao coeficiente de correlação ponto bisserial do item **i** utilizado na TC;

π_i é a proporção verdadeira de respostas corretas do item **i**;

Φ é a função de distribuição associada à N(0,1);

b_i é o parâmetro de dificuldade do item **i**.

- parâmetro de acerto casual (**c**):

$$\hat{c}_i = \frac{1}{m_i}, \tag{6.3}$$

onde m_i é o número de alternativas do item **i**.

Embora essas sejam as estimativas iniciais usuais, os melhores resultados obtidos em simulações utilizaram, para o parâmetro **a**, números aleatórios gerados a partir de uma distribuição log-normal com média 0 e desvio-padrão em torno de 0,3, para o parâmetro **b**, números aleatórios gerados a partir de uma distribuição normal (0,1) e para o parâmetro **c**, a partir de uma beta (6,16) (Matos, 2001). Na implementação, o usuário tem a opção de escolher quais serão os valores das estimativas iniciais dos parâmetros, podendo, inclusive, optar por utilizar uma constante, por ele definida, para os parâmetros **a** ou **b**. Existem 5 possibilidades disponíveis para tais chutes que estão descritos no apêndice A.8. A vantagem da utilização dos "chutes" propostos pela literatura é que para um mesmo conjunto de dados obteremos sempre as mesmas estimativas para os parâmetros, porém, em nosso conjunto de dados, os resultados não ficaram tão bons. Optando-se pelo uso da geração de números aleatórios a desvantagem é que, para um mesmo conjunto de dados, em duas execuções do programa poderemos obter 2 estimativas diferentes, embora, de modo geral, estejam próximas. Esse método gera valores satisfatórios para uma amostra grande (acima de 1500). Observamos, também, que quando o número de respondentes é muito baixo (inferior a 1500), ocorrem mais problemas de estabilidade do processo o que acaba por gerar erros-padrão mais elevados.

A seguir, apresentaremos um exemplo de saída dos dados, utilizando nosso banco de dados do SARESP. Na Figura 6.13 podemos ver o exemplo de saída contendo as estimativas dos parâmetros dos itens (arquivo *saída-fase3-TRI-x.txt*).

```
Calibracao - TRI     Teste X
Metodo               Max. veross.
Modelo               Logistico          (3 parametros)
Ciclos EM:           51
Num. itens:          32
Itens desconsider.:  0
Respondentes:        640

Método p/'chutes':   2

Estimativas
(erro padrao)
item                 discriminacao      dificuldade        acerto casual
29                    1.182              1.070              0.054
                     0.265              0.145              0.050
23                    1.640             -1.094              0.038
                     0.288              0.168              0.099
31                    1.527             -1.235              0.355
                     0.373              0.176              0.110
32                    1.505             -0.873              0.183
                     0.298              0.148              0.093
33                    1.226             -0.539              0.035
                     0.221              0.195              0.095
.                      .                  .                  .
.                      .                  .                  .
.                      .                  .                  .
```

Figura 6.13: Exemplo do arquivo *saída-fase3-TRI-x.txt.*

Nessa análise, utilizamos o método número 2 de "chutes" das estimativas iniciais: nela, foram atribuídos para o parâmetros **a** de todos os itens uma estimativa igual a 1, para o parâmetro **b** as estimativas foram obtidas a partir de (6.2). Podemos observar, na Figura 6.13, que foram analisados 32 itens e o número de respondentes foi igual a 640. O número de ciclos EM foi de 51, ou seja, é o número de ciclos do algoritmo de Newton-Raphson que foram realizados. Mais abaixo aparecem quatro colunas que contém os nomes dos itens e as estimativas para os parâmetros **a** (discriminação), **b** (dificuldade) e **c** (acerto casual), onde a primeira linha indica as estimativas dos parâmetros e a linha seguinte mostra o erro-padrão calculado.

No arquivo *parametrosx.txt* encontramos os valores dos parâmetros **a**, **b** e **c** e os nomes dos itens, nessa ordem, separados por vírgulas. Esse arquivo poderá ser utilizado por outros programas, conforme veremos mais adiante.

6.7.7. Programa de estimação na TRI – estimação das habilidades

Este programa realiza a estimação das habilidades dos respondentes de um teste quando os parâmetros dos itens são conhecidos, conforme vimos em 5.3.3. É utilizado o modelo logístico com 3 parâmetros da TRI e as estimações são feitas através do método de máxima verossimilhança. Maiores informações sobre este método podem ser vistos em Azevedo (2003) e Andrade, et al. (2000). Como já foi mencionado em 6.7.6., o método de máxima verossimilhança pode gerar problemas nas estimações dos parâmetros, como, por exemplo, a não convergência das estimativas. Segundo Andrade, et al. (2000), a estimação das habilidades de um grupo pequeno de indivíduos é mais confiável se forem utilizados itens já calibrados. No desenvolvimento do algoritmo, foi utilizado o processo iterativo de Newton-Raphson. Para isso, é preciso ter estimativas iniciais das habilidades. Segundo Andrade et al. (2000) tais estimativas são dadas por:

$$\hat{\theta}_j = \frac{T_j - m}{s},$$

onde

$\hat{\theta}_j$ representa a estimativa inicial do parâmetro θ para o indivíduo **j**;

T_j é o escore obtido pelo indivíduo **j**;

m é o escore médio;

s é o desvio padrão dos escores.

As soluções para os problemas de estimação quando o indivíduo possui escore nulo ou perfeito podem ser vistas em 5.3.2.

A saída principal do programa está no arquivo *saída-fase3-habil-x.txt* e, parte dela, está representada na Figura 6.14.

```
Aluno 1 2 3
Estimacao de  habilidades   - Teste X
parametros    (conhecidos)
Num. respond.: 640

identificacao Habilidade    Erro-padrao   Num. ciclos

13118031      0.733         2.676         3
13118021      -1.004        2.891         2
13118011      -0.718        2.95          2
13115251      0.152         2.876         2
13115241      -0.942        2.912         2
13111251      0.872         2.601         3
13111241      0.584         2.743         3
13111231      -0.624        2.956         2
    .             .             .             .
    .             .             .             .
    .             .             .             .
```

Figura 6.14: Exemplo do arquivo *saída-fase3-habil-x.txt.*

Nesse arquivo observamos os seguintes elementos: nome do teste analisado (**X** ou **Y**), o número de respondentes do teste e uma tabela que contém, nas suas colunas, a identificação do indivíduo, as habilidades estimadas, o erro-padrão associado à essa estimativa e o número de ciclos realizados de Newton-Raphson até atingir um dos critérios de convergência. Em algumas situações podem aparecer na estimativa e no erro padrão a sigla **NaN** (*Not a Number*) que significa que o programa não conseguiu estimar a habilidade do indivíduo. Novamente, salientamos que problemas de estimação através do método da máxima verossimilhança podem ocorrer. Nesse caso, um método bayesiano de estimação pode ser mais apropriado.

6.7.8. Programa de equalização na TRI – 2 testes

Esta implementação permite equalizar dois testes utilizando o método a posteriori: transforma os parâmetros dos itens e as habilidades dos respondentes (opcional), supondo-se um modelo logístico de 3 parâmetros, da escala do teste **X** para a métrica do teste **Y**. Conforme vimos em 5.4.1., para a realização da equalização, utilizamos uma transformação linear. Para isso, basta calcular os valores dos coeficientes dessa transformação que havíamos chamado de α e β. Tais

coeficientes dependem do método escolhido: média-média ou média-desvio. Esta implementação permite a escolha de um dos dois métodos. O programa sempre gera um arquivo de saída chamado *saída-fase4-param2testes.txt*, que contém informações dos testes e os resultados das equalizações dos parâmetros e um arquivo de imagem chamado *saída-fase4-graf.jpg*, que contém dois gráficos: um dos parâmetros **a** equalizados do teste **Y** *versus* teste **X** e outro dos parâmetros **b**. Opcionalmente, o usuário pode querer que sejam gerados os valores das habilidades equalizadas. Neste caso, é gerado um segundo arquivo chamado *saída-fase4-habil2testes.txt*. A seguir, vemos um exemplo de saída do arquivo *saída-fase4-param2testes.txt* mostrado na Figura 6.15 com os respectivos gráficos gerados em *saída-fase4-graf-A.jpg* e *saída-fase4-graf-B.jpg* nas Figuras 6.16a e 6.16b.

Nesse exemplo, observamos, inicialmente, algumas estatísticas da equalização: são mostrados para o teste **X** e teste **Y** os valores das médias dos parâmetros de discriminação (**a**), dos parâmetros de dificuldade (**b**) e dos desvios-padrão dos parâmetros de dificuldade com relação aos itens comuns. Em seguida, aparecem os valores dos coeficientes angular (**alfa**) e linear (**beta**) responsáveis pela transformação das escalas. No caso, α=1,033 e β=–0,156. Assim, as equações de (5.6) a (5.9) ficam:

$$a^{*}_{Yi} = \frac{1}{1,033} a_{Xi},$$

$$b^{*}_{Yi} = 1,033.b_{Xi} - 0,156,$$

$$c^{*}_{Yi} = c_{Xi},$$

e

$$\theta^{*}_{Yj} = 1,033.\theta_{Xj} - 0,156.$$

```
Parametros          equalizados       - TRI
Modelo com          3 parametros
Metodo:             Media-desvio

Estatisticas -      itens comuns
------------        Teste X           Teste Y
Media (a)           1.326             1.478
Media (b)           0.004             -0.151
Desvio-padrao (b)   0.794             0.820

alfa:               1.033
beta:               -0.156

Itens considerados
------------------
Teste X:            32
Teste Y:            30
Comuns:             21

Equalizacao
-----------

                    discriminacao     dificuldade       acerto casual
Item                    (a)               (b)               (c)

1                   1.614             -1.137            0.033
10                  1.207             -0.801            0.010
11                  1.797             -0.983            0.034
12                  1.327             -1.718            0.094
13                  1.478             -0.630            0.028
14                  1.055             0.917             0.011
15                  1.514             0.812             0.054
.                     .                 .                 .
.                     .                 .                 .
.                     .                 .                 .
```

Figura 6.15: Exemplo de saída do arquivo *saida-fase4-param2testes.txt*.

Com essas equações, podemos transformar os parâmetros dos itens e as habilidades da métrica do teste **X** para a do teste **Y**. Mais abaixo, são mostrados o total de itens considerados de cada um dos testes e o número de itens comuns. Por fim, aparecem os resultados da equalização em ordem alfabética. Assim, por exemplo, temos que o item 1, após equalizado, passou a ter seu parâmetro de discriminação igual a 1,614, o parâmetro de dificuldade igual a -1,137, e o parâmetro de acerto casual ficou igual a 0,033.

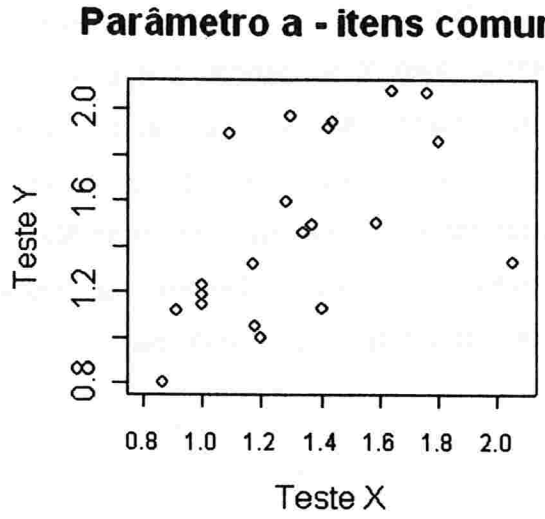


Figura 6.16a: Gráfico do arquivo *saída-fase4-graf-A.jpg*.

