Universidade Federal de Juiz de Fora
Pós-Graduação em Educação

Wellington Silva

Eficácia dos processos de *linkagem* na Avaliação Educacional em Larga Escala

Juiz de Fora
2010

Wellington Silva

Eficácia dos processos de *linkagem* na Avaliação Educacional em Larga Escala

Dissertação apresentada ao Programa de Pós-Graduação em Educação da Universidade Federal de Juiz de Fora, como requisito parcial para a obtenção do título de Mestre.

Orientador: Prof. Dr. Tufi Machado Soares

Juiz de Fora
2010

Wellington Silva

EFICÁCIA DOS PROCESSOS DE *LINKAGEM* NA AVALIAÇÃO EDUCACIONAL EM LARGA ESCALA

Dissertação apresentada ao Programa de Pós-Graduação em Educação da Universidade Federal de Juiz de Fora, como requisito parcial para a obtenção do título de Mestre.

BANCA EXAMINADORA

_____

Prof. Dr. Tufi machado Soares (Orientador)

Programa de Pós-Graduação em Educação da UFJF

_____

Prof. Dr. Eduardo Magrone

Programa de Pós-Graduação em Educação da UFJF

_____

Prof. Dr. Jose Ignacio Cano Gestoso

Programa de Pós-Graduação em Sociologia da UERJ

_____

Prof. Dr. Amaury Patrick Gremaud

Programa de Pós-Graduação em Administração e Economia da USP

À Josiane, companheira de longa jornada, pelo apoio e confiança sempre depositados em sonhos que já não são apenas meus.

A Lucas e à Elisa, filhos tão maravilhosamente diferentes em se relacionar com o mundo, fontes constantes de sentimentos bons, que me inspiram a novos projetos.

Aos meus tios Ivan e Heloisa, pelo despertar de uma profissão através dos estudos.

A meus pais, Pedro e Ivanilde, pela confiança e empenho dedicados aos meus estudos em Juiz de Fora. Hoje, também sinto, talvez, a mesma dor no peito ao permitir a realização do sonho de um filho.

A meu primeiro chefe, Dr. João Paulo do Amaral Braga, pela oportunidade de exercer minha formação inicial como engenheiro eletricista e ferroviário e pelos ensinamentos de uma visão sempre confiante em relação às adversidades.

À minha atual chefe, Professora Lina Kátia Mesquita de Oliveira, pela oportunidade de exercer minha atual formação como analista de políticas públicas educacionais, pela amizade e pelos ensinamentos de que temos que fazer a diferença.

Aos meus avós José Etienne e Francisco Caju e suas Marias na ordem inversa fechando o ciclo: Uma italiana, chamada Nega a outra José, assim como tantas. A benção!

AGRADECIMENTOS

Ao Centro de Políticas Públicas e Avaliação da Educação da Universidade Federal de Juiz de Fora – CAEd/UFJF - pela confiança depositada em meu trabalho ao longo desses anos.

À professora Jane Azevedo, pela ajuda em momento de dificuldade que proporcionou minha atual atividade profissional.

Ao Prof. Tufi Machado Soares, meu orientador, pela oportunidade de ingressar no CAEd, pela companhia em trabalhos realizados em avaliação educacional e pela orientação na realização deste trabalho.

Ao professor Manuel Palácios, pelas orientações nessa dissertação e em minhas atividades profissionais.

Aos colegas da Coordenação de Medidas Educacionais do CAEd/UFJF, Ailton, Rafael, Clayton e Roberta, não apenas pela ajuda nas preparações de arquivos utilizados nessa dissertação, mas, principalmente pela presença constante ao longo de anos de trabalho conjunto.

À professora Mônica e seu pai Fernando Prado, pelos ensinamentos que de alguma forma transformaram ou criaram algumas de minhas realizações e convicções.

Aos estatísticos do Ministério da Educação da França, Thierry Rocher e Pierre Vrignaud pela cordial colaboração nas discussões técnicas sobre as avaliações realizadas pelo CAEd.

Ao meu amigo Marcos Baeta pela parceria na construção desta minha história.

Aos meus amigos Cyril e Laurance que me despertaram para a importância da formalização do agradecimento.

RESUMO

Em 1997, através do Sistema Nacional de Avaliação da Educação Básica – SAEB, definiu-se a escala de proficiência para o Brasil. A partir de então, praticamente todas as avaliações em larga escala realizadas por diversos estados brasileiros têm procurado manter uma comparabilidade de resultados com essa escala, por meio da Metodologia da Teoria da Resposta ao Item – TRI. Entretanto observa-se uma diversidade de situações ao se analisar as diferentes avaliações realizadas pelos Estados brasileiro e até mesmo no próprio SAEB. Nesse trabalho, apresentaremos alguns aspectos técnicos necessários para se garantir a comparabilidade nos procedimentos de *linkagem* de avaliações, bem como as características das avaliações do SAEB e de alguns estados brasileiros ao longo do tempo.

Palavras-chave: Teoria da Resposta ao Item (TRI), Avaliação em Larga Escala, Métodos de Equalização (*Linkagem*).

ABSTRACT

EFFECTIVENESS OF THE LINKING PROCESSES IN EDUCATIONAL LARGE SCALE ASSESSMENT

In 1997, through the National System of Basic Education Evaluation (*SAEB*), the proficiency scale for Brazil was defined. From that time on, almost all the assessment realized by several Brazilian states have tried to keep a result comparability with this scale through Item Response Theory Methodology (*IRT*). However, a variety of situations is observed when different assessments realized in Brazilian states or even at *SAEB* are analyzed.In this article, some technical aspects needed for ensuring the comparability in the assessment linking procedures are presented, as well as the characteristic of *SAEB´s* assessment and some Brazilian states´ assessment throughout time.

Key- words: Item Response Theory (*TRI), Large Scale* Assessment , Equalization Methods, Linkage.

SUMÁRIO

Lista de Ilustrações

INTRODUÇÃO

Em 1997, o presidente americano Bill Clinton propôs a unificação de todas as escalas de proficiência produzidas pelos diversos programas Estaduais, distritais e comercias aplicados nos EUA. Este projeto consistiria em uma tentativa de agrupar todas essas escalas ao Sistema de Avaliação Nacional Americano, NAEP, com a possibilidade, através da criação de testes voluntários, fornecer o desempenho de cada aluno em 4 níveis de desempenho: Abaixo do Básico, Básico, Proficiente e Avançado.

Para estudar tal proposta, foi contratado o National Research Council - NRC, o qual foi produzido por Feuer em 1999, relatório intitulado *Uncommon Measures: Equivalence and linkage among educational tests.*

A proposta de Bill Clinton de se ter uma escala única está atrelada ao lema nacional americano "*e pluribus unum*" - De muitos, um. Entretanto, de acordo com as análises do NRC, este sonho americano não foi tecnicamente aprovado.

Este caso americano, com suas recomendações técnicas, será utilizado nessa dissertação para estudarmos e refletirmos sobre diversas situações de equiparação de escores dos sistemas de avaliações aplicados no Brasil.

Um ponto forte e que diferencia a realidade brasileira da americana é o fato de, no Brasil, já existir uma cultura de escala única referenciada ao Sistema Nacional de Avaliação da Educação Básica – SAEB. Essa característica nacional está fundamentada em certos fatores técnicos, como por exemplo, as avaliações realizadas, até então, pelos diversos estados brasileiros terem mantido uma mesma matriz de referência com o SAEB, o que é bem diferente da realidade americana, em que cada estado tem autonomia para elaborar sua própria matriz e também ao fato de utilizarem os mesmos modelos e procedimentos matemáticos utilizados pelo SAEB os quais serão explicitados ao longo dessa dissertação.

Entretanto, temos observado ao longo dos anos, tanto nas avaliações estaduais como no próprio SAEB, certas variações nas características dos testes no que se refere às disciplinas avaliadas e à quantidade de itens a serem respondidos pelos alunos. E também a certas divergências entre as avaliações estaduais e o SAEB no que se refere às populações utilizadas

para a calibração de itens. Diante desse cenário, será que estamos realmente comparando os resultados de nossas avaliações de forma confiável através de uma mesma escala?

Tendo como foco essa questão, estruturamos essa dissertação em cinco capítulos. Assim, faz-se necessário inicialmente definirmos alguns conceitos básicos relativos às avaliações educacionais em larga escala, os quais serão trabalhados no capítulo 1.

O propósito do capítulo 2 é fornecer um breve histórico da avaliação educacional através da utilização da Teoria da Resposta ao Item – TRI e as características das avaliações em larga escala no Brasil. Essas informações serão essenciais para o desdobramento de todo o restante do trabalho, pois buscaremos retratar as variações ocorridas nas avaliações ao longo dos anos as quais nos levam a questionar a confiabilidade na comparabilidade de seus resultados.

Nos capítulos 3 e 4 abordaremos os aspectos técnicos referentes à TRI, modelo matemático norteador de todo este trabalho. Assim, no capítulo 3 apresentaremos os modelos logísticos e métodos de estimação de parâmetros de itens e escores de alunos e, no capítulo 4, os procedimentos de equiparação de escores ou *linkagem*, que têm como objetivo a construção de uma mesma escala entre as populações envolvidas nas avaliações.

No capítulo 5, através de estudos de casos reais de avaliações em larga escala no Brasil, discutiremos como certas diferenças relativas ao nível de proficiência das populações e/ou características dos testes podem afetar a comparabilidade dos resultados. Para isso, agregaremos evidências observadas nas diversas análises e simulações de *linkagem*.

Nos anexos deste trabalho, apresentaremos os procedimentos para linkagem utilizando métodos não-lineares através do software BILOG-MG, bem como planilhas com parâmetros de itens, obtidos em diferentes simulações de *linkagen*s, cujas informações serão trabalhadas no capítulo 5.