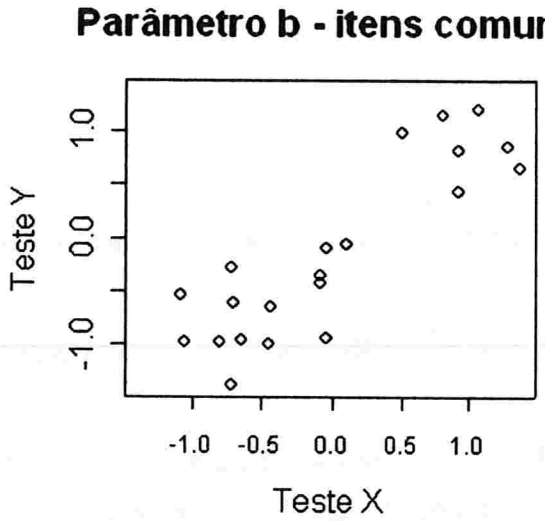


Figura 6.16b: Gráfico do arquivo *saída-fase4-graf-B.jpg*.

Caso o usuário faça opção pela equalização das habilidades, será gerado o arquivo *saída-fase4-habil2testes.txt*. Um exemplo de saída desse arquivo pode ser visto na Figura 6.17. Através desses gráficos, o usuário deve perceber se existe uma relação linear entre os parâmetros do teste **X** e do teste **Y** (o ideal seria que todos os pontos do gráfico estivessem sobre a reta identidade).

Nesse arquivo são exibidos os valores das habilidades (equalizadas) dos indivíduos que fizeram o teste **X** e, depois, dos que fizeram o teste **Y**. Quando algum valor das habilidades não foi estimado, ou seja, quando uma das habilidades dentro do arquivo *habilidadesx.txt* (ou *habilidadesy.txt*) aparecer como **NaN**, o valor equalizado também será exibido como **NaN**, conforme podemos ver na Figura 6.17. A seqüência das habilidades exibidas nos arquivos é a mesma dos indivíduos que aparecem nos arquivos de entrada.

```
Habilidades equalizadas
Metodo:                  Media-desvio

Individuos - teste X
--------------------
0.029
-2.014
-1.678
-0.464
-1.941
0.193
-0.147
-1.568
    .
    .
    .
```

Figura 6.17: Exemplo de saída do arquivo *saída-fase4-habil2testes.txt*.

6.7.9. Programa de equalização na TRI – escala pré-definida

Este programa realiza uma equalização a posteriori dos parâmetros dos itens e/ou habilidades dos respondentes de um dos testes para uma escala pré-definida utilizando um dos métodos de equalização já apresentados: média-média ou média-desvio. Estão disponíveis duas versões do programa: uma que equaliza o teste **X** e outra que equaliza o teste **Y**. Na saída, sempre é gerado um arquivo chamado *saída-fase4-param-escala-x.txt* ou *saída-fase4-param-escala-y.txt* que contém informações dos testes, da escala e os resultados das equalizações dos parâmetros. Opcionalmente, o usuário pode querer que sejam gerados os valores das habilidades equalizadas. Neste caso, é gerado um segundo arquivo chamado

saída-fase4-habil-escala-x.txt (ou *saída-fase4-habil-escala-y.txt*). A seguir, vemos um exemplo de saída do arquivo *saída-fase4-param-escala-x.txt* mostrado na Figura 6.18. Nela, podemos observar os seguintes elementos:

- Método: mostra qual método foi utilizado na análise: média-desvio ou média-média.
- Num.itens consid.: número de itens considerados do teste X e número de itens comuns.
- Escala: fornece informações sobre a escala pré-definida. No caso, o método escolhido foi média-desvio, então aparecem os valores referentes aos itens comuns da escala pré-definida para a média do parâmetro **b** (**Media b**) e para o desvio-padrão do parâmetro **b** (**Desvio b**). Em nosso exemplo, suponhamos que a média seja 0,5 e desvio 1,0 para o parâmetro de dificuldade.

Mais abaixo, temos uma tabela contendo os itens e os parâmetros **a**, **b** e **c** já equalizados. Assim, podemos ver que, por exemplo, o item 29, após ser equalizado para a nova métrica, apresentou para o índice de discriminação o valor de 0,938, para o índice de dificuldade, 1,843 e para o acerto casual, 0,054.

```
Parametros        equalizados       - TRI            Teste X
Modelo com        3 parametros
Metodo:           Media-desvio

Num.itens consid.
-----------------
Teste X:          32
Comuns:           21

Escala
------
Media b:          0.5
Desvio b:         1

Equalização
-----------
Nome do           discriminacao     dificuldade      acerto casual
Item              a                 b                c

29                0.938             1.843            0.054
23                1.302             -0.884           0.038
31                1.212             -1.062           0.355
32                1.194             -0.606           0.183
.                   .                 .                .
.                   .                 .                .
.                   .                 .                .
```

Figura 6.18: Exemplo do arquivo *saída-fase4-param-escala-x.txt*.

No arquivo *saída-fase4-habil-escala-x.txt* encontramos, na parte superior, a informação do método utilizado e, abaixo, as habilidades transformadas para a métrica pré-definida. A seqüência dos valores apresentados corresponde à mesma seqüência em que os indivíduos aparecem no arquivo *entradax.txt*. Um exemplo dessa saída está mostrado na Figura 6.19. Conforme já mencionamos, habilidades que não puderam ser estimadas e aparecem como **NaN** no arquivo *habilidadesx.txt* (ou *habilidadesy.txt*) também aparecerão como **NaN** neste arquivo de equalização.

```
Habilidades equalizadas
              Metodo: Media-desvio

Individuos - teste X
--------------------
2.281
-0.476
-0.022
1.616
-0.378
2.503
  .
  .
  .
```

Figura 6.19: Exemplo do arquivo *saída-fase4-habil-escala-x.txt*.

7. Conclusões e sugestões

Neste trabalho foram desenvolvidos programas escritos para serem utilizados juntamente com o software **R** apresentando, para isso, os conceitos estatísticos necessários para a compreensão da utilização e análise dos resultados fornecidos pelas implementações feitas. Dessa forma, um usuário tem a oportunidade de, a partir de um conjunto de dados, testar e comparar, na prática, métodos de equalização na TC e na TRI. Para detalhes teóricos sugere-se Senno (2006).
Dentre as implementações desenvolvidas, temos análise de itens através da TC, métodos de equalização pela TC e pela TRI, além de processos de estimação pela TRI. Tais implementações, quando rodadas com algum conjunto de dados e comparadas com resultados existentes desse conjunto em livros ou quando comparados com resultados obtidos através de programas comerciais, mostraram-se satisfatórias.

Atualmente existe uma carência de softwares livres e gratuitos que possibilitem estudos relacionados a TRI o que, de certa forma, dificulta a expansão do método. Pensando nisso, damos um primeiro passo no que diz respeito à inclusão digital não comercial da TRI utilizando uma linguagem R na elaboração das implementações e alocando tais programas em um único ambiente chamado EstatR.exe que está em desenvolvimento no programa de Mestrado em Ciência da Computação no INE-UFSC. Muitas implementações devem ser realizadas para que tenhamos um pacote computacional completo relacionados à TRI. Assim, como sugestões para trabalhos futuros, temos implementações relacionadas à outros métodos de calibração, como os bayesianos, estimação conjunta, além de implementações relacionadas a outros modelos como o de resposta nominal e resposta gradual, como está detalhado em Azevedo (2003), modelos longitudinais, que permitem o estudo de um indivíduo ao longo do tempo, como pode ser visto em Tavares e Andrade (2005) e Andrade e Tavares (2005), e para grupos múltiplos, como pode ser visto em Andrade (2001). No campo da

equalização, sugerimos implementações que realizem a equalização juntamente ao processo de estimação, além de equalização para modelos de grupos equivalentes. Um processo que não foi abordado aqui, mas que é muito importante no campo educacional é a questão dos itens âncora, com os quais é possível construir uma escala de habilidades e interpretá-la. Como referências sobre esse assunto temos Valle (1999) e Beaton & Allen (1992).

Para finalizar, devemos ressaltar a importância da integração entre os novos métodos estatísticos (TRI), de recursos computacionais e de especialistas ligados às áreas da estatística e da educação de maneira que os processos de avaliações educacionais possam ser melhorados e divulgados para um número crescente de profissionais ligados à avaliação.

Apêndice – Manual de utilização dos programas

Apresentaremos, neste Apêndice, um manual de referência das implementações desenvolvidas neste trabalho. Neste manual são explicados em detalhes como configurar as implementações de modo a funcionarem corretamente no software **R** e descrevemos, também, as saídas geradas por cada programa, além de explicar como realizar a instalação e utilização do programa *EstatR.exe*.

A.1. *EstatR.exe*

Para a utilização do software *EstatR.exe* é necessário que o **R** já esteja instalado no computador e também fazer a instalação prévia do software RDCOM que pode ser baixado a partir do endereço: **http://cran.r-project.org/contrib/extra/dcom/RSrv135.exe**. Este programa permite que o programa *EstatR.exe* e o **R** passem a se comunicar. A instalação deste software deve ser feita no mesmo diretório em que o **R** foi instalado (geralmente é **C:/Arquivos de programas/R**). Feito isso, basta abrir o programa *EstatR.exe*. Na janela que aparece, no canto superior esquerdo há um botão semelhante a um plug, conforme figura A.1. Deve-se clicar nesse botão de modo que o programa rode as implementações do **R**.

Figura A.1: *Plug* para o EstatR.exe se comunicar com o R.

Após isso, o *EstatR.exe* já está apto a executar as implementações. No menu lateral esquerdo, deve-se escolher a implementação que se deseja carregar. Clicando sobre a implementação desejada com o botão esquerdo do mouse, aparecerá no campo direito uma tela contendo duas guias na parte superior. A primeira contém o nome do programa e algumas instruções de uso. A outra guia é chamada de **Detalhes** e contém o nome da pasta que armazena o programa a ser executado.

Em todas as implementações, o usuário deve fazer algumas configurações iniciais de parâmetros a serem utilizados e do local (diretório) em que se encontram os arquivos de entrada e onde serão salvos os arquivos de saída. Para se fazer tais configurações, o *EstatR.exe* possui um recurso de edição dos arquivos onde estão os programas desenvolvidos. Basta clicar com o botão direito do mouse sobre a implementação que se deseja rodar e selecionar, no menu que se abre, a opção **Configurar**. Uma janela será aberta contendo, no campo da direita o programa que deverá ser editado. Após fazer as alterações necessárias, deve-se clicar no botão **Salvar** que se encontra na parte inferior direita da tela. Feito isso, a janela irá se fechar e a tela principal do programa voltará a ser exibida.

Para rodar a implementação há duas possibilidades: na guia que contém o nome do programa, clique no botão **Executar** no canto inferior direito da tela. Alternativamente, pode-se clicar sobre o nome da simulação com o botão direito do mouse (no menu lateral esquerdo) e selecionar a opção **Executar**. No *EstatR.exe* o console do **R** não é exibido em nenhum momento, fazendo com que o usuário só saiba o que está ocorrendo ao longo do processamento através de janelas de comunicação emitidas pelos programas que estão sendo executados.

A.2. Configuração dos arquivos de entrada

Cada um dos programas desenvolvidos tem como requisito para ser executado de um a seis arquivos de entrada que devem ser criados pelo usuário e atribuídos exatamente os mesmos nomes como citaremos a seguir. Alguns arquivos podem tanto ser criados pelo usuário como podem ser gerados automaticamente a partir de algumas das implementações, conforme citaremos mais adiante. Como já foi dito, pelo fato de se tratar do uso de 2 testes ($\mathbf{X}$ e $\mathbf{Y}$), alguns arquivos, embora tenham nomes diferentes, possuem mesma configuração. A diferença entre eles está apenas nos conteúdos, mas não no modo de configurá-los. Assim, os

exemplos apresentados se referem a um teste **X** (note que os arquivos correspondentes a este teste têm sempre o nome com a terminação x – *aaax.txt,* enquanto que para o teste **Y** os arquivos tem sempre a terminação y – *aaay.txt*). Para o teste **Y**, a configuração é idêntica.

A.2.1. Arquivos *entradax.txt* e *entraday.txt*

Estes arquivos são básicos e necessários (ao menos um deles) em todas as implementações. Eles contêm dados sobre o teste, tais como informações dos itens e respostas dos alunos. As respostas destes arquivos podem ser dados numéricos (0 indicando item errado ou 1 para item correto) ou literais (A,B,C,D,E...). A limitação para o uso de respostas literais é de até 20 letras, ou seja, de A até T. Na figura A.2 apresentamos um exemplo contendo parte de um arquivo literal, baseados em dados do SARESP, chamado *entradax.txt* .

```
ANSWER,C,C,C,C,C,C,C,C,B,B,C,D,C,C,C,D,A,C,C,C,C,A,D,A,B,B,D,D,C,D,B,D
4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4
S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S
Aluno,29,23,31,32,33,34,35,36,24,11,13,14,37,38,25,26,27,39,40,41,17,18,19,20
,21,1,3,4,7,28,15,16
13118031,C,C,C,C,C,C,C,C,B,B,A,B,B,B,C,D,A,C,C,C,B,B,D,A,B,B,D,D,C,A,B,B
13118021,D,C,C,B,B,C,C,C,C,B,D,A,B,B,A,D,C,B,C,B,B,A,A,A,B,B,A,B,C,A,A,C
13118011,A,A,C,C,C,C,C,C,B,D,C,B,B,C,A,B,B,C,C,C,B,A,D,A,A,D,*,B,D,B,C,A
13115251,C,C,C,C,C,C,C,C,A,B,C,D,C,C,C,A,B,C,C,C,B,A,D,C,C,B,D,B,C,C,A,A
13115241,C,B,C,C,C,C,C,C,A,B,A,D,B,C,D,C,B,C,B,C, , , , , ,D,D,C,C,C,A,A
13111251,C,C,C,C,C,C,C,C,B,B,C,D,C,B,B,D,A,B,C,C,C,A,D,A,B,B,D,D,A,C,C,D
13111241,D,C,C,C,B,C,C,C,B,B,C,A,C,C,C,D,C,C,C,C,B,A,D,A,A,B,D,D,C,B,D,D
13111231,A,C,C,C,C,C,C,C,B,D,C,C,C,C,C,A,A,B,C,C,B,D,C,A,D,A,A,C,C,A,D,
13111221,D,C,C,C,C,C,C,C,B,B,D,D,C,B,*,C,A,C,C,*,B,A,D,B,D,B,D,D,A,B,A,A
13111211,D,C,C,C,B,C,C,C,A,B,C,A,B,C,C,D,A,B,C,C,A,A,D,C,D,B,D,A,D,C,A,D
13111201,D,C,C,C,C,C,C,C,B,B,C,A,C,C,C,B,A,C,C,C,B,A,D,A,B,B,D,D,A,C,D,D
```

Figura A.2: Exemplo de um arquivo literal *entradax.txt.*

Nesse exemplo podemos observar como devem ser configurados esses arquivos:

- Na primeira linha, o nome do gabarito, no caso, "ANSWER" e as alternativas corretas de cada item separadas por vírgula (C,C,C,...).

- Na segunda linha deve ser informado o número de alternativas de cada um dos itens, ou seja, se um item contém as alternativas A,B,C e D, o número a ser informado é 4, que é o caso do nosso exemplo. É possível termos um teste que contenha itens com número de alternativas diferentes, por exemplo, se o primeiro item tiver 5 alternativas, o segundo 6, o terceiro 4 e o quarto item contiver 5 alternativas, essa linha deverá ser 5,6,4,5. Os programas reconhecem quantas são as alternativas e emitem análises fazendo essas considerações. Como já dissemos, o número máximo de alternativas é igual a 20. Caso ultrapasse esse valor, os programas não conseguirão realizar as análises e uma janela de erro será emitida ao usuário.
- Na terceira linha, devemos colocar, para cada item, separado por vírgulas e em letras maiúsculas as letras S ou N que representam uma resposta Sim ou Não à questão: "O item deve ser considerado na análise?". Este recurso é bastante interessante visto que permite facilmente incluir ou excluir itens das análises. Assim, um item ruim, pode ser removido sem a necessidade de ser criar um novo arquivo de entrada.
- Na quarta linha colocamos um nome para a identificação da coluna dos respondentes (no caso, "Aluno") e, em seguida, separado por vírgulas, os nomes de cada um dos itens a serem analisados (29, 23, 31,...). Deve-se tomar o cuidado de atribuir, no arquivo *entradax.txt* e *entraday.txt*, os mesmos nomes para os itens comuns (ou itens âncoras). Em contrapartida, deve-se ter cuidado de atribuir nomes diferentes a itens distintos em cada um dos dois arquivos, visto que os programas entendem que itens com mesmo nome deve ser tratados como comuns. Se em um dos arquivos tivermos um item chamado 5 que será considerado na análise ("S" na terceira linha), mas no arquivo correspondente ao outro teste tivermos um item também chamado 5 mas que deve ser desconsiderado da análise ("N" na terceira linha), os programas entenderão o item chamado 5 como sendo um item não comum.
- Da quinta linha em diante, deve ser colocado o nome ou identificação do respondente (no caso da quinta linha, 13118031) seguido das alternativas marcadas por esse indivíduo (C,C,C,C,...). Dois aspectos são muito importantes:

caso o indivíduo tenha deixado a questão em branco ou tenha assinalado mais de uma alternativa, o que invalidaria esse item para esse aluno, deve-se deixar um espaço em branco separado por vírgulas ou colocar um caractere alternativo, como um asterisco (*). Por exemplo, podemos observar que o quinto indivíduo (identificado como 13115241) deixou em branco alguns itens e que o terceiro indivíduo (identificado como 13118011) anulou (veja o asterisco) o item identificado como 3. Outro ponto importante é que os programas criados diferenciam maiúsculas de minúsculas, ou seja, "A" é diferente de "a". Por isso, é necessário que os gabaritos bem como as respostas dos indivíduos estejam todos em maiúsculas ou todos em minúsculas a fim de serem reconhecidos pelas implementações.

Uma outra possibilidade de configuração dos arquivos de entrada é a utilização de "zeros e uns", muito útil quando já se tem o teste corrigido (0 representa resposta incorreta e 1, resposta correta). Neste caso, a configuração do arquivo é a mesma, conforme ilustra a Figura A.3. Para este exemplo, utilizamos um outro conjunto de dados que não corresponde àquele apresentado na Figura A.2. Porém, algumas considerações devem ser feitas:

- na primeira linha, deve contar o nome do gabarito seguido do mesmo, que no caso são os números "1" (que indicam que o item foi acertado);
- na segunda linha, o número de alternativas deve ser igual a 2, visto que estamos considerando apenas certo ou errado;
- da quinta linha em diante, nas respostas dos indivíduos, itens que foram deixados em brancos ou foram anulados pelo respondente devem ser corrigidos como "0" (item incorreto), não devendo aparecer outros caracteres nem espaços em branco.

```
Gaba,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2
S,S,N,S,S,N,S,S,S,S,S,S,N,S,S,S,N,S
Aluno,I1,I2,I3,I4,I5,I6,I7,I8,I9,I10,I11,I12,I13,I14,I15,I16,I17,I18
Aluno1,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Aluno2,1,1,1,0,0,1,0,0,0,0,0,1,0,0,1,0,1,0
Aluno3,1,1,1,0,1,1,0,1,0,0,1,0,0,0,0,0,0,0
Aluno4,1,0,0,1,1,1,1,0,0,0,0,0,1,1,0,0,0,1
Aluno5,1,1,1,0,1,1,0,1,1,1,1,0,1,1,0,0,0,0
Aluno6,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,1,0,0,1
Aluno7,1,0,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1
Aluno8,1,1,1,0,1,1,1,1,1,1,1,0,0,0,1,0,1,1,0
Aluno9,1,1,1,0,1,1,1,1,1,1,1,0,0,0,0,1,0,0,1,0
Aluno10,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0
```

Figura A.3: Exemplo de arquivo numérico *entradax.txt.*

A.2.2. Arquivos *parametrosx.txt* e *parametrosy.txt*

Esses arquivos são necessários apenas para se realizar equalizações na TRI e contém os parâmetros de cada um dos itens considerados em uma análise. Esse arquivo pode ser criado pelo usuário ou pode ser gerado através da implementação que realiza a calibração dos parâmetros dos itens. A configuração deste arquivo consiste em colocar em cada linha os parâmetros dos itens, previamente calibrados em uma escala qualquer, de um modelo logístico com 3 parâmetros, separados por vírgulas e os nomes dos itens. Ou seja, nas três primeiras colunas temos os parâmetros: índice de discriminação (a), índice de dificuldade (b) e acerto casual (c). Na quarta coluna, os nomes dos itens, conforme ilustra a Figura A.4. Este arquivo, quando gerado por outro programa (e não pelo usuário) pode apresentar uma quantidade de casas decimais elevadas. Tais resultados, entretanto, não serão mostrados ao usuário e são usados apenas internamente, em outras implementações.

```
1.19065000300047    ,1.05894362196974    ,0.0577525312294606  ,29
1.63170324934713    ,-1.15064544562281   ,0.0249504056896564  ,23
1.27101386879851    ,-0.350566819487568  ,0.0916349908420657  ,33
1.51335134064236    ,-1.57325792844795   ,0.499569186021201   ,34
1.49029859844648    ,-1.8658995149446    ,0.445946237744564   ,35
1.42016428579667    ,-2.15817539509318   ,0.3630521156034     ,36
0.967888483683066   ,0.110143379298512   ,0.0777770051239651  ,24
1.87384011095949    ,-0.820206927196946  ,0.0277360174494469  ,11
```

Figura A.4: Arquivo de exemplo *parametrosx.txt.*

Uma configuração importante deste arquivo é que, nele, devem constar apenas os parâmetros dos itens considerados na análise. Assim, por exemplo, suponhamos que a configuração de *entradax.txt* fosse feita conforme a Figura A.5:

```
ANSWER,C,C,C,C,C,C,C,C,B,B,C,D,C,C,C,D,A,C,C,C,C,A,D,A,B,B,D,D,C,D,B,D
4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4
S,S,N,N,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S,S
Aluno,29,23,31,32,33,34,35,36,24,11,13,14,37,38,25,26,27,39,40,41,17,18,19,20
,21,1,3,6,7,28,15,16
```

Figura A.5: Suposição de configuração para *entrada.txt*.