1 CONCEITOS BÁSICOS EM MEDIDAS EDUCACIONAIS

A comparação do desempenho de diferentes grupos de estudantes ao longo do tempo é o principal objetivo das avaliações educacionais e um dos grandes desafios para a psicometria. Para realizar esta tarefa, podemos adotar dois procedimentos: aplicar o mesmo teste ou comparar formas diferentes de testes. No primeiro caso, não temos o erro de medida, mas em compensação os resultados poderão ser inflacionados pelo fato de os grupos passarem informações entre si, sobre o conteúdo dos testes. No segundo caso, que é o utilizado nas avaliações educacionais e que será o foco deste trabalho, eliminamos este efeito, mas, em contrapartida, estamos sujeitos aos erros inerentes aos processos de comparação entre diferentes testes.

Para podermos evoluir em nosso objetivo de comparabilidade de resultados entre avaliações, é fundamental o conhecimento dos seguintes termos: *linkagem*, equalização e escalonamento, o que será abordado nesse capítulo.

1.1 *LINKAGEM*

Utilizaremos neste nosso trabalho a expressão *linkagem*, proveniente de uma variação das palavras em inglês *linking* e *linkage* normalmente empregadas, para caracterizar métodos e procedimentos para equiparação ou equivalência de escores entre diferentes avaliações. Normalmente, no Brasil, a equiparação de escores é dita como equalização o que gera confusão ao estudarmos certos autores internacionais , em que a equalização é definida como uma situação particular de *linkagem*, já utilizando este termo para nos referirmos a equiparação de escores.

Desta forma, seguiremos a taxonomia de Mislevy e Linn, tal como apresentada por Kolen e Brennan (2004). Segundo estes autores, a *linkagem* subdivide-se em quatro tipos, em função da precisão da comparação que se deseja obter, assim, a equalização, seria um destes tipos. Quanto maior for o rigor na estruturação dos testes cujos resultados desejam ser comparados, maior será esta precisão. Dessa forma, as melhores comparações são obtidas quando, nas diferentes

avaliações: os testes medem o mesmo constructo, possuem a mesma estrutura, mesmos descritores distribuídos em testes paralelos, mesmo método estatístico para cálculo das proficiências e populações equivalentes. Variações nessas características influenciarão na qualidade da precisão da *Linkagem*, ou seja, no nível de robustez (força) entre as comparações das medidas obtidas nesses processos.

1.1.1. Métodos de *Linkagem*

Apresentamos, a seguir, os quatro tipos de *linkagem* em ordem decrescente de robustez no que se refere à comparabilidade dos resultados entre as avaliações:

- Equalização: termo utilizado quando se compara os resultados de diferentes formas de um mesmo teste que foi projetado para ser paralelo. Dessa forma, os testes medem os mesmos conteúdos, possuem os mesmos descritores, mesma estrutura, mesma forma de aplicação, pequena variação na dificuldade de itens similares que compõem as diferentes formas dos testes e as populações são equivalentes. Os resultados obtidos nesse processo de *linkagem* são os melhores possíveis, ou seja, têm-se o mesmo nível de confiabilidade para as diferentes formas.

Os escores obtidos por este processo são intercambiáveis, ou seja, se um teste X é equalizado a um teste Y, as interpretações obtidas através do teste X são equivalentes às obtidas no teste Y. Dizendo de uma outra forma, os resultados para o grupo que respondeu ao teste X seriam os mesmos se este grupo tivesse respondido ao teste Y. No entanto, conforme observado por Lord (1980), na prática, essa intercambialidade não é perfeita, pois, podem existir pequenas variações estruturais entre as diferentes formas do teste.

Para garantir a qualidade da equalização, devemos verificar: i) se a correspondência entre escores equalizados é simétrica, ou seja, uma única tabela de correspondência deve ser usada para obter os escores da forma X em Y e vice-versa; ii) a invariância de grupo: a função de equalização deve ser a mesma para qualquer subgrupo da população, por exemplo, sexo, raça, região ou política educacional adotada; iii) a invariância no tempo: não faz diferença se a equalização é baseada, por exemplo, em dados obtidos em 2000 ou 2005.

É importante ressaltar que a equalização ajusta diferenças de dificuldades entre testes que foram projetados para serem paralelos (similares em dificuldade e conteúdo), mas não ajusta diferença de conteúdos entre os mesmos.

- Equalização vertical ou Calibração: Este tipo de *linkagem* fornece mecanismos de comparação de escores de testes em que requerimentos como conteúdos, estrutura e formas de aplicação não são tão rigorosos quanto na equalização. Neste caso, a qualidade da *linkagem* é menor que no caso anterior, pois as medidas podem não possuir o mesmo nível de confiabilidade, ou seja, uma interpretação dos resultados de uma forma não é exatamente a mesma interpretação em outra forma do teste.

Contudo, esse tipo de *linkagem* não se apresenta sob uma única forma. Na verdade, podemos distinguir dois tipos de equalização vertical. O primeiro, quando temos diferentes formas de testes não necessariamente com a mesma estrutura e com diferentes conteúdos, aplicados a populações equivalentes. Isto ocorre, por exemplo, quando, no 5º ano de Matemática, se utiliza uma estrutura de Blocos Incompletos Balanceados (BIB) de 13 blocos com 13 itens, 26 cadernos com 3 blocos, sendo que cada caderno possui blocos comuns entre si.

O segundo, quando se deseja medir a performance dos alunos em diferentes níveis de escolaridades. Neste caso, a calibração é comumente denominada de equalização vertical ou escalonamento vertical (*Vertical equating* ou *vertical scaling*). Essa situação ocorre ao colocarmos em uma mesma escala alunos do 5º ano EF com alunos do 9° ano EF e alunos do 3º ano EM.

A principal diferença desse tipo de *linkagem* com relação à equalização, é que as formas de testes não são intercambiáveis entre os diferentes grupos, por exemplo, há uma perda na precisão dos resultados do 5º ano EF ao se aplicar nesse grupo de respondentes testes do 9º ano EF e esta perda também é observada na situação inversa, ou seja, aplicação de testes do 9º ano EF a respondentes do 5º ano EF.

Como esse tipo de *linkagem* é o normalmente utilizado nas avaliações estaduais com o SAEB, devemos ressaltar que, para se atingir tanto o objetivo de estimar as proficiências individuais dos alunos, quanto os percentuais de alunos em determinados níveis de proficiência, é essencial que as duas avaliações estejam ajustadas com relação à mesma abrangência de conteúdo, às mesmas demandas cognitivas exigidas dos alunos e às mesmas condições em que os

testes são administrados. Variações no grau de similaridades dessas condições influenciarão na confiabilidade das medidas obtidas.

Neste trabalho, utilizaremos, para este tipo de *linkagem*, o termo equalização vertical e deixaremos termo calibração para descrever o processo do cálculo dos parâmetros dos itens em uma mesma escala nos diferentes processos de *linkagem* (quase sempre em modelos da Teoria da Resposta ao Item).

- Projeção: forma unidirecional de *linkagem* aplicada a um mesmo grupo de respondentes em que os escores de um teste são projetados, por exemplo, via regressão (linear ou não-linear) para se obter os escores de outro teste, sem a expectativa de que os mesmos estejam medindo exatamente a mesma coisa, como, por exemplo, "linka" os escores de Língua Portuguesa com os de Matemática.

É importante mencionar que a projeção de A em B não é necessariamente a mesma que B em A. Assim, a precisão da projeção depende do quão forte é a relação entre os testes e necessita ser reavaliada frequentemente, pois a projeção é muito sensível e dependente do contexto, grupo utilizado para estabelecer a relação e tempo.

- Moderação: é o tipo mais fraco dos processos de *linkagem*, em que testes com especificações técnicas diferentes são comparáveis através de suas respectivas distribuições de escores, razão pela qual este método é também denominado de ajustamento por distribuição. Diferentemente da projeção, a moderação pode ser aplicada a grupos diferentes.

Distinguem-se dois tipos de moderação: (a) Moderação estatística: quando são utilizados procedimentos para ajustar as distribuições dos diferentes grupos através dos escores; (b) Moderação social: nesse tipo de *linkagem*, usam-se julgamentos obtidos de informações externas às situações dos testes. Os resultados obtidos por esses tipos de moderação servem apenas para comparações superficiais entre os grupos.

No Quadro 1.1 apresentamos os tipos de *linkagem* e suas principais características.

Característica da avaliação	Tipo de *linkagem*			
	Equalização	Calibração	Projeção	Moderação
Mede o mesmo conteúdo (constructo)	sim	sim	Não	Não
Mesma confiabilidade	sim	não	Não	Não
Mesma precisão da medida através dos diferentes níveis de conhecimento dos alunos	sim	não	Não	Não
Diferentes conversões para obter os resultados do teste X em Y com os resultados do teste Y em X.	não	talvez	Sim	Não
Diferentes conversões para as estimativas das distribuições individuais e de grupo	não	sim	Sim	Não
Checagens frequentes para verificar a estabilidade das conversões dos resultados no que diz respeito a diferentes conteúdos, diferentes grupos e diferentes períodos de aplicação da avaliação.	não	sim	Sim	Sim
Consenso em padrões de desempenho	não	não	Não	Sim

Quadro 1.1: Características dos diferentes tipos de *linkagem.* Fonte: Kolen e Brennan (2004)

1.2 DESIGNS OU DELINEAMENTOS PARA COLETA DE DADOS

Diferentes tipos de designs ou delineamentos para coleta de dados são utilizados em *linkagens* do tipo equalização e do tipo equalização vertical, com o objetivo de manter uma escala única entre diferentes formas dos testes e/ou grupos e, consequentemente, a comparabilidade entre seus resultados. Passaremos, a seguir, a descrever esses variados delineamentos para coleta de dados encontrados na literatura, conforme Kolem e Brennan (2004) e Pasquali (2004).