No arquivo de entrada apresentado na Figura A.5 temos que o item 31 e 32 não devem ser considerados na análise. Então, nesse caso, na Figura A.4 teríamos que a linha 1 corresponde aos parâmetros do item 29, a linha 2 ao item 23, a linha 3 ao item 33, a linha 4 ao item 34, a linha 5 ao item 35 e assim por diante, ou seja, não devem constar os itens desconsiderados na análise.

A.2.3. Arquivos *habilidadesx.txt* e *habilidadesy.txt*

Estes arquivos possuem os valores das habilidades estimadas para cada um dos indivíduos na mesma métrica dos testes **X** e **Y**, respectivamente. Em cada linha, conforme podemos ver na Figura A.6, devemos colocar os valores das habilidades para cada indivíduo. Tais valores podem ser incluídos pelo usuário ou obtidos através da implementação que realiza a estimação das habilidades pela TRI.

```
 0.732962070
-1.003800439
-0.717748062
 0.313986020
-0.941743301
 0.872487050
 0.583760022
-0.624258819
```

Figura A.6: Exemplo de um arquivo *habilidadesx.txt*.

Uma observação importante é que os valores das habilidades podem ser colocados com números decimais, conforme mostra a Figura A.6 ou colocados em forma de notação científica (7.32962070e-01). Os programas que fazem as leituras dos arquivos de habilidades conseguem reconhecer as duas formas de se entrar com os números.

A.2.4. Arquivos *itenscomunsx.txt* e *itenscomunsy.txt*

Estes arquivos são necessários apenas quando se deseja realizar uma equalização de um dos testes para uma métrica que já está pré-definida. Neste caso, faz-se necessário saber quais itens devem ser tratados como comuns. Assim, estes arquivos contém apenas uma linha com os nomes dos itens comuns separados por vírgulas, conforme ilustra a Figura A.7.

```
29,23,24,11,13,14,25,26,27,17,18,19,20,21,1,3,6,7,28,15,16
```

Figura A.7: Exemplo de um arquivo *itencomunsx.txt* .

A.3. Programa – Análise Clássica de Itens

Nome do programa em *EstatR.exe*:	Fase 1 – Análise de itens – Teste X
Nome do arquivo:	fase1-x.txt
Versão:	Teste X
Arquivos de entrada necessários:	entradax.txt
Arquivos de saída gerados:	saida-fase1-x.txt saida-fase1-hist-x.jpg

Nome do programa em *EstatR.exe*:	Fase 1 – Análise de itens – Teste Y
Nome do arquivo:	fase1-y.txt
Versão:	Teste Y
Arquivos de entrada necessários:	entraday.txt
Arquivos de saída gerados:	saida-fase1-y.txt saida-fase1-hist-y.jpg

Estas implementações realizam uma análise clássica dos itens de um teste X ou de um teste Y. Os programas *fase1-x.txt* e *fase1-y.txt* devem ser configurados antes de se realizar as análises. Um exemplo de configuração é mostrado na Figura A.8:

```
# Configure o local onde estão os arquivos e onde será gravado o
arquivo de saída.
setwd("C:/TRI/exemplo")

# respostas são literais (A,B,C...) ou são numéricas (corrigidas:
1=acerto, 0=erro)?
# Escolha "L" para literais ou "N" para numericas
tiporesposta<-"L"
```

Figura A.8: Exemplo de configuração do arquivo *fase1-x.txt*.

Devemos configurar apenas o que está destacado em negrito na Figura A.8. Ou seja, devemos informar em que diretório se encontram os arquivos de entrada e onde serão salvos os arquivos gerados na saída. Em nosso exemplo, tais arquivos encontram-se no diretório **C:/TRI/exemplo**. Cuidado na digitação do endereço: o programa faz distinção entre maiúsculas e minúsculas. Mais abaixo, devemos configurar se desejamos trabalhar com dados literais (**L**), conforme exemplo da Figura A.2 ou numéricos (**N**), conforme exemplo da Figura A.3.Após executar o programa, são gerados dois arquivos de saída: um arquivo de texto chamado *saída-fase1-x.txt* (ou *saída-fase1-y.txt*) e um arquivo de imagem chamado *saida-fase1-hist-x.jpg* (ou *saida-fase1-hist-y.jpg*).

No arquivo *saída-fase1-x.txt* podemos observar no seu início uma identificação do teste (X ou Y) e, logo a seguir, temos todos os itens presentes no teste e que foram considerados na análise ("S" na terceira linha de *entradax.txt*). Para cada item, podemos observar os seguintes elementos, que já foram discutidos no Capítulo 3:

- ITEM: nome do item analisado
- Prop. corr.: é a proporção de acertos do item, também conhecido como índice de dificuldade do item.
- Ind. discrim.: é o índice de discriminação do item.
- Pto. biss.: é o coeficiente de correlação ponto bisserial.

Dentro de cada item, temos as seguintes colunas:

- Alter.: alternativas de cada item acrescido da opção "Outros" que contém os itens deixados em branco ou com resposta diferente da informada. Por exemplo, se foi informado que um item contém 5 alternativas mas nas respostas de um dos indivíduos apareceu a letra "F", ou estava em branco, ou apareceu um outro caracter (como *), essa resposta é considerada na opção "Outros". No caso de respostas numéricas (do tipo 0 ou 1), na saída, a alternativa "A" corresponde ao "1", ou seja, representa a alternativa correta, enquanto que a alternativa B representa o "0", ou seja, a alternativa errada. Visto que as respostas já estão corrigidas, não espera-se encontrar dados omissos. Por isso, espera-se que, em todos itens analisados, a opção "Outros" apareça com valores iguais a zero em todos os campos.
- Prop.: proporção de acerto da referida alternativa. A soma dos valores desta coluna, dentro do processamento do programa, é igual a 1. Porém, por motivos de arredondamento dos valores, em alguns itens o somatório pode ser diferente de 1.
- GI: proporção de acerto de cada uma das alternativas no grupo inferior, ou seja, no grupo dos 27% de indivíduos que obtiveram os escores mais baixos.

- GS: proporção de acerto de cada uma das alternativas no grupo superior, ou seja, no grupo dos 27% de indivíduos que obtiveram os escores mais altos.
- Pto.biss.: correlação ponto bisserial para cada uma das alternativas.
- Gabarito: o asterisco indica qual alternativa foi indicada como sendo o gabarito do item.

Ao final desse arquivo são apresentadas algumas estatísticas gerais do teste:

- N. itens: número total de itens considerados na análise.
- N. descartados: número de itens descartados da análise.
- N. indivíduos: número de indivíduos considerados na análise.
- escore médio: escore médio dos indivíduos para os itens considerados.
- desvio padrão: desvio padrão amostral dos escores.
- maximo: escore máximo obtido.
- minimo: escore mínimo obtido.
- mediana: mediana dos escores.
- ALFA: coeficiente alfa de fidedignidade do teste
- EPM: erro padrão de medida do teste.
- num. Indiv. GI: número de indivíduos pertencentes ao grupo dos respondentes com os menores escores.
- escore max. GI: maior escore obtido no grupo GI.
- num. Indiv. GS: número de indivíduos pertencentes ao grupo dos respondentes com os maiores escores.
- escore min. GS: menor escore obtido no grupo GS.

Os arquivos de imagem (*saida-fase1-hist-x.jpg* e *saida-fase1-hist-y.jpg*) produzem um histograma dos escores obtidos pelos respondentes. O número de classes desse histograma é automaticamente definido pelo **R**.

A.4. Programa de equalização na TC – método de Tucker

Nome do programa em *EstatR.exe*:	Fase 2 – Equalização Clássica - Tucker
Nome do arquivo:	fase2-Tucker.txt
Versão:	única
Arquivos de entrada necessários:	entradax.txt entraday.txt
Arquivos de saída gerados:	saida-fase2-Tucker.txt saida-fase2-graf-Tucker.jpg

Para rodar este programa, é necessário que o usuário configure, no início do programa (arquivo *fase2-Tucker.txt*), o local (diretório) onde estão os arquivos de entrada e onde serão criados os arquivos de saída que, no caso do exemplo, é **C:/TRI/exemplo**, e configure os valores dos pesos da população sintética (w_1 e w_2). Na realidade, basta alterar o valor de **w1**, pois **w2** é calculado como sendo 1−w1. O usuário também deve definir se deseja trabalhar com dados literais (**L**), conforme exemplo da Figura A.2 ou numéricos (**N**), conforme exemplo da Figura A.3. As alterações possíveis estão destacadas em negrito na Figura A.9.

```
# Configure o local onde estão os arquivos e onde será gravado o
arquivo de saída.
setwd("C:/TRI/exemplo")

# atribui os pesos para as populações sintéticas
# w1 = peso para a população X
# w2 = peso para a população Y

w1<-1
w2<-1-w1

# respostas são literais (A,B,C...) ou são numéricas (corrigidas:
1=acerto, 0=erro)?
# Escolha "L" para literais ou "N" para numericas
tiporesposta<-"L"
```

Figura A.9: Exemplo de configuração do arquivo *fase2-Tucker.txt*.

Este programa tem algumas sub-rotinas de erros implementadas, como por exemplo a que detecta a existência de algum item que ultrapasse 20 alternativas e outra que informa ao usuário caso não sejam encontrados itens comuns nos testes **X** e **Y**. Como já foi explicado anteriormente, o programa consegue detectar os itens comuns a partir de seus nomes (quarta linha dos arquivos *entradax.txt* e *entraday.txt*). Portanto, itens comuns devem necessariamente ter o mesmo nome, para que possam ser identificados pelo programa e, assim, poder realizar a equalização.

Na saída, são gerados dois arquivos: um de texto chamado *saída-fase2-Tucker.txt* e outro de imagem chamado *saída-fase2-graf-Tucker.jpg*. O primeiro arquivo contém os resultados e algumas estatísticas, enquanto que, no segundo, encontramos gráficos de regressão através do método dos mínimos quadrados.

No arquivo de saída *saída-fase2-Tucker.txt*, temos as informações do método utilizado (Tucker), e de que a amostra da população 1 realizou a prova tipo **X** e, da população 2, tipo **Y**. Em seguida, vemos os pesos w_1 e w_2 correspondentes à população sintética e que foram definidos nas configurações iniciais do programa pelo usuário. Após isso, são informados o número de itens não comuns da prova **X**, o número de itens não comuns presentes na prova **Y** e, finalmente, o número de itens comuns aos dois testes, chamado de teste **V**.

A seguir, são mostrados os números de respondentes de cada um dos dois testes. Depois, aparecem algumas estatísticas como média e desvio-padrão correspondentes aos:

– G1 – teste **X**: indivíduos provenientes da população 1 e que realizaram os itens não comuns da prova **X**.

– G1 – teste **V**: indivíduos provenientes da população 1 e que realizaram os itens comuns da prova **X**.