1.2.1 Design de grupos aleatórios

Nesse design cada examinando responde a apenas uma forma de teste, o que minimiza o tempo de aplicação, em relação a designs que exigem que todas as formas sejam respondidas pelo mesmo examinando. Para garantir uma distribuição igualitária das formas (podendo inclusive ser mais de duas formas), os cadernos de testes das diferentes formas são montados e distribuídos em 'espiral', ou seja: o primeiro examinado recebe o caderno 1, o segundo o caderno 2, o terceiro o caderno 1 e assim sucessivamente.

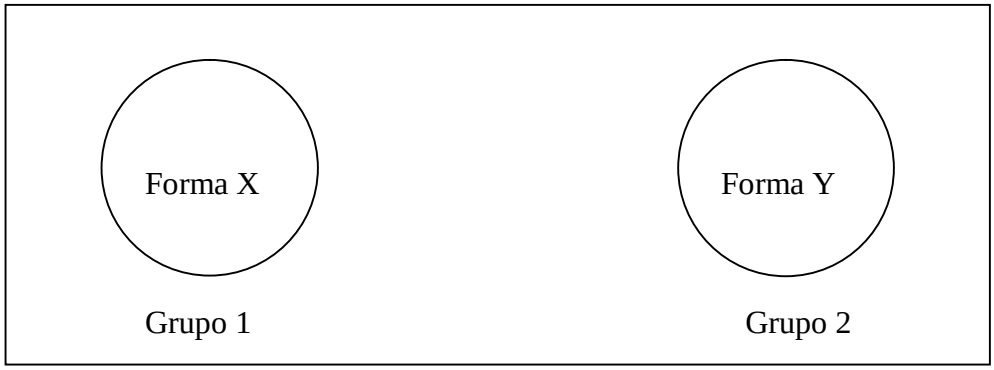

Figura. 1.1: Design de grupos aleatórios. Fonte: Kolem e Brennan (2004)

Uma vez garantidas a equidade na distribuição das diferentes formas de testes, assim como um significativo número de respondentes para estas formas, a diferença nas performances entre as formas, é um indicativo da variabilidade de dificuldade entre as mesmas.

Como desvantagem deste design, podemos destacar o fato de que as diferentes formas têm que ser aplicadas ao mesmo tempo o que inviabiliza a comparação de resultados em diferentes períodos de tempo.

1.2.2 Design de grupo simples

Em avaliações que utilizam esse design, um mesmo examinando responde a duas formas de testes, primeiro, por exemplo, à forma X e, em seguida, à forma Y. Nesse sentido, podemos ressaltar três características negativas nesse design, o que o torna pouco utilizada na prática.

A primeira característica é o efeito cansaço, ou seja, o fato de responder à forma Y depois de ter respondido à forma X faz com que o desempenho na forma Y seja prejudicado. A segunda diz respeito ao fato de o examinando levar uma certa vantagem ao fazer a segunda forma, uma vez que ele pode se beneficiar de alguma informação contida na primeira forma. A terceira refere-se ao número limitado de formas que podem ser utilizadas, geralmente duas, uma vez que os examinados serão submetidos a todas as formas do teste.

1.2.3 Design de grupo simples com balanceamento entre as formas

Esse design é, de certa maneira, semelhante ao design anterior. Nele, porém a ordem das formas são alternadas entre os respondentes, ou seja, em um teste com duas formas X e Y, metade da população responde ao teste na seqüência X – Y e a outra metade responde aos testes na ordem Y – X. Com esse procedimento, se houver efeito de cansaço entre as formas, o mesmo

será observado, e na análise dos resultados deverão ser considerados apenas os resultados das formas que foram respondidas em primeiro lugar.

Na prática, esse design deve ser usado em relação ao design de grupos aleatórios, quando: i) a aplicação das duas formas é viável operacionalmente; ii) não ocorre o efeito de diferença de performance dos respondentes devido à ordem das formas e iii) há necessidade de uma menor quantidade de respondentes para a obtenção dos resultados.

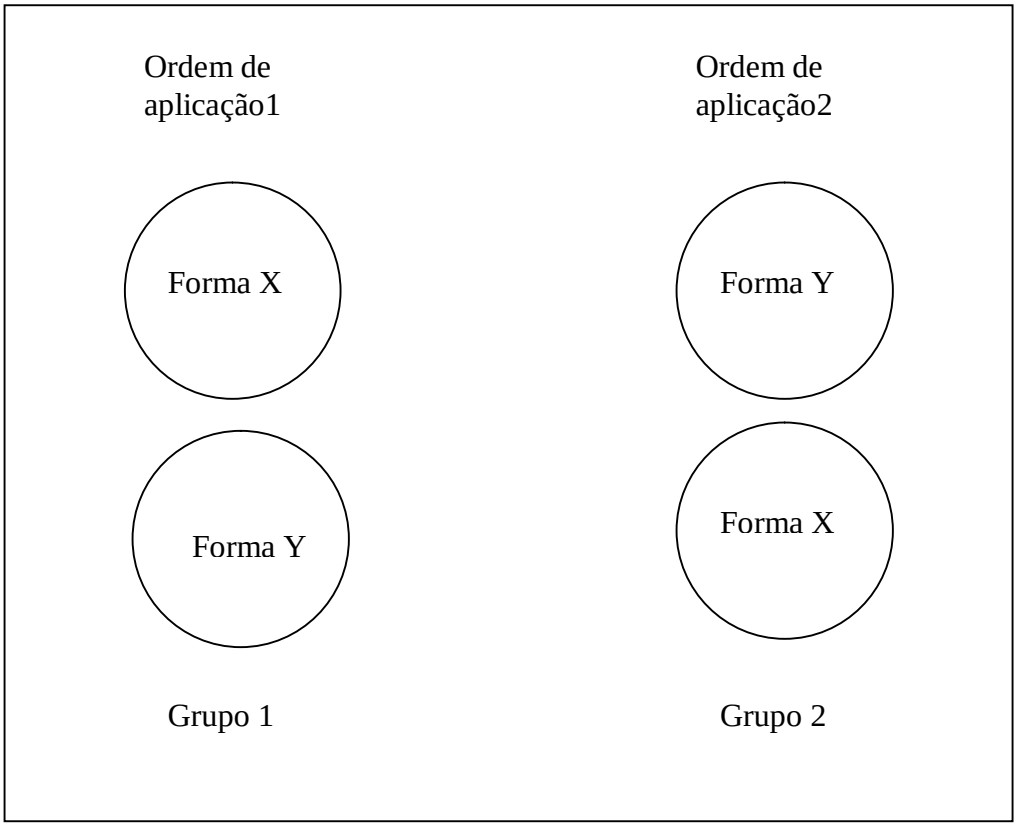

Figura. 1.2: Design de grupo simples com balanceamento entre as formas. Fonte: Kolen e Brennan (2004).

1.2.4 Design para grupos não equivalentes através de itens comuns

Diferentemente dos casos anteriores em que os grupos eram equivalentes, nesse design é possível haver grupos com diferentes características, por exemplo grupos em que as médias e distribuições dos escores não sejam iguais.

Nesse caso, o que possibilita o escalonamento entre os grupos é a inclusão de itens comuns entre as diferentes formas de testes, conforme apresentado na figura 1.3:

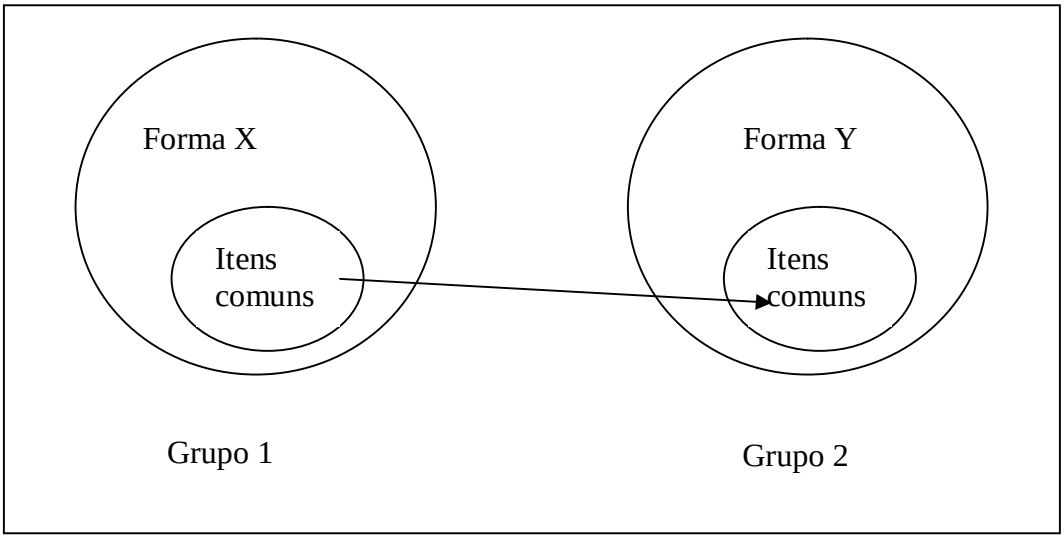

Figura. 1.3: Design para grupos não equivalentes com itens comuns. Fonte: Kolen e Brennan (2004).

Nessa situação, uma condição a ser observada para um bom escalonamento entre as formas, é garantir que os itens comuns sejam, por assim dizer uma miniatura do teste. Isto é, esses itens devem ser selecionados de maneira a se constituírem como uma amostra representativa do teste, garantindo uma boa representatividade no que se refere aos níveis de dificuldade e descritores do teste original, ou dizendo de outra forma, o conteúdo e as características estatísticas do teste total.

Para o cálculo do desempenho dos respondentes, os resultados provenientes dos itens comuns podem ou não ser considerados. Assim caso os resultados dos itens comuns sejam

considerados, o escalonamento é dito como interno e, no caso em que os itens comuns são desconsiderados o escalonamento é dito externo.

Devido a tanta flexibilidade na aplicação, esse é o design mais utilizadas nas avaliações em larga escala, entretanto, mesmo com uma grande variedade de métodos de *linkagens* existentes e que serão abordados no capítulo 4, nenhum deles terá uma boa confiabilidade se os grupos de examinandos e estruturas de testes forem muito diferentes.

Se tomarmos como exemplo a *linkagem* entre o SAEB 2007 com o SAEB 2009, em que temos 6 grupos representados pelos 5º EF, 9º EF e 3º EM dos anos de 2007 e 2009 e dentro de cada grupo temos 21 formas ou modelos de cadernos com as disciplinas de Língua Portuguesa e Matemática, conforme estrutura apresentada no capítulo 2, é possível observar a presença dos designs de itens comuns, ligando diferentes grupos e cadernos e de grupo simples com balanceamento entre as formas uma vez que as disciplinas são alternadas entre os cadernos.

O design de itens comuns é o mais utilizado nas avaliações nacionais e também o mais fácil de ser implementado. Entretanto, esse design está sujeito a um problema denominado de efeito de contexto: ao colocarmos itens comuns entre níveis de escolaridades adjacentes, o fato de se ter estes itens, por exemplo, no final do teste para o nível inferior e no início do teste para o nível superior, que é a situação mais frequentemente utilizada na prática, poderá provocar um erro sistemático no processo de *linkagem*.

1.2.5 Design teste de ancoragem

Nesse tipo de design, após os alunos responderem a testes com itens apenas de seu nível de escolaridade respondem também a um teste único com itens representativos de todas as séries avaliadas.

Dessa maneira, a construção da escala é feita através do teste de ancoragem e os escores de cada nível são linkados entre si, através desse mesmo teste de ancoragem.

O design de teste de ancoragem é o que garante o melhor resultado no que se refere ao cálculo do desenvolvimento do conhecimento ao longo das séries, entretanto é o mais difícil de

ser implementado devido à dificuldade de se construir um teste alinhado à matriz de avaliação em todas as séries e, ao mesmo tempo, que se ajuste bem para toda a população. Fica evidente que quanto mais distantes estiverem os períodos de escolaridade, mais difícil se torna o processo de *linkagem*.

1.3 ESCALONAMENTO

Nos processos de avaliação em larga escala, após a coleta de dados e a aplicação do método de *linkagem*, o passo seguinte é a construção de uma escala, ou seja, o escalonamento, que é o processo de transformação dos escores brutos, encontrados por meio da Teoria Clássica dos Testes (TCT), ou habilidades (TRI), em escores de escala. O principal objetivo do escalonamento é facilitar a interpretação dos resultados do teste aos usuários.

Normalmente, a escala é estabelecida usando uma única forma de teste e, para as formas subsequentes de testes, a escala é mantida através dos procedimentos de *linkagem* abordados na seção 1.1. Dessa forma, a escala permanece com o mesmo significado independente da forma de teste aplicado e do grupo testado. Tipicamente, escores brutos de uma nova forma são linkados aos escores brutos de uma velha forma e os resultados assim obtidos são convertidos em escores de escalas, utilizando-se transformações lineares ou não-lineares.

2 HISTÓRICO DAS AVALIAÇÕES EM LARGA ESCALA ATRAVÉS DA TEORIA DA RESPOSTA AO ITEM

Dedicaremos este capítulo ao processo evolutivo das avaliações educacionais tal qual a conhecemos hoje, fazendo um histórico sobre sua origem nos EUA no ano de 1925 e seu surgimento e evolução no Brasil a partir de 1995, com as avaliações do SAEB.

As abordagens aqui apresentadas são fundamentais para o entendimento de nossa preocupação com os rumos das *linkagens* realizadas no Brasil e que, em grande parte, são referências para as análises deste trabalho.

2.1 ANTECEDENTES

Podemos considerar como a primeira iniciativa no sentido de se construir um modelo de construção de escala nos moldes da TRI, o trabalho realizado por Louis Leon Thurstone em 1925 intitulado "Um método para escalonamento de testes psicológicos e educacionais". Thurstone solucionou um grande problema para a época: qual a melhor forma de se colocar os itens de Binet e Simon (1905), desenvolvidos para mensurar o desenvolvimento mental das crianças em uma escala graduada em função da idade?

Com a intenção de compararmos o modelo proposto por Thurstone e o modelo atual da TRI, o que em muito contribui para um bom entendimento das características e evolução da TRI, reproduzimos estes dois modelos utilizando uma base de dados da avaliação do Sistema Permanente da Avaliação do Ceará – SPAECE, no ano de 1998.

A base de dados do SPAECE apresenta o desempenho dos alunos em Língua Portuguesa em três períodos de escolaridade: 9º ano do Ensino Fundamental (alunos com 14 anos de idade), 1ª e 3ª séries do Ensino Médio (alunos com 15 e 17 anos de idade, respectivamente). Ao todo estaremos verificando o desempenho de 12.991 alunos levando-se em consideração apenas 15 itens comuns aos três períodos de escolaridades mencionados[1].

[1] Itens específicos de um único período de escolaridade foram retirados da análise.

O escalonamento proposto por Thustone, tal qual apresentado por Bock (1997) em *A brief history of item response theory* apresenta duas análises: i) percentual de alunos que responderam corretamente a cada item em função da idade e ii) posicionamento dos itens, em função de sua dificuldade média, na mesma escala do percentual de acerto.

Para a primeira análise, utilizaremos as informações contidas no quadro 2.1 e seu respectivo gráfico 2.1 apresentados a seguir.

ITEM	IDADE		
	14	15	17
1	0.537	0.575	0.592
2	0.514	0.532	0.590
3	0.713	0.748	0.778
4	0.508	0.544	0.575
5	0.644	0.720	0.762
6	0.379	0.394	0.426
7	0.463	0.575	0.640
8	0.566	0.625	0.660
9	0.478	0.553	0.572
10	0.296	0.366	0.429
11	0.613	0.650	0.699
12	0.586	0.660	0.685
13	0.435	0.511	0.554
14	0.288	0.312	0.386
15	0.598	0.643	0.691

Quadro 2.1: Percentual de resposta correta por item em função da idade dos alunos. Fonte SPAECE 1998.

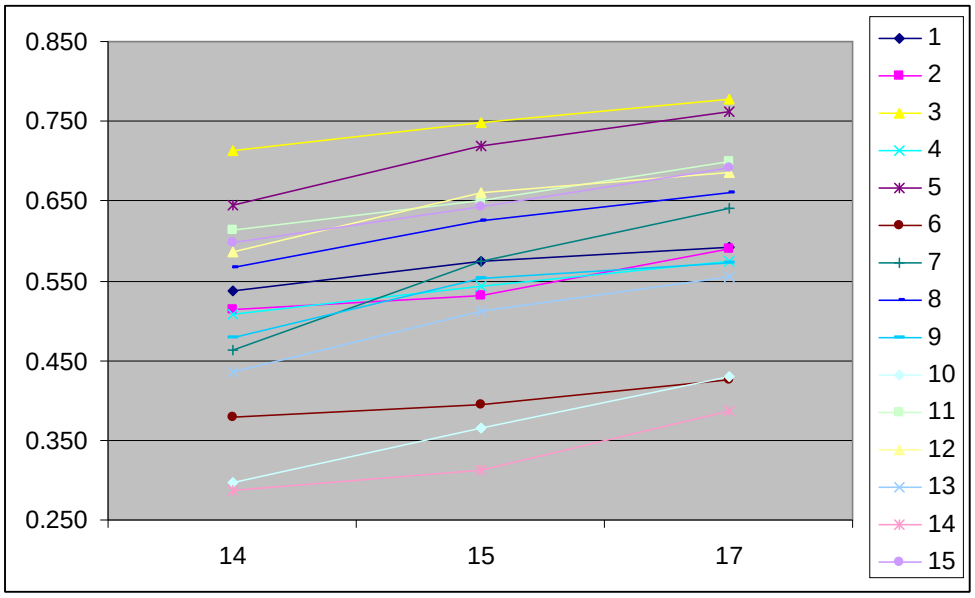

Gráfico 2.1: Percentual de resposta correta por item em função da idade dos alunos. Fonte SPAECE 1998.

Podemos observar através do gráfico 2.1 uma tendência monotônica crescente do percentual de acerto dos itens em função da idade do aluno. Cabe ressaltar que se tivéssemos estes itens sendo respondidos em mais faixas etárias, no caso apresentado por Thurstone em 1925, os itens eram respondidos por crianças de 3 a 15 anos de idade, estas curvas seriam mais parecidas com a curva característica do item obtida através da TRI.

Para a segunda análise, nos moldes apresentados por Thurstone, transformamos a distribuição do percentual de acerto dos alunos de 14, 15 e 17 anos de idade, de forma que os alunos de 14 anos passaram a ter uma distribuição normal (0,1), apresentamos no quadro 2.2, a seguir, os novos valores das médias e desvios-padrão dos acertos aos 15 itens, para as idades de 14, 15 e 17 anos nessa escala transformada ou escala padrão:

IDADE	escala original		escala transformada		alunos
	percentual médio de acerto	desvio padrão	percentual médio de acerto	desvio padrão	
14	0.501	0.216	0	1	4921
15	0.590	0.225	0.41	1.04	4878
17	0.620	0.228	0.55	1.057	3192

Quadro 2.2: Valores dos percentuais médios de acerto e desvios padrões dos 3 grupos, na escalas original e padronizada. Fonte SPAECE 1998.