– G2 – teste **Y**: indivíduos provenientes da população 2 e que realizaram os itens não comuns da prova **Y**.

– G2 – teste **V**: indivíduos provenientes da população 2 e que realizaram os itens comuns da prova **Y**.

Mais abaixo, no mesmo arquivo, podemos ver algumas estatísticas do teste, que são obtidos a partir das fórmulas apresentadas no Quadro-resumo 1 do Capitulo 4:

- gama1: valor de γ_1.
- gama2: valor de γ_2.
- media pop. sint. X: média dos escores da população sintética para o teste **X**.
- media pop. sint. Y: média dos escores da população sintética para o teste **Y**.
- desvio-padrao X: desvio padrão dos escores da população sintética para o teste **X**.
- desvio-padrao Y: desvio padrão dos escores da população sintética para o teste **Y**.

Em seguida, são apresentados os resultados da equalização que transforma os escores do teste **X** na escala de escores do teste **Y** dada por $l_Y(x) = Ax + B$:

- A (inclinação): corresponde ao coeficiente angular da reta que ajusta a equalização;
- B (intercepto): corresponde ao coeficiente linear da reta que ajusta a equalização.

Em seguida, é apresentada uma tabela onde na coluna da esquerda aparecem os valores dos possíveis escores do teste **X**, admitindo que o teste âncora seja interno, e, na coluna da direita, aparecem os correspondentes escores ajustados para a escala do teste **Y**. Em alguns casos, ocorrem problemas nos extremos dos escores equalizados, ou seja, podem ser calculados escores que não existam na escala do teste **Y**, como, por exemplo, escores negativos ou escores superiores ao total de itens do teste **Y**. Em razão disso, utilizou-se uma sub-rotina que

transforma escores equalizados negativos em 0 e escores superiores ao máximo possível em valores iguais ao total de itens do teste **Y**. É gerado também um arquivo chamado *saída-fase2-graf-Tucker.jpg* contendo os gráficos de regressão dos escores do teste **X** (não comuns) sobre **V** e do teste **Y** (não comuns) sobre **V**. Abaixo dos mesmos, apresenta-se um gráfico de resíduos para ajudar na verificação da suposição de regressão linear do método de Tucker.

A.5. Programa de equalização na TC – método de Levine

Nome do programa em *EstatR.exe*:	Fase 2 – Equalização Clássica - Levine
Nome do arquivo:	fase2-Levine.txt
Versão:	única
Arquivos de entrada necessários:	entradax.txt entraday.txt
Arquivos de saída gerados:	saida-fase2-Levine.txt saida-fase2-graf-Levine.jpg

Para rodar este programa, é necessário que o usuário configure, no início do programa (arquivo *fase2-Levine.txt*), o local (diretório) onde estão os arquivos de entrada e onde serão criados os arquivos de saída que, no caso do exemplo, é C:/TRI/exemplo, e configure os valores dos pesos da população sintética (w_1 e w_2). Na realidade, basta alterar o valor de **w1**, pois **w2** é calculado como sendo 1-w1. O usuário também deve definir se deseja trabalhar com dados literais (**L**), conforme exemplo da Figura A.2 ou numéricos (**N**), conforme exemplo da Figura A.3. Além disso, é necessário definir se o teste âncora utilizado é interno (**I**) ou externo (**E**). As alterações possíveis estão destacadas em negrito na Figura A.10.

```
# Configure o local onde estão os arquivos e onde será gravado o
arquivo de saída.
setwd("C:/TRI/exemplo")

# atribui os pesos para as populações sintéticas
# w1 = peso para a população X
# w2 = peso para a população Y

w1<-1
w2<-1-w1

# Define se o teste âncora é interno (I) ou externo (E)
testeancora<-"I"

# respostas são literais (A,B,C...) ou são numéricas (corrigidas:
1=acerto, 0=erro)?
# Escolha "L" para literais ou "N" para numericas
tiporesposta<-"L"
```

Figura A.10: Exemplo de configuração do arquivo *fase2-Tucker.txt*.

Algumas sub-rotinas de erro foram implementadas da mesma forma que as descritas em A.2. Na saída, são gerados dois arquivos: um de texto chamado *saída-fase2-Levine.txt* e outro de imagem chamado *saída-fase2-graf-Levine.jpg*. O primeiro arquivo contém os resultados e algumas estatísticas, enquanto que, no segundo, encontramos gráficos de regressão através do método dos mínimos quadrados. No arquivo *saída-fase2-Levine.txt* encontramos os seguintes elementos:

- w1 e w2: pesos das populações sintéticas definidos pelo usuário.
- Teste âncora: pode ser interno ou externo e é definido pelo usuário.
- Itens consider. X: número de itens considerados do teste **X**.
- Itens consider. Y: número de itens considerados do teste **Y**.
- Num. Itens comuns: número de itens comuns aos dois testes.

Em seguida, são mostrados os números de indivíduos que responderam os testes **X** e **Y**. Mais abaixo, no mesmo arquivo, podemos ver algumas estatísticas do teste, que são obtidos a partir das fórmulas apresentadas no Quadro-resumo 2 do Capitulo 4:

- gama1: valor de γ_1.

- gama2: valor de γ_2.

- media pop. sint. X: média dos escores da população sintética para o teste **X**.

- media pop. sint. Y: média dos escores da população sintética para o teste **Y**.

- desvio-padrao X: desvio padrão dos escores da população sintética para o teste **X**.

- desvio-padrao Y: desvio padrão dos escores da população sintética para o teste **Y**.

Em seguida, são apresentados os resultados da equalização que transforma os escores do teste **X** na escala de escores do teste **Y** dada por $l_Y(x) = Ax + B$:

- A (inclinação): corresponde ao coeficiente angular da reta que ajusta a equalização;

- B (intercepto): corresponde ao coeficiente linear da reta que ajusta a equalização.

Em seguida, é apresentada uma tabela onde na coluna da esquerda aparecem os valores dos possíveis escores do teste **X** e, na coluna da direita, aparecem os correspondentes escores ajustados para a escala do teste **Y**. Em alguns casos ocorrem problemas nos extremos dos escores equalizados e a solução dada foi a mesma descrita em A.4. É gerado também um arquivo chamado *saída-fase2-graf-Levine.jpg* contendo os mesmos gráficos de *saída-fase2-graf-Tucker.jpg* descritos em A.4.

A.6. Programa de equalização na TC – método eqüipercentil

Nome do programa em *EstatR.exe*:	Fase 2 – Equalização clássica - Equipercentil
Nome do arquivo:	fase2-Equipercentil.txt
Versão:	única
Arquivos de entrada necessários:	entradax.txt entraday.txt
Arquivos de saída gerados:	saida-fase2-Equipercentil.txt

Para rodar este programa, é necessário que o usuário configure, no início do programa (arquivo *fase2-Equipercentil.txt*), o local (diretório) onde estão os arquivos de entrada e onde serão criados os arquivos de saída que, no caso do exemplo, é **C:/TRI/exemplo**. O usuário também deve definir se deseja trabalhar com dados literais (**L**), conforme exemplo da Figura A.2 ou numéricos (**N**), conforme exemplo da Figura A.3. Além disso, é necessário definir se o teste âncora utilizado é interno (**I**) ou externo (**E**). As alterações possíveis estão destacadas em negrito na Figura A.11.

```
# Configure o local onde estão os arquivos e onde será gravado o
arquivo de saída.
setwd("C:/TRI/exemplo")

# Define se o teste âncora é interno (I) ou externo (E)
testeancora<-"I"

# respostas são literais (A,B,C...) ou são numéricas (corrigidas:
1=acerto, 0=erro)?
# Escolha "L" para literais ou "N" para numericas
tiporesposta<-"L"
```

Figura A.11: Exemplo de configuração do arquivo *fase2-Equipercentil.txt*.

Algumas sub-rotinas de erro foram implementadas da mesma forma que as descritas em A.2. Na saída, é gerado um arquivo de texto chamado *saída-fase2-Equipercentil.txt*, que contém os resultados da equalização. No início do arquivo

encontramos o nome do método utilizado (equipercentil) e, a seguir, duas tabelas contendo as informações sobre os testes **X** e **Y**:

- num. indivíduos: número de alunos que realizaram o teste mencionado.
- num. itens comuns: número de itens comuns a ambos testes.
- num. itens não comuns: número de itens não comuns no teste mencionado.

Nas colunas das tabelas encontramos os seguintes elementos:

- escore (x) ou escore (y): valores dos possíveis escores em cada uma das provas consideradas;
- f(x) ou g(y): proporção de indivíduos que obtiveram o escore da primeira coluna.
- F(x) ou G(y): distribuição acumulada dos acertos de cada escore.
- P(x) (%) ou Q(y) (%): *rank* percentil de cada escore.

Mais adiante, encontramos uma tabela que nos fornece os valores da equalização através do método equipercentil. Antes dela, há indicação se o teste âncora é interno ou externo e, a seguir, são apresentados os possíveis escores do teste **X** e, ao lado, os valores obtidos pela equalização. Em alguns casos ocorrem problemas nos extremos dos escores equalizados e a solução dada foi a mesma descrita em A.4.

No final do arquivo, temos uma tabela que apresenta os 4 primeiros momentos (média, desvio-padrão, assimetria e curtose) para o teste **Y**, teste **X** e para os escores do teste equalizado $e_Y(x)$.

A.7. Programa de equalização na TC – método de Braun-Holland

Nome do programa em *EstatR.exe*:	Fase 2 – Equalização Clássica – Braun-Holland
Nome do arquivo:	fase2-Braun-Holland.txt
Versão:	única
Arquivos de entrada necessários:	entradax.txt entraday.txt
Arquivos de saída gerados:	saida-fase2-BraunHol.txt saída-fase2-graf-BraunHol.jpg

Para rodar este programa, é necessário que o usuário configure, no início do programa (arquivo *fase2-Braun-Holland.txt*), o local (diretório) onde estão os arquivos de entrada e onde serão criados os arquivos de saída que, no caso do exemplo, é **C:/TRI/exemplo**. É necessário definir se o teste âncora utilizado é interno (**I**) ou externo (**E**). Deve ser configurado o valor do peso w_1 da população sintética. O valor de w_2 é automaticamente calculado fazendo-se $1-w_1$. O usuário também deve definir se deseja trabalhar com dados literais (**L**), conforme exemplo da Figura A.2 ou numéricos (**N**), conforme exemplo da Figura A.3. As alterações possíveis estão destacadas em negrito na Figura A.11.

```
# Configure o local onde estão os arquivos e onde será gravado o
arquivo de saída.
setwd("C:/TRI/exemplo")

# Define se o teste âncora é interno (I) ou externo (E)
testeancora<-"I"

# atribui os pesos para as populações sintéticas
# w1 = peso para a população X
# w2 = peso para a população Y
w1<-1
w2<-1-w1

# respostas são literais (A,B,C...) ou são numéricas (corrigidas:
1=acerto, 0=erro)?
# Escolha "L" para literais ou "N" para numericas
tiporesposta<-"L"
```

Figura A.11: Exemplo de configuração do arquivo *fase2-Braun-Holland.txt.*

Algumas sub-rotinas de erro foram implementadas da mesma forma que as descritas em A.2. Na saída, é gerado um arquivo de texto chamado *fase2-Braun-Holland.txt*, que contém os resultados da equalização e um arquivo gráfico denominado *saída-fase2-graf-BraunHol.jpg*. No início do arquivo *fase2-Braun-Holland.txt* encontramos o nome do método utilizado (Braun-Holland) e, a seguir, algumas informações sobre os testes:

- Teste âncora: informa qual tipo de teste utilizado na análise, que pode ser interno ou externo e é definido pelo usuário.
- w1 e w2: pesos das populações sintéticas definidos pelo usuário.
- Itens consider. X: número de itens considerados do teste **X**.
- Itens descart. X: número de itens descartados do teste **X**.
- Itens consider. Y: número de itens considerados do teste **Y**.
- Itens descart. Y: número de itens descartados do teste **Y**.
- Num. Itens comuns: número de itens comuns aos dois testes.