Em seguida, calculamos a dificuldade dos itens na escala padronizada através da expressão abaixo:

Bi_padronizado = (Pi – média14)/dp14 (2.1)

Onde,

Bi_padronizado - Dificuldade do item i na escala padronizada

Pi - percentual médio de acerto original do item i nos 3 grupos

média14 - Média dos percentuais de acerto dos alunos de 14 anos na escala original= 0.501

dp14- desvio padrão dos percentuais de acerto dos alunos de 14 anos na escala original = 0.216

Os resultados destas transformações estão apresentados no quadro 2.3:

item	Dificuldade escala transformada
1	0.295
2	0.179
3	1.116
4	0.169
5	0.925
6	-0.489
7	0.216
8	0.508
9	0.128
10	-0.680
11	0.680
12	0.633
13	-0.039
14	-0.837
15	0.633

Quadro 2.3: Dificuldade dos itens na escala padronizada. Fonte SPAECE 1998.

Por meio dos procedimentos acima, construímos o gráfico 2.2, no qual constatamos a grande inovação proposta por Thurstone, ou seja: posicionar os itens na mesma métrica do desempenho dos alunos. Desta forma podemos fazer uma análise critica do quanto os itens estão alinhados com o desempenho dos alunos. Neste estudo, verificamos uma boa distribuição dos itens ao longo da escala de desempenho dos alunos, embora haja uma concentração de itens em torno dos valores 0.2 e a.6 e poucos itens nos extremos da escala, o que de certa forma está coerente com a base de dados utilizada, pois nos casos de avaliação em larga escala, na elaboração dos testes há uma concentração de itens em torno da média, onde a precisão da medida precisa ser maior e poucos itens nos extremos da escala, onde temos poucos alunos.

Através deste estudo, fica evidente que para uma boa medida do conhecimento do examinando, não é suficiente verificar a quantidade de itens que o mesmo acerta, como estipulado pela Teórica Clássica dos Testes – TCT, mas sim, quais itens, em função de sua dificuldade que o mesmo acerta ou erra. Para tanto cabe aos elaboradores de testes especificar uma quantidade suficiente de itens bem distribuídos ao longo da escala de conhecimento.

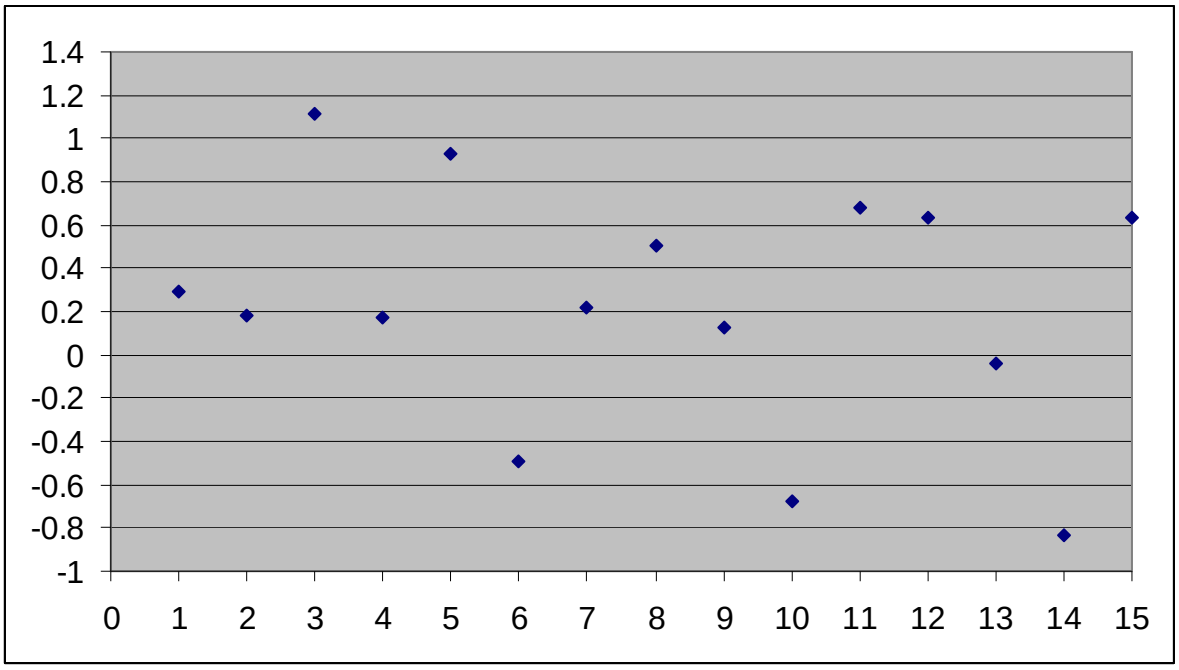

Gráfico 2.2: Posicionamento dos itens em função de sua dificuldade média na mesma escala do percentual de acerto reescalonada para uma normal (0,1) para os alunos de 14 anos. Fonte SPAECE 1998.

Para a análise pela TRI utilizamos o software BILOGMG, através do processamento de calibração simultânea utilizando o modelo logístico de 3 parâmetros. Para o cálculo dos parâmetros dos itens utilizamos o método MMAP e para as proficiências o método EAP. Os detalhes dos modelos matemáticos da TRI são abordados no capítulo III. Desta forma obtivemos as proficiências dos alunos e parâmetros dos itens em uma mesma escala onde o grupo de referência, ou seja, o grupo com uma distribuição normal padronizada são os alunos de 14 anos. Nos quadros 2.4 e 2.5 abaixo apresentamos estes valores:

IDADE	escala padronizada		alunos
	proficiência média	desvio padrão	
14	0	1	4921
15	0.415	1.081	4878
17	0.562	1.070	3192

Quadro 2.4: Médias de acerto e desvios-padrão das proficiências dos 3 grupos, na padronizada. Fonte SPAECE 1998

ITEM	PARÂMETROS		
	A	B	C
1	.436	-.146	.008
2	.552	.498	.226
3	.877	-.630	.028
4	.581	.320	.161
5	.586	-.613	.031
6	.844	.984	.201
7	1.070	.234	.160
8	.680	-.157	.067
9	.707	.495	.225
10	1.110	.854	.109
11	.759	.014	.211
12	.645	-.117	.124
13	.379	.569	.118
14	.698	1.620	.204
15	.430	-.366	.046

Quadro 2.5: Parâmetros dos itens pela TRI. Fonte SPAECE 1998.

No gráfico 2.3 apresentamos as curvas características dos itens para os itens considerados. Diferente do modelo de Thurstone, onde as curvas dos itens são discretas, pois possuem pontos apenas nas idades dos respondentes, no modelo da TRI temos curvas contínuas, o que se ajusta melhor à realidade, pois alunos com a mesma idade possuem diferentes habilidades e proficiências.

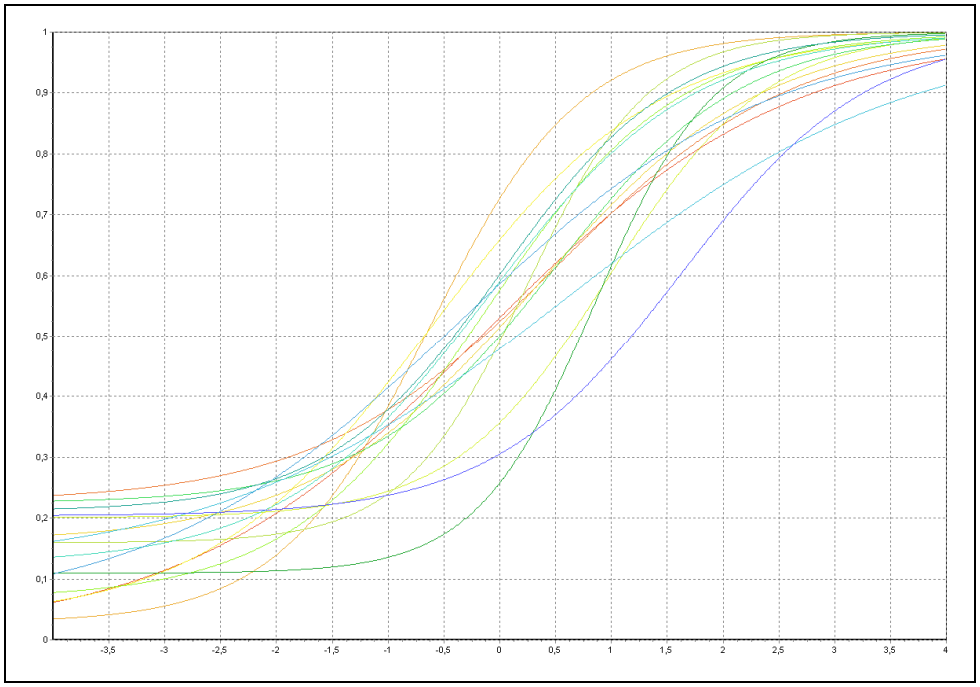

Gráfico 2.3: Curvas Características dos Itens – CCI. Fonte: SPAECE 1998.

Analogamente ao modelo de Thurstone, também podemos posicionar os itens, em função de sua dificuldade, representada pelo parâmetro *B*, na mesma escala da proficiência, conforme apresentado no gráfico 2.4.

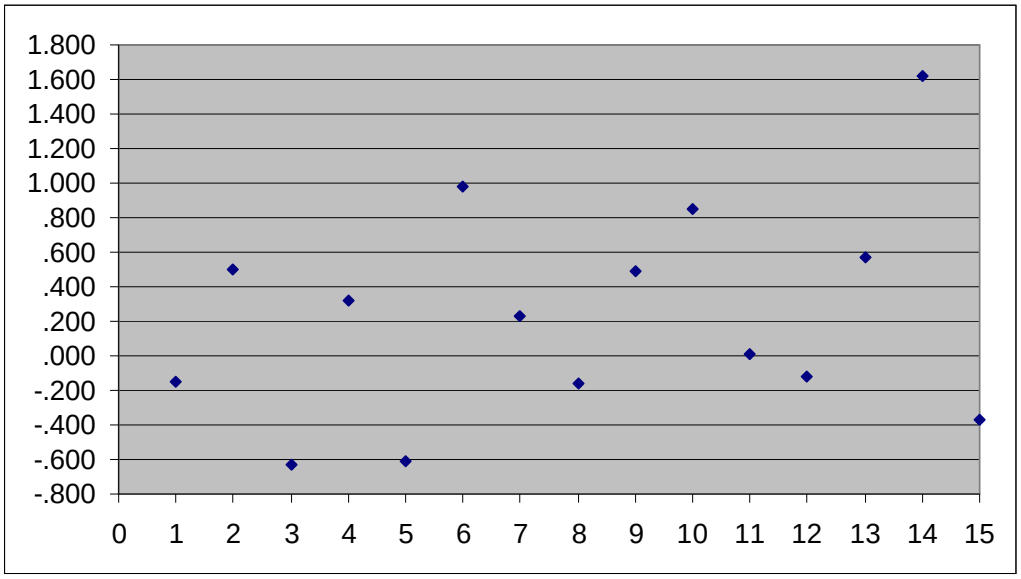

Gráfico 2.4: Posicionamento dos itens em função do parâmetro *B* na mesma escala da proficiência para uma normal (0,1) para os alunos de 14 anos. Fonte: SPAECE 1998.

Entretanto, pela TRI, conseguimos uma análise mais significativa, através da curva de informação do teste, conforme pode ser verificado no gráfico 2.5. Neste gráfico, observamos que o teste possui a melhor informação, ou seja, possui estimativas mais precisas para alunos com proficiência no intervalo compreendido entre -1,8 e 2,6.

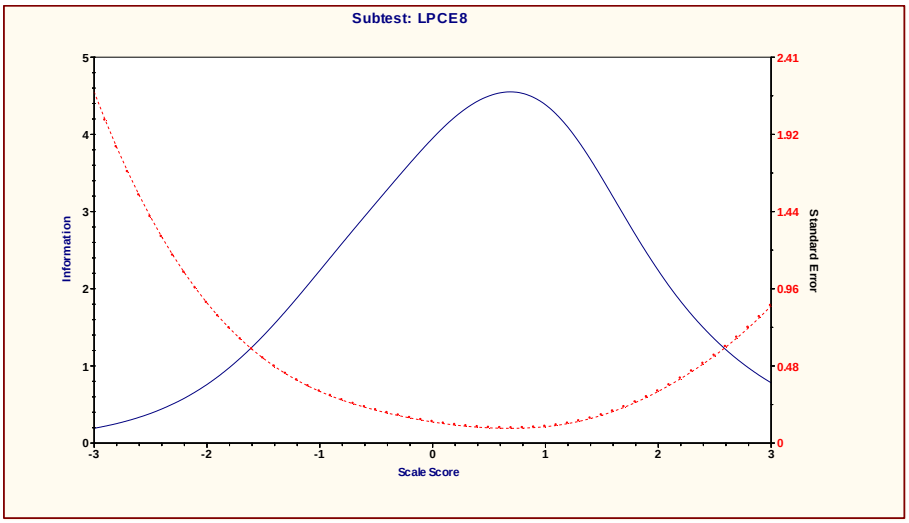

Gráfico 2.5: Curva de informação do teste. Fonte: SPAECE 1998.

Pelas análises apresentadas, verificamos similaridades entre os dois modelos e podemos realmente concluir que o modelo de Thurstone está mais próximo da TRI que os procedimentos da TCT. Entretanto, o fato do Modelo de Thurstone trabalhar com variável observável, no caso a idade dos alunos, não faz com que seu modelo resolva o problema da influência do teste na medida. Somente no início dos anos de 1950 que o psicometristas começaram a resolver esse problema com a introdução do traço latente de Lazersfeld (1950).

Porém, a origem da TRI moderna surgiu com os trabalhos de Lord (1952, 1953), os quais definiram a teoria e o modelo matemático para os cálculos dos parâmetros dos itens utilizando o modelo da ogiva normal para testes com itens do tipo dicotômico.

A partir da década de 1950, uma série de trabalhos matemáticos foram desenvolvidos dando origem aos modelos matemáticos para tratamento de dados dentro da TRI. Dentre estes trabalhos podemos citar o modelo de Samejima (1969, 1972) para itens politômicos e dados contínuos, utilizados em testes de personalidade e o trabalho de Birnbaum (1957) que ao substituir no modelo da TRI as curvas de ogiva por curvas logísticas, tornou as análises matemáticas de tratamento dos dados bem mais fácil.

Embora os modelos matemáticos para trabalhar com a TRI já estivessem consolidados no final da década de 1950, somente na década de 1980, com o desenvolvimento da informática e desenvolvimento de *softwares* específicos que a TRI teve sua expansão e estabelecimento como instrumento de medida dentro da psicometria.

O primeiro *software* para as análises da TRI foi o BICAL desenvolvido, em 1979, por Wright, Mead e Bell, em seguida, surgiram o LOGIST (Wingersky, Barton e Lord, 1982), BILOG (Mislevy e Bock, 1984) e o BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996).

Paralelamente à evolução dos aspectos técnicos que culminaram para o estabelecimento da TRI como modelo estatístico no campo da Psicologia e da Educação, podemos ressaltar a substituição dos modelos de testes do tipo ensaio, com itens escritos, por testes de múltipla escolha, o que facilitou a coleta de dados, principalmente com o surgimento em 1936 da primeira máquina de leitora de cartões criada pela IBM, denominada Type 80, cuja primeira utilização em

larga foi feita em 1936 pela Americam Council on Education´s Cooperative Test Service, a qual deu origem em 1947 à Educational Testing Service – ETS[2],

Atualmente, utilizam-se máquinas de *scanners* para processamento dos cartões de respostas, as quais geram arquivos digitais com as respostas dos respondentes possibilitando uma nova tendência nas avaliações, que é a utilização de itens em que os respondentes escrevem as respostas nos testes, as quais são capturadas e enviadas para corretores através de meios eletrônicos, tornando possível a correção dos itens com agilidade e confiabilidade. De certa forma, estamos voltando à concepção inicial de medição de conhecimento através de itens abertos, porém com equipamentos e técnicas estatísticas mais apropriadas. Parece contraditório, mas a evolução da tecnologia que provocou praticamente a eliminação dos itens abertos dos processos avaliativos na década de 1930 está direcionando para uma tendência à volta desse tipo de avaliação.

[2] A ETS é, atualmente, o maior centro educacional privado do mundo, desenvolvendo, administrando e contabilizando mais de 24 milhões de testes todos os anos, em mais de 180 países. Nos Estados Unidos, é a instituição responsável pelo NAEP e também pela aplicação do Scolastic Aptitude Test ou Scholastic Assessment Test – SAT, teste utilizado para avaliação de todos os estudantes que buscam o ingresso na universidade.

2.2 CARACTERÍSTICAS DAS AVALIAÇÕES EM LARGA ESCALA NO BRASIL

Até 1993, o SAEB utilizou a Teoria Clássica de Testes (TCT) para a construção dos instrumentos, atribuição dos escores e análise dos resultados, não havendo planejamento para uma comparação dos resultados. A partir de 1995, o SAEB introduz a Teoria de Resposta ao Item (TRI), passando a ter as seguintes características:

- Avaliações amostrais com representatividade de agregação de resultados para todos os Estados brasileiros.
- Participação das redes de ensino estaduais, municipais, federais e particulares.
- Avaliações em Língua Portuguesa e em Matemática nas 4ª e 8ª séries do Ensino Fundamental e 3° ano do Ensino Médio.
- Criação de escalas de habilidades para Língua Portuguesa e para Matemática, através da técnica estatística da Teoria da Resposta ao Item, tendo como referência a 8ª série do Ensino Fundamental de 1997 com média 250 pontos e desvio padrão de 50 pontos, garantindo, portanto, a comparabilidade de resultados entre anos avaliados. Essa média e esse desvio padrão são a referência de escala de habilidades.
- Avaliações realizadas a cada dois anos: 1995, 1997, 1999, 2001, 2003, 2005, 2007 e 2009.

Em 2005, procurando um mapeamento maior da Educação Básica, foi instituída a Prova Brasil, com característica censitária, avaliando todos os alunos, apenas da rede pública, nas disciplinas de Matemática e de Língua Portuguesa, na 4ª e 8ª séries do Ensino Fundamental de oito anos. O Ministério da Educação - MEC, através do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira -INEP, elabora, aplica e entrega os resultados da Prova Brasil, cabendo às escolas a participação na aplicação dos testes e o devido uso de seus resultados, tornando essa avaliação um importante instrumento para gestão dentro de cada unidade escolar.