Em seguida, aparecem algumas estatísticas:

- media X: média dos escores do teste **X**.
- media Y: média dos escores do teste **Y**.
- desvio-padrao X: desvio-padrão dos escores do teste **X**.
- desvio-padrao Y: desvio-padrão dos escores do teste **Y**.

Em seguida, são apresentados os resultados da equalização que transforma os escores do teste **X** na escala de escores do teste **Y** dada por $l_Y(x) = Ax + B$:

- A (inclinação): corresponde ao coeficiente angular da reta que ajusta a equalização;
- B (intercepto): corresponde ao coeficiente linear da reta que ajusta a equalização.

E, finalmente, é mostrada uma tabela contendo os possíveis escores do teste X e os correspondentes escores equalizados para a escala do teste Y utilizando a

equação de equalização apresentada. Em alguns casos ocorrem problemas nos extremos dos escores equalizados e a solução dada foi a mesma descrita em A.4. É gerado também um arquivo chamado *saída-fase2-graf-BraunHol.jpg* contendo os mesmos gráficos de *saída-fase2-graf-Tucker.jpg* descritos em A.4.

A.8. Programa de estimação na TRI – estimação dos parâmetros dos itens

Nome do programa em *EstatR.exe*:	Fase 3 – Estimação dos parâmetros – Teste X
Nome do arquivo:	fase3-estima parametros-x.txt
Versão:	Teste X
Arquivos de entrada necessários:	entradax.txt
Arquivos de saída gerados:	saida-fase3-TRI-x.txt parametrosx.txt

Nome do programa em *EstatR.exe*:	Fase 3 – Estimação dos parâmetros – Teste Y
Nome do arquivo:	fase3-estima parametros-y.txt
Versão:	Teste Y
Arquivos de entrada necessários:	entraday.txt
Arquivos de saída gerados:	saida-fase3-TRI-y.txt parametrosy.txt

Esta implementação realiza a calibração dos parâmetros dos itens de um teste quando as habilidades dos respondentes são desconhecidas. Utiliza, para isso, o método de máxima verossimilhança marginal. Para rodar este programa, é necessário que o usuário realize diversas configurações, as quais estão divididas em duas partes: configurações básicas e configurações avançadas, que muitas vezes não precisam ser alteradas, podendo ser consideradas *default* do programa. Nas figuras A.12 e A.13, vemos quais são as configurações básicas que devem ser alteradas pelo usuário (em negrito). Lembramos que as configurações para os

arquivos *fase3-estima parametros-x.txt* e *fase3-estima parametros-y.txt* são idênticas.

```
# Configure o local onde estão os arquivos e onde será gravado o
arquivo de saída.
setwd("C:/TRI/exemplo")

# respostas são literais (A,B,C...) ou estão numéricas (corrigidas:
1=acerto, 0=erro)?
# Escolha "L" para literais ou "N" para numericas
tiporesposta<-"L"
```

Figura A.12: Configurações básicas do arquivo *fase3-estima parametros-x.txt*.

Na Figura A.12 vemos que o usuário, assim como foi feito em alguns programas anteriores, deve configurar o local onde estão os arquivos de entrada e onde serão salvos os arquivos de saída. Deve, também, informar se os dados do teste contidos nos arquivos de entrada estão no formato de dados literais (**L**), conforme exemplo da Figura A.2 ou numéricos (**N**), conforme exemplo da Figura A.3.

Na Figura A.13 temos a continuação das configurações básicas. O usuário deve informar como deseja que sejam os "chutes iniciais" ou estimativas iniciais para os parâmetros dos itens. São dadas 5 opções ao usuário que deve colocar o número correspondente à opção no campo "chuteinicial". A definição dessas estimativas iniciais é um passo muito importante na estimação pela TRI e deve ser escolhido com cautela. Percebemos que, no método de máxima verossimilhança marginal, os valores dos chutes iniciais podem gerar alterações bastante significativas nas estimativas. As opções disponíveis para escolha das estimativas iniciais dos parâmetros de discriminação (**a**), de dificuldade (**b**) e acerto casual (**c**) são:

```
### Define os chutes iniciais ('chuteinicial') de acordo com os
números a seguir (de 1 a 4):
# (Em todas as possibilidades, c é obtido a partir do número de
alternativas do item).
#
# 1 - chute padrão recomendado na literatura:
#     a = obtido a partir da correlação ponto bisserial
#     b = obtido a partir da curva normal + ponto bisserial
#
# 2 - a = todos os valores de a recebem uma constante (defina o valor
de 'chute.a') (sugestão: 1)
#     b = obtido a partir da curva normal + ponto bisserial
#
# 3 - a = todos os valores de a recebem uma constante (defina o valor
de 'chute.a') (sugestão: 1)
#     b = todos os valores de b recebem uma constante (defina o valor
de 'chute.b') (sugestão: 0)
#
# 4 - a = valores gerados de uma log-normal com média 'lognor.media'
e desvio-padrão
#         'lognor.dp' (sugestão: média 0 e variância 0.3)
#     b = obtido a partir da curva normal (0,1)
#
# 5 - a = valores gerados de uma log-normal com média 'lognor.media'
e desvio-padrão
#         'lognor.dp' (sugestão: média 0 e variância 0.3)
#     b = gerados a partir da curva normal N(0,1)
#     c = obtido a partir de uma beta (defina parametros c1 e c2)
(sugestão: c1=6 e c2=16)

chuteinicial<-2

chute.a<-1
chute.b<-0

lognor.media<-0
lognor.dp<-0.3

c1<-6
c2<-16
```

Figura A.13: Continuação das configurações básicas do arquivo *fase3-estima parametros-x.txt.*

1 – corresponde às estimativas iniciais recomendadas em Andrade et al (2000). A escolha dos parâmetros é feita da seguinte forma:

- **a** é obtido a partir da correlação ponto bisserial;
- **b** é obtido a partir da normal associada à correlação ponto bisserial.
- **c** é obtido a partir do número de alternativas ($1/m_i$ onde m_i é o número de alternativas do item i).

2 – possibilita que o usuário atribua uma constante única para as estimativas iniciais de **a**, enquanto que **b** fica estimado da mesma maneira que na opção **1**. Ou seja:

- todos valores de **a** recebem uma constante definida pelo usuário no campo **chute.a**;
- **b** é obtido a partir da normal associada à correlação ponto bisserial.
- **c** é obtido a partir do número de alternativas ($1/m_i$ onde m_i é o número de alternativas do item i).

3 – nesta opção, pode-se atribuir constantes tanto para o parâmetro **a** quanto para o parâmetro **b**:

- todos valores de **a** recebem uma constante definida pelo usuário no campo **chute.a**;
- todos valores de **b** recebem uma constante definida pelo usuário no campo **chute.b**.
- **c** é obtido a partir do número de alternativas ($1/m_i$ onde m_i é o número de alternativas do item i).

4 – nesta opção, são gerados valores aleatórios para os parâmetros **a** e **b**:

- os parâmetros **a** são gerados aleatoriamente a partir de uma distribuição log-normal com média definida pelo usuário em **lognor.media** e desvio-padrão definido em **lognor.dp**.
- **b** é gerado a partir da curva normal (0,1).
- **c** é obtido a partir do número de alternativas ($1/m_i$ onde m_i é o número de alternativas do item i).

5 – nesta opção, são gerados valores aleatórios para os parâmetros **a**, **b** e **c**:

- os parâmetros **a** são gerados aleatoriamente a partir de uma distribuição log-normal com média definida pelo usuário em **lognor.media** e desvio-padrão definido em **lognor.dp**.

- **b** é gerado a partir da curva normal (0,1).
- **c** é gerado a partir de uma beta com parâmetros **c1** e **c2** definidos pelo usuário.

Depois disso, pode-se fazer algumas configurações que chamamos de avançadas. Como já dissemos, tais configurações não precisam ser necessariamente alteradas, podendo ser consideradas como padrão do programa. Tais configurações podem ser vistas na Figura A.14.

```
###
### Configurações avançadas
###

### Geração dos pontos de quadratura:

# número de pontos de quadratura
npq<-10

# amplitude do intervalo para pontos de quadratura
limite<-4

# média da curva normal para pontos de quadratura
m.norm<-0

# variância da curva normal para pontos de quadratura
var.norm<-1

### atualização condicional do algoritmo EM:

# limite inferior para parâmetro a
lia<-0

# limite superior para parâmetros a
lsa<-2.2

# limite para parâmetro d
ld<-6

### critérios de parada:

# 1 - número de ciclos do algoritmo EM
n.ciclo.EM<-50

# 2 - precisão das estimatimas dos itens
prec.item.EM<-0.001
```

Figura A.14: Configurações avançadas do arquivo *fase3-estima parametros-x.txt.*

Nas configurações avançadas, podemos fazer algumas definições com relação à geração dos pontos de quadratura:

- npq: número de pontos de quadratura a serem gerados;
- limite: amplitude do intervalo onde serão gerados os pontos de quadratura. Assim, por exemplo, um valor igual a 4 indica que os pontos de quadratura serão gerados no intervalo [-4,4];
- m.norm: média da curva normal utilizada para a geração dos pontos de quadratura;
- var.norm: variância da curva normal utilizada para a geração dos pontos de quadratura.

A seguir, aparecem algumas definições a respeito da atualização condicional do algoritmo EM. Por questões de convergência dos valores, em algumas simulações foi observado que poderiam ocorrer problemas com as estimativas dos parâmetros **a**. Por isso, pode-se restringir o intervalo de seus possíveis valores. Caso **a** ultrapasse os valores definidos pelo usuário, o que o programa faz é não atualizar, no algoritmo EM, o valor desse parâmetro, ficando o valor anterior. Da mesma forma, foi colocada, dentro do programa, uma atualização condicional que não permite que o valor de **c** ultrapasse o intervalo [0,1]. Foi utilizada uma reparametrização, por questões computacionais, que facilita o cálculo de algumas derivadas (do vetor escore e da matriz hessiana) tal que $d_i=-a_i.b_i$ e esse parâmetro d também possui uma atualização condicional. Se, por exemplo, esperamos que o parâmetro **a** varie em [0,2] e **b** varie em [-3,3], esperamos que o parâmetro **d** varie de -6 até 6. Logo, se for definido 6 para os limites desse parâmetro, valores estimados que estejam fora do intervalo [–6,6] serão rejeitados e o programa manterá a última estimativa compreendida no intervalo definido. Maiores informações sobre essa reparametrização podem ser encontradas em Baker (1992). Assim:

- lia: limite inferior para o parâmetro de discriminação (**a**);

- lsa: limite superior para o parâmetro de discriminação (**a**);
- ld: limite para o parâmetro **d**.

Depois, existem duas opções que compõem os critérios de parada do algoritmo EM:

- n.ciclo.EM: número máximo de ciclos EM executados antes de interromper o processo de estimação;
- prec.item.EM: precisão do item dentro do algoritmo EM. Essa precisão é calculada através da diferença entre os valores das estimativas no passo **t** e os valores no passo **t+1**. Caso a diferença seja menor que prec.item.EM, o algoritmo é interrompido.

Quando algum dos dois critérios for alcançado, o algoritmo EM é interrompido. Dentro do algoritmo EM, utilizou-se o algoritmo Newton-Raphson (NR) para se fazer a maximização da função de máxima verossimilhança. Neste programa, utilizou-se apenas 1 laço de NR, ou seja, a cada ciclo EM, o algoritmo NR é rodado apenas uma única vez. Optou-se por este método devido ao fato dele apresentar os melhores resultados (em relação à utilização de mais de um laço de NR) em simulações. Algumas sub-rotinas de erro foram implementadas da mesma forma que as descritas em A.2. Na saída, são gerados dois arquivos de texto chamados *saída-fase3-TRI-x.txt* e *parametrosx.txt*. O primeiro arquivo mostra na saída o número de ciclos do algoritmo EM rodados até se obter a convergência ou atingir o critério de parada, mostra o número de itens considerados e desconsiderados na análise e o total de respondentes. Podemos ver, também, qual método para se obter as estimativas iniciais foi utilizado, dentre os 4 possíveis apresentados. É apresentada uma tabela contendo quatro colunas: nome do item, estimativas dos parâmetros de discriminação (**a**), estimativas dos parâmetros de dificuldade (**b**) e estimativa dos parâmetros de acerto casual (**c**). Em cada linha, para cada um dos itens, são mostrados as estimativas dos itens e, na linha de baixo, os erros-padrão correspondentes. Além disso, é gerado um

segundo arquivo chamado *parametrosx.txt* (ou *parametrosy.txt*) que possuem os valores dos parâmetros **a**, **b** e **c** dos itens considerados na análise. Esse arquivo é criado com o único propósito de ser utilizado por outros programas como o de calibração de habilidades e equalizações pela TRI.

A.9. Programa de estimação na TRI – estimação das habilidades

Nome do programa em *EstatR.exe*:	Fase 3 – Estimação das habilidades – Teste X
Nome do arquivo:	fase3-estima habilidades-x.txt
Versão:	Teste X
Arquivos de entrada necessários:	entradax.txt parametrosx.txt
Arquivos de saída gerados:	saida-fase3-habil-x.txt habilidadesx.txt

Nome do programa em *EstatR.exe*:	Fase 3 – Estimação das habilidades – Teste Y
Nome do arquivo:	fase3-estima habilidades -y.txt
Versão:	Teste Y
Arquivos de entrada necessários:	entraday.txt parametrosx.txt
Arquivos de saída gerados:	saida-fase3-habil-y.txt habilidadesy.txt

Esta implementação realiza a estimação das habilidades dos respondentes quando os parâmetros dos itens são conhecidos. Utiliza, para isso, o método de máxima verossimilhança. Para rodar este programa, é necessário que o usuário realize algumas configurações nos arquivos *fase3-estima habilidades-x.txt* e/ou *fase3-estima habilidades -y.txt*, conforme podemos ver na Figura A.15.

```
# Configure o local onde estão os arquivos e onde será gravado o
arquivo de saída.
setwd("C:/TRI/exemplo")

# número máximo de iterações do algoritmo Newton Raphson
tmax<-50

# critério de parada para o algoritmo Newton Raphson (valor em
módulo)
crit<-0.01
```

Figura A.15: Configurações básicas do arquivo *fase3-estima habilidades-x.txt.*

O usuário deve definir apenas o local onde estão os arquivos de entrada e onde serão salvos os arquivos de saída e deve definir os critérios de parada do algoritmo Newton-Raphson:

- tmax: número máximo de iterações que podem ser realizadas pelo algoritmo Newton-Raphson;
- crit: o algoritmo é interrompido quando o valor, em módulo, da diferença entre a estimativa da habilidade no passo **t** e no passo **t+1** for menor que o valor definido em **crit**.

No arquivo de saída (*saida-fase3-habil-x.txt)*, é apresentado o número de respondentes do teste em questão e uma tabela contendo a identificação do respondente, sua habilidade estimada, o erro padrão associado e o número de ciclos do algoritmo Newton-Raphson até se atingir a estimativa. Observou-se, em algumas situações, problemas na estimação de alguns indivíduos, o que ocasiona, na saída, um valor igual a **NaN** (*Not a Number*). Isso indica que o programa obteve valores muito pequenos com os quais não conseguiu realizar todos os cálculos e, por isso, é informado que o valor obtido não é um número. Também observou-se que esse problema depende, não só das respostas do indivíduo, mas também do conjunto de dados. Ou seja, se num conjunto de dados um indivíduo recebeu **NaN**, em outro conjunto, o mesmo indivíduo pode ter sua habilidade calculada normalmente. Quanto a quantidade de casas decimais, pode ser que sejam exibidas mais do que 3 casas. Isso ocorre geralmente quando o valor é muito

pequeno, ou muito grande, e o **R**, automaticamente transforma esse valor para notação científica. É gerado também um arquivo de texto chamado *habilidadesx.txt* que contém os valores das habilidades estimadas para os indivíduos. Esta saída serve apenas para ser usada em outras implementações, como as de equalização.

A.10. Programa de equalização na TRI – 2 testes

Nome do programa em *EstatR.exe*:	Fase 4 – Equalização entre 2 testes
Nome do arquivo:	fase4-equalizacao 2 teste.txt
Versão:	Única
Arquivos de entrada necessários:	entradax.txt entraday.txt parametrosx.txt parametrosy.txt habilidadesx.txt (opcional) habilidadesy.txt (opcional)
Arquivos de saída gerados:	saida-fase4-param2testes.txt saida-fase4-graf.jpg saida-fase4-habil2testes.txt (opcional)

Este programa realiza a equalização a posteriori entre dois testes: transforma os parâmetros dos itens e as habilidades dos respondentes (opcional) da escala do teste **X** para a métrica do teste **Y**. Admite-se o modelo ML3. Como configurações iniciais, o usuário deve informar o local onde estão os arquivos de entrada e onde estão os arquivos de saída. Deve decidir se deseja ou não que o programa faça a equalização das habilidades colocando **S** para equalizar ou **N** para não equalizar (opção **equahabi**). Deve definir, também, qual método deseja utilizar: **1** para média-desvio ou **2** para média-média (opção **método**). Finalmente, deve escolher o que deseja fazer com os valores dos parâmetros de itens comuns. Ou seja, para um item comum do teste **Y** temos os valores dos parâmetros; para o item correspondente no teste **X** temos outros valores que, quando equalizados, devem

se aproximar dos valores do teste **Y**. Temos, portanto, dois valores diferentes para os parâmetros de um mesmo item. O programa disponibiliza duas alternativas na opção **metcomum**:

1 – obter a média entre os valores dos parâmetros de ambos testes;

2 – descartar os valores equalizados e manter os valores do teste **Y**.

As configurações mencionadas podem ser vistas na Figura A.16.

```
# Configure o local onde estão os arquivos e onde será gravado o arquivo
de saída.
setwd("C:/TRI/pacote")

# Você deseja equalizar as habilidades (necessário arquivos
habildiadesx.txt
# e habilidadesy.txt) ? Configure "S" para sim ou "N" para não.
equahabi<-"N"

# escolha o método de equalização a posteriori:
# 1 para média-desvio ou 2 para média-média
metodo<-1

# Para os parâmetros dos itens comuns deve-se
# 1 - obter a média dos parâmetros de Y com os equalizados de X
# 2 - manter os parâmetros do teste Y
metcomum<-1
```

Figura A.16: Configurações iniciais do arquivo *fase4-equalizacao 2 testes.txt*.

O programa sempre gera a saída contendo os parâmetros equalizados (arquivo *saida-fase4-param2testes.txt*). Nesse arquivo, podemos encontrar os seguintes elementos:

Media (a): média dos parâmetros de discriminação (**a**) para os testes **X** e **Y**.

Media (b): média dos parâmetros de dificuldade (**b**) para os testes **X** e **Y**.

Desvio-padrão (b): desvio-padrão dos parâmetros de dificuldade (**b**) para os testes **X** e **Y**.

alfa: valor do parâmetro α utilizado na equalização e definido em 5.4.1.

beta: valor do parâmetro β utilizado na equalização e definido em 5.4.1.

Mais adiante, são mostrados o número de itens considerados na análise para o teste **X** e para o teste **Y** e o número de itens comuns a ambos testes. Depois é

apresentada uma tabela contendo, em suas colunas, os nomes dos itens e os valores já equalizados dos parâmetros **a**, **b** e **c**. Também é gerado um arquivo chamado *saida-fase4-graf.jpg* que contém os gráficos de dispersão dos parâmetros a e b entre os testes **X** e **Y**. A idéia é que o usuário possa verificar se existe, a partir dos gráficos, uma relação linear entre os valores equalizados dos parâmetros para os dois testes. Opcionalmente, se definido pelo usuário, o programa pode gerar um arquivo contendo as habilidades dos respondentes equalizadas: primeiro aparecem as habilidades dos indivíduos que realizaram o teste X e, depois, que fizeram o teste Y, na mesma seqüência em que os indivíduos foram colocados no arquivo *entradax.txt* e *entraday.txt*. Foi implementada uma sub-rotina de erro que avisa, através de janelas, caso não seja encontrado nenhum item em comum nos dois testes. Como já foi dito, para que o programa reconheça os itens comuns é preciso que os nomes atribuídos aos itens sejam os mesmos nos arquivos de entrada, ou seja, mesmo item deve ter o mesmo nome.

A.11. Programa de equalização na TRI – escala pré-definida

Nome do programa em *EstatR.exe*:	Fase 4 – Equalização – escala X
Nome do arquivo:	fase4-equalizacao escala-x.txt
Versão:	Teste X
Arquivos de entrada necessários:	entradax.txt parametrosx.txt Itenscomunsx.txt habilidadesx.txt (opcional)
Arquivos de saída gerados:	saida-fase4-param-escala-x.txt saida-fase4-habil-escala-x.txt (opcional)

Nome do programa em *EstatR.exe*:	Fase 4 – Equalização – escala Y
Nome do arquivo:	fase4-equalizacao escala-y.txt
Versão:	Teste Y
Arquivos de entrada necessários:	entraday.txt parametrosy.txt Itenscomunsy.txt habilidadesy.txt (opcional)
Arquivos de saída gerados:	saida-fase4-param-escala-y.txt saida-fase4-habil-escala-y.txt (opcional)

Estes programas permitem a equalização a posteriori transformando os parâmetros dos itens e as habilidades dos respondentes (opcional) da escala de um dos testes para uma escala pré-definida (obtida a partir de uma equalização prévia entre 2 testes). Isso permite comparar um conjunto de itens com outro conjunto de itens que já haviam sido calibrados para uma determinada métrica. Admite-se o modelo ML3. Como configurações iniciais, o usuário deve informar o local onde estão os arquivos de entrada e onde estão os arquivos de saída. Deve decidir se deseja ou não que o programa faça a equalização das habilidades colocando **S** para equalizar ou **N** para não equalizar (opção **equahabi**). Deve definir, também, qual método deseja utilizar: **1** para média-desvio ou **2** para média-média (opção **método**). Deve escolher o que deseja fazer com os valores dos parâmetros de itens comuns, conforme explicado em A.10, na opção **metcomum**:
1 – obter a média entre os valores dos parâmetros de ambos testes;
2 – descartar os valores equalizados e manter os valores do teste Y.