2.2.1 Diferenças entre o Saeb e a Prova Brasil

Verificamos, no Quadro 2.6, as principais características dessas duas avaliações que vêm ocorrendo de forma simultânea, no Brasil, a partir de 2005:

Prova Brasil	SAEB
A prova foi criada em 2005.	A primeira aplicação ocorreu em 1990.
Sua primeira edição foi em 2005, e em 2007 houve nova aplicação.	É aplicado de dois em dois anos. A última edição foi em 2005. Em 2007 houve nova prova.
A Prova Brasil avalia as habilidades em Língua Portuguesa (foco em leitura) e Matemática (foco na resolução de problemas)	Alunos fazem prova de Língua Portuguesa (foco em leitura) e Matemática (foco na resolução de problemas)
Avalia apenas estudantes de ensino fundamental, de 4ª e 8ª séries.	Avalia estudantes de 4ª e 8ª séries do ensino fundamental e também estudantes do 3º ano do ensino médio.
A Prova Brasil avalia as escolas públicas localizadas em área urbana.	Avalia alunos da rede pública e da rede privada, de escolas localizadas nas áreas urbana e rural.
A avaliação é quase universal: todos os estudantes das séries avaliadas, de todas as escolas públicas urbanas do Brasil com mais de 20 alunos na série, devem fazer a prova.	A avaliação é amostral, ou seja, apenas parte dos estudantes brasileiros das séries avaliadas participam da prova.
Por ser universal, expande o alcance dos resultados oferecidos pelo Saeb. Como resultado, fornece as médias de desempenho para o Brasil, regiões e unidades da Federação, para cada um dos municípios e escolas participantes.	Por ser amostral, oferece resultados de desempenho apenas para o Brasil, regiões e unidades da Federação.
Aplicação em 2007: 5 a 20 de novembro.	Aplicação em 2007: 5 a 20 de novembro.
Parte das escolas que participarem da Prova Brasil ajudará a construir também os resultados do Saeb, por meio de recorte amostral.	Todos os alunos do Saeb e da Prova Brasil farão uma única avaliação.

Quadro 2.6: Características Prova Brasil e SAEB até o ano de 2007. Fonte: www.inep.gov.br

2.2.2 Características do SAEB e da Prova Brasil

Como um dos objetivos do SAEB foi a criação de uma escala de conhecimento para o Brasil, nas disciplinas de Língua Portuguesa e Matemática, utilizou-se uma estrutura de Blocos Incompletos Balanceados (BIB), na construção dos testes. Esta estrutura de montagem possibilita a aplicação de uma grande quantidade de itens, permitindo aos especialistas das disciplinas avaliadas a construção e interpretação das escalas de habilidades. A montagem dos blocos nos cadernos segue uma estrutura em espiral, com blocos comuns entre os cadernos de forma a possibilitar a *linkagem* dos mesmos.

Até 2005, o BIB utilizado pelo SAEB era composto por 26 modelos diferentes de cadernos por disciplina, sendo que cada caderno era composto por 3 blocos de 13 itens. A posição dos diferentes blocos nos cadernos é apresentada no quadro 2.7.

CADERNO	BLOCOS		
	POS1	POS2	POS3
1	1	2	5
2	2	3	6
3	3	4	7
4	4	5	8
5	5	6	9
6	6	7	10
7	7	8	11
8	8	9	12
9	9	10	13
10	10	11	1
11	11	12	2
12	12	13	3
13	13	1	4

CADERNO	BLOCOS		
	POS1	POS2	POS3
14	1	3	8
15	2	4	9
16	3	5	10
17	4	6	11
18	5	7	12
19	6	8	13
20	7	9	1
21	8	10	2
22	9	11	3
23	10	12	4
24	11	13	5
25	12	1	6
26	13	2	7

Quadro 2.7: BIB de 26 cadernos. Fonte: Arquivos SAEB.

A partir de 2007, tanto o SAEB quanto a Prova Brasil passaram a adotar um mesmo BIB, composto por 21 modelos diferentes de cadernos, sendo cada caderno composto por 4 blocos de itens e 2 disciplinas, conforme mostra o quadro 2.8.

cadernos ímpares						cadernos pares				
caderno	blocos					caderno	blocos			
	lp	mat	lp	mat			mat	lp	mat	lp
1	1	1	2	2		2	2	2	3	3
3	3	3	4	4		4	4	4	5	5
5	5	5	6	6		6	6	6	7	7
7	7	7	1	1		8	1	1	3	3
9	2	2	4	4		10	3	3	5	5
11	4	4	6	6		12	5	5	7	7
13	6	6	1	1		14	7	7	2	2
15	1	1	4	4		16	2	2	5	5
17	3	3	6	6		18	4	4	7	7
19	5	5	1	1		20	6	6	2	2
21	7	7	3	3						

Quadro 2.8: BIB de 21 cadernos. Fonte: Arquivos SAEB.

Faremos, a seguir, uma descrição mais detalhada desses designs utilizados ao longo das aplicações do SAEB e Prova Brasil, onde poderemos observar três momentos distintos de metodologias empregadas na construção dos testes por esses dois sistemas de avaliação.

1º momento: Design SAEB até 2005

O aluno era avaliado em apenas uma disciplina, ou Língua Portuguesa ou Matemática, havia, portanto, um design para Língua Portuguesa e outro para Matemática, conforme apresentado no quadro 2.9.

Série	Língua Portuguesa					Matemática				
	Nº de blocos	Itens por bloco	Itens por caderno	Nº de cadernos	Total de itens	Nº de blocos	Itens por bloco	Itens por caderno	Nº de cadernos	Total de itens
5 EF	13 lp	13	39	26	169	13 mat	13	39	26	169
9 EF	13 lp	13	39	26	169	13 mat	13	39	26	169
3 EM	13 lp	13	39	26	169	13 mat	13	39	26	169

Quadro 2.9: Design de montagem dos Blocos de itens nas versões do SAEB até 2005. Fonte: Arquivos SAEB.

2º momento: Design Prova Brasil 2005

Por meio desse design, em um mesmo teste, o aluno foi avaliado em Língua Portuguesa e em Matemática, conforme podemos observar no quadro 2.10.

Língua Portuguesa e Matemática					
Série	Nº de blocos	Itens por bloco	Itens por caderno	Nº de cadernos	Total de itens
5 EF	7 lp e 7 mat	10	40	21	70 lp e 70 mat
9 EF	7 lp e 7 mat	12	48	21	84 lp e 84 mat

Quadro 2.10: Design de montagem dos Blocos de itens na Prova Brasil 2005. Fonte: Arquivos SAEB

Verificamos, neste design, que os cadernos ímpares começaram com 2 blocos de Língua Portuguesa e terminaram com 2 blocos de Matemática, enquanto que, nos cadernos pares, a montagem das disciplinas foi invertida, ou seja, começaram com 2 blocos de Matemática e terminaram com 2 blocos de Língua Portuguesa.

3º momento: Design Prova Brasil 2007 e SAEB 2007

A partir de 2007, o SAEB mudou o design de seus testes, passando a utilizar, assim como na Prova Brasil, cadernos de testes com as duas disciplinas juntas. Ressaltamos que o design do SAEB, assim como os itens e cadernos de testes, passaram a ser os mesmo da Prova Brasil, apenas com a diferença que no SAEB é avaliado o 3º ano do Ensino Médio e na Prova Brasil esta série não é avaliada.

Também houve uma mudança nesse design com relação ao utilizado em 2005. Conforme podemos observar nos quadros 2.11 e 2.12 o número de itens aumentou no 5º ano e a disposição dos blocos de Língua Portuguesa e de Matemática nos cadernos foi alterada com relação à versão anterior.

Língua Portuguesa e Matemática					
Série	Nº de blocos	Itens por bloco	Itens por caderno	Nº de cadernos	Total de itens
5 EF	7 lp e 7 mat	11	44	21	77 lp e 77 mat
9 EF	7 lp e 7 mat	13	52	21	91 lp e 91 mat

Quadro 2.11: Design de montagem dos Blocos de itens na Prova Brasil 2007. Fonte: Arquivos SAEB

Língua Portuguesa e Matemática					
Série	Nº de blocos	Itens por bloco	Itens por caderno	Nº de cadernos	Total de itens
5 EF	7 lp e 7 mat	11	44	21	77 lp e 77 mat
9 EF	7 lp e 7 mat	13	52	21	91 lp e 91 mat
3 EM	7 lp e 7 mat	13	52	21	91 lp e 91 mat

Quadro 2.12: Design de montagem dos Blocos de itens no SAEB 2007. Fonte: Arquivos SAEB

Nessas avaliações do SAEB e Prova Brasil há cadernos ímpares, começando com Língua Portuguesa e cadernos pares começando com Matemática, assim como no design da Prova Brasil de 2005, entretanto, diferentemente desta versão em que os blocos de uma mesma disciplina eram apresentados juntos, os blocos das disciplinas, no ano de 2007, foram alternados dentro do caderno, ou seja, nos cadernos ímpares os blocos foram montados na ordem LP/MAT/LP/MAT e nos cadernos pares a ordem foi MAT/LP/MAT/LP.

Alguns aspectos dessas diferenças de designs serão objeto de estudos no capítulo 5 desta dissertação, no que diz respeito à comparabilidade de resultados entre o Saeb, a Prova Brasil e os estados brasileiros, pois, em diferentes anos e projetos, os modelos adotados nessas instâncias não foram os mesmos.

2.2.3 Avaliações em larga escala nos estados brasileiros

Alguns estados brasileiros, dentre os quais destacamos Minas Gerais, Rio de Janeiro, Rio Grande do Sul, Mato Grosso do Sul, Bahia, Ceará, Paraná, Pernambuco e São Paulo, realizam avaliações censitárias de suas escolas, visando, principalmente, o direcionamento de políticas públicas no sentido de melhorar a qualidade de suas redes de ensino e a melhoria da prática docente.