Finalmente, devem ser configuradas informações relativas à métrica da escala pré-definida. Caso o método escolhido seja o média-desvio (opção **1** de **metodo**), deve-se configurar o valor da média do parâmetro **b** na opção **media.bpd** e do desvio-padrão desse parâmetro em **dp.bpd**, correspondentes aos itens comuns utilizados na composição da escala pré-definida. A equalização ajustará os parâmetros do teste para essa escala. Neste caso, deixe qualquer valor na opção **media.apd**, visto que ela não será utilizada nos cálculos. Caso o método

escolhido seja o média-média (opção **2** de **metodo**), deve-se definir o valor da média dos itens comuns que compõem a escala pré-definida do parâmetro **a** na opção **media.apd** e da média do parâmetro **b** na opção **media.bpd**. Neste caso, deixe qualquer valor na opção **desvio.bpd**, visto que ela não será utilizada nos cálculos. As configurações mencionadas podem ser vistas na Figura A.17.

```
# Configure o local onde estão os arquivos e onde será gravado o arquivo
de saída.
setwd("C:/TRI/exemplo")

# Você deseja equalizar as habilidades (necessário arquivos
habilidadesx.txt)?
# Configure "S" para sim ou "N" para não.
equahabi<-"N"

# escolha o método de equalização a posteriori:
# 1 para média-desvio ou 2 para média-média
metodo<-1

# Valor da média do parâmetros de discriminação (a) da métrica pré-
definida
# (necessário apenas no método média-média; caso o método escolhido seja
# média-desvio, deixar qualquer valor no campo abaixo).
media.apd<-1

# Valor da média do parâmetros de dificuldade (b) da métrica pré-definida
media.bpd<-0.5

# valor do desvio-padrão da métrica pré-definida
# (necessário apenas no método média-desvio; caso o método escolhido seja
# média-média, deixar qualquer valor no campo abaixo).
dp.bpd<-1
```

Figura A.17: Configurações iniciais do arquivo *fase4-equalizacao escala-x.txt*.

O programa sempre gera a saída contendo os parâmetros equalizados (arquivo *saida-fase4-param-escala-x.txt* ou *saida-fase4-param-escala-y.txt*). Esses arquivos contém os seguintes elementos: em **Num.itens consider.** temos o número de itens considerados do teste inteiro ($\mathbf{X}$ ou $\mathbf{Y}$) e número de itens comuns (tais itens devem ser informados no arquivo *itenscomunsx.txt* ou *itencomunsy.txt*). Depois, aparecem informações sobre a escala pré-definida: média (**media b**) e desvio-padrão (**desvio b**) do parâmetro de dificuldade quando o método escolhido for média-desvio ou aparecem as médias dos parâmetros de discriminação e de dificuldade (**media a** e **media b**) quando o método escolhido for média-média. Por

fim, há uma tabela contendo, em suas colunas, os nomes dos itens e os valores equalizados dos parâmetros **a**, **b** e **c** para a escala pré-definida.

O arquivo *saída-fase4-habil-escala-x.txt* (ou *saída-fase4-habil-escala-y.txt*) possui os valores das habilidades equalizadas para a métrica pré-definida. A seqüência em que os valores aparecem corresponde a mesma seqüência de identificação dos indivíduos no arquivo de entrada.

A.12. Resumo dos arquivos

A seguir, mostramos a Tabela A.1 que poderá ser útil na hora de localizar os arquivos de entrada necessários para cada simulação e quais são os arquivos de saída. Nela estão disponíveis os nomes do programa dentro do software EstatR.exe, os nomes dos arquivos originais dos programas (arquivo mãe), os arquivos de entrada necessários para que o arquivo mãe rode e, finalmente, os arquivos de saída gerados.

Tabela A.1: Resumo dos arquivos das implementações

Nome em EstatR.exe	Arquivo mãe	Arquivos de entrada	Arquivos de saída
Fase 1 – Análise de itens – Teste X	fase1-x.txt	entradax.txt	saida-fase1-x.txt saida-fase1-hist-x.jpg
Fase 1 – Análise de itens – Teste Y	fase1-y.txt	entraday.txt	saida-fase1-y.txt saida-fase1-hist-y.jpg
Fase 2 – Equalização clássica - Tucker	fase2-Tucker.txt	entradax.txt entraday.txt	saida-fase2-Tucker.txt saida-fase2-graf-Tucker.jpg
Fase 2 – Equalização clássica - Levine	fase2-Levine.txt	entradax.txt entraday.txt	saida-fase2-Levine.txt saida-fase2-graf-Levine.jpg
Fase 2 – Equalização clássica - Equipercentil	fase2-Equipercentil.txt	entradax.txt entraday.txt	saida-fase2-Equipercentil.txt
Fase 2 – Equalização clássica - Braun-Holland	fase2-Braun-Holland.txt	entradax.txt entraday.txt	saida-fase2-BraunHol.txt saida-fase2-graf-BraunHol.jpg
Fase 3 – Estimação dos parâmetros – Teste X	fase3-estima parametros-x.txt	entradax.txt	saida-fase3-TRI-x.txt parametrosx.txt
Fase 3 – Estimação dos parâmetros – Teste Y	fase3-estima parametros-y.txt	entraday.txt	saida-fase3-TRI-y.txt parametrosy.txt
Fase 3 – Estimação das habilidades – Teste X	fase3-estima habilidades-x.txt	entradax.txt parametrosx.txt	saida-fase3-habil-x.txt habilidadesx.txt
Fase 3 – Estimação das habilidades – Teste Y	fase3-estima habilidades-y.txt	entraday.txt parametrosy.txt	saida-fase3-habil-y.txt habilidadesy.txt

Tabela A.1(Cont.): Resumo dos arquivos das implementações

Nome em EstatR.exe	Arquivo mãe	Arquivos de entrada	Arquivos de saída
Fase 4 – Equalização entre 2 testes	fase4-equalização 2 testes.txt	entradax.txt	saida-fase4-param2testes.txt
		entraday.txt	saida-fase4-graf-A.jpg
		parametrosx.txt	saida-fase4-graf-B.jpg
		parametrosy.txt	saida-fase4-habil2testes.txt (opcional)
		habilidadesx.txt (opcional)	
		habilidadesy.txt (opcional)	
Fase 4 – Equalização – escala X	fase4-equalização escala-x.txt	entradax.txt	saida-fase4-param-escala-x.txt
		parametrosx.txt	saida-fase4-habil-escala-x.txt (opcional)
		itenscomunsx.txt	
		habilidadesx.txt (opcional)	
Fase 4 – Equalização – escala Y	fase4-equalização escala-y.txt	entraday.txt	saida-fase4-param-escala-y.txt
		parametrosy.txt	saida-fase4-habil-escala-y.txt (opcional)
		itenscomunsy.txt	
		habilidadesy.txt (opcional)	

Bibliografia

Andrade, D.F. (2001). Comparando desempenhos de grupos de alunos por intermédio da Teoria da Resposta ao Item. In: **Estudos em Avaliação Educacional** (n° 23, pp. 31-69). Fundação Carlos Chagas: São Paulo.

Andrade, D.F., Tavares, H.R. & Valle, R.C. (2000). **Teoria da Resposta ao Item: Conceitos e Aplicações**. ABE: São Paulo.

Angoff, W.A. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), **Educational measurement** (2ª ed., pp. 508-600). Washington, DC: American Council on Education. (Reprinted as W.A. Angoff, **Scales, norms, and equivalent scores**. Princeton, NJ: Educational Testing Service, 1984.)

Azevedo, C.L.N. (2003). **Métodos de Estimação na Teoria de Resposta ao Item**. Dissertação de Mestrado – Instituto de Matemática e Estatística da Universidade de São Paulo (USP), São Paulo.

Beaton, A.E. and Allen, N.L. (1992). Interpreting scales through scale anchoring. **Journal of Educational Statistics, 17**, 191-204.

Aitkin, M. & Bock, R.D. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. **Psychometrika, 46**, 443-459.

Baker, F.B. (1992). **Item Response Theory**. New York: Marcel Dekker, Inc.

Braun, H.I. & Holland, P.W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland & D.B. Rubin (Eds.), **Test equating**. New York: Academic Press.

Bussab, W.O. (1988). **Análise de variância e de regressão**. São Paulo: Atual.

Bussab, W.O. & Morettin, P.A. (1987). **Estatística básica**. São Paulo: Atual.

Dantas, C.A.B. (1997). **Probabilidade: um curso introdutório**. São Paulo: Edusp.

Garret, H.E. (1958). **A estatística na psicologia e na educação**. Rio de Janeiro: Editora Fundo de Cultura.

Gulliksen, H (1950). **Theory of mental tests**. New York: Wiley.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). **Fundamentals of Item Response Theory**. Newbury Park: SAGE Publications.

Holland, P.W. & Thayer, D.T. (1989). **The kernel method of equating score distribuitions** (Techinical Report n° 89-84). Princeton, NJ: Educational testing service.

Jarjoura, D. & Kolen, M.J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. **Journal of Educational Statistics**, 10, 143-160.

Kolen, M.J. & Brennan, R.L. (1995). **Test equating – methods and practices**. New York: Springer-Verlag.

Lord, F.M. (1980). **Applications of item responde theory to practical testing problems**. Hillsdale, New Jersey: Lawrence Erlbaum.

Lodr, F.M. & Novick, M.R. (1968). **Statistical theories of mental test scores**. Reading, MA: Addison-Wesley.

Lucero, I. & Meza, S. (2002). Validación de instrumentos para medir conocimientos. Disponível em: <http://www.unne.edu.ar/cyt/2002/09-Educacion/D-027.pdf>. Acesso em fevereiro/2006.

Magalhães, M.N. & Lima, A.C.P. (2001). **Noções de probabildiade e estatística**. São Paulo: IME-USP.

Nunnally, J.C. (1964). **Educational Measurement and Evaluation**. New York: McGraw-Hill Book Company.

Paradis, E. **R para Principiantes**. Disponível em: <http://cran.r-project.org/doc/contrib/rdebuts_es.pdf>. Acesso em fevereiro/2005.

Petersen, N.S., Kolen, M.J. & Hoover, H.D. (1989). Scaling, norming and equating. In R.L. Linn (Ed.), **Educational measurement** (3ª ed., pp. 221-262). New York: Macmillan.

Senno, R.M. (2006). **Métodos de Equalização na Teoria Clássica e na Teoria da Respostas ao Item.** Dissertação de mestrado – IME-USP, São Paulo.

Spiegel, M,R, (1975). **Estatística**. São Paulo: McGraw-Hill.

Valle, R.C. (1999). **Teoria da Resposta ao Item**. Dissertação de Mestrado – Instituto de Matemática e Estatística da Universidade de São Paulo (USP), São Paulo.

Vianna, H.M. (1982). **Testes em educação**. São Paulo: IBRASA.