Uma característica importante nessas avaliações estaduais, e não observada, nos EUA, como mencionado anteriormente, é a preocupação de comparabilidade desses resultados com os resultados do país. Para tanto, a parceria com o INEP, através de disponibilização de itens e bases de dados das avaliações do SAEB foram imprescindíveis.

A partir de 2005, observa-se, através da criação da Prova Brasil, a tendência do INEP avaliar, também de forma censitária, os alunos da rede pública de ensino, com o objetivo de fornecer maiores subsídios para os estados, municípios e escolas. Apesar disso, o que se observa é que alguns estados continuam com seus sistemas de avaliação, pois esses possuem algumas características diferentes da Prova Brasil e, a princípio, não querem abandonar toda uma metodologia empregada ao longo dos anos. Como, por exemplo:

- Minas Gerais, que possui uma série histórica constituída pelas avaliações nos anos de 2000, 2002, 2003, 2006, 2007, 2008 e 2009, aplicando em dias distintos, testes para Língua Portuguesa e para Matemática, para todas as escolas do estado independente do número de alunos. Diferenciando-se, portanto, da Prova Brasil que aplica Língua Portuguesa e Matemática no mesmo caderno de teste e não aplica testes em escolas com menos de 15 alunos.
- Rio Grande do Sul, cujo coorte distingui-se daquele pela Prova Brasil pois avalia a 5ª série/6º ano do Ensino Fundamental e 1º ano do Ensino Médio.
- Ceará, que, além do 5º e 9º anos do Ensino Fundamental, avalia todo o Ensino Médio, fornecendo os resultados por aluno avaliado, contrapondo-se à Prova Brasil cuja menor unidade de avaliação é a escola.

Além desses estados, Rio de Janeiro, Pernambuco, São Paulo, Espírito Santo e Bahia continuam com seus sistemas de avaliação.

3 MODELOS E MÉTODOS MATEMÁTICOS UTILIZADOS NA TRI

Existem inúmeras possibilidades de se expressar matematicamente a probabilidade de acerto de um indivíduo com uma certa habilidade a um determinado item. A definição do modelo matemático mais adequado a ser utilizado para esta finalidade vai depender das características dos testes. Assim, tais modelos são classificados quanto à dimensionalidade dos testes e à estrutura dos itens.

Com relação à dimensionalidade, os testes podem ser unidimensionais ou multidimensionais, dependendo se medem apenas uma habilidade ou traço latente do respondente, ou mais de uma habilidade, respectivamente.

Quanto à estrutura dos itens, estes podem ser do tipo dicotômico ou politômico. Os dicotômicos são itens de múltipla escolha com duas possibilidades de respostas, certo ou errado, envolvendo geralmente 4 ou 5 opções de respostas. Já para os itens politômicos, não há uma única opção ou situação como correta e sim diferentes possibilidades de acerto ou concordância. Os itens politômicos podem ser trabalhados em escalas do tipo nominal, sem uma ordem de grandeza entre as opções de respostas ou em escalas graduadas, onde as respostas podem ser ordenadas da mais errada à mais correta ou em níveis de concordância, por exemplo, utilizando escala do tipo Likert. Algumas escalas graduadas são também acumulativas, ou seja, para se atingir um nível mais alto de acerto é necessário possuir os conhecimentos dos níveis anteriores.

Iremos, nesse capítulo, restringir nossos estudos aos modelos logísticos da TRI e métodos de estimação de parâmetros de itens e escores, envolvendo testes unidimensionais compostos por itens dicotômicos, pois esta é a característica das avaliações em larga escala, realizadas pelo SAEB e pelos estados brasileiros, com as quais iremos trabalhar realizando estudos de casos, mais especificamente no capítulo 5 desta dissertação.

3.1 MODELOS LOGÍSTICOS UTILIZADOS NA TRI

Nessa situação de unidimensionalidade envolvendo itens dicotômicos, a relação entre a probabilidade de acerto ao item e a habilidade do respondente pode ser expressa por uma relação linear ou não-linear.

No caso de uma relação linear, como apresentado no gráfico 3.1, verificamos que a probabilidade de se acertar ao item é representada por uma reta.

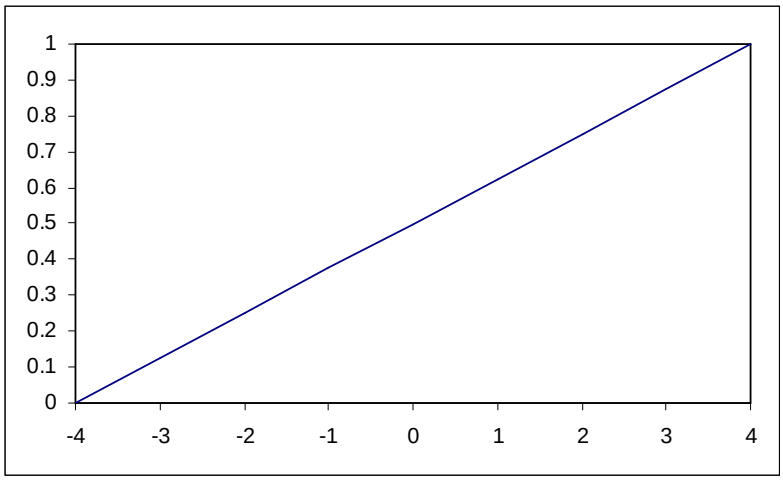

Gráfico 3.1 – Relação entre habilidade e probabilidade de acerto em um modelo linear.

Escrevendo através de uma equação matemática, temos a seguinte função:

$$P_i(\theta) = a_i\theta + b_i \qquad (3.1)$$

Onde,

$P_i(\theta)$ – probabilidade do indivíduo com habilidade θ acertar ao item i

θ – nível de habilidade do indivíduo

a_i – coeficiente de inclinação da reta

b_i – dificuldade do item i

Já no caso de uma representação matemática através de um modelo não-linear, uma opção seria utilizar uma função ogiva, dada pela equação:

$$P_i = e^{(\theta - b_i)} \qquad (3.2)$$

Sendo e = 2,72

Observamos no gráfico 3.2 que para este caso, a probabilidade de se acertar ao item apresenta a forma de um “s” (uma signóide):

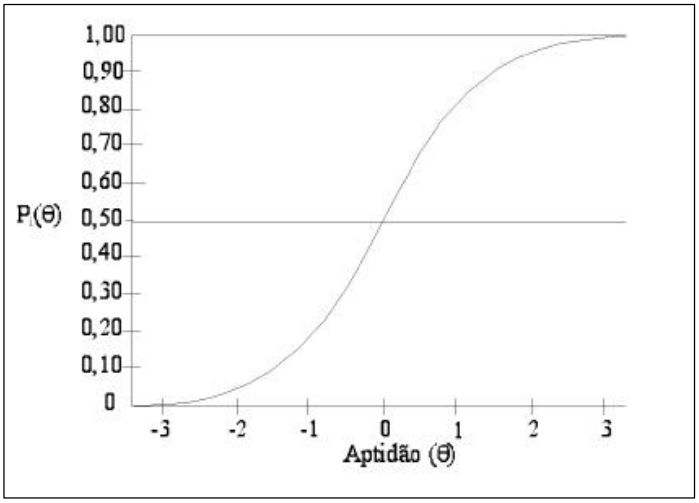

Gráfico 3.2: Relação entre habilidade e probabilidade de acerto em um modelo não-linear. Fonte: CAEd

Constatamos que, embora esse segundo modelo seja mais complexo que o primeiro e que pelo princípio da parcimônia na matemática, este seria preterido em função do primeiro, isto não acontece, pois este modelo representa melhor a realidade, haja vista que, no modelo de uma reta, a probabilidade de acerto aumentaria indefinidamente com o aumento da habilidade e sabemos que isto não acontece, pois o item tem uma região de saturamento, ou seja, a partir de determinada habilidade do respondente a probabilidade de acerto tende a se estabilizar.

Devido a este melhor ajuste a função ogiva é empregada na TRI e recebe o nome de curva característica do item – CCI.

Essa função, pode ser representada tanto por um modelo normal, quanto por um modelo logístico. No caso do modelo normal a função toma a forma a seguir:

$$P = \int_{-\infty}^{z} \frac{1}{2\pi} e^{\frac{-z^2}{2}} dz \qquad (3.3)$$

Uma vez que a solução desta equação é complexa, pois envolve a utilização de uma integral, os estatísticos utilizam modelos de soluções mais simples, que são os modelos logísticos.

São três os modelos logísticos utilizados na TRI, os quais se distinguem entre si pelo número de parâmetros utilizados para descrever o item. O primeiro modelo logístico foi proposto por Birnbaum (1968), no qual podemos distinguir em sua formulaçao a presença de dois parâmetros , dando origem ao modelo logístico de dois parâmetros 2 – 2 LP, conforme apresentado na equação 3.4 abaixo:

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} \qquad (3.4)$$

e= constante =2,72

b= parâmetro de dificuldade do item

a= parâmetro de discriminação do item

θ= habilidade

a(θ-b)= logit

A fim de se aproximar melhor os resultados do modelo logístico ao modelo normal, multiplica-se o fator logit por uma constante D, igual a 1,7:

$$\text{Logit} = Da(\theta - b) \qquad (3.5)$$

A inclusão deste fator torna as diferenças de resultados entre os modelos normal e logístico, menores que 0.01 para todos os valores de θ. Assim, o modelo logístico para dois parâmetros passa a ser representado pela seguinte expressão:

$$P(\theta) = \frac{1}{1 + e^{-aD(\theta - b)}} \qquad (3.6)$$

O segundo modelo logístico, modelo de 1 parâmetro – 1 LP, foi desenvolvido por Wright (1977) e como se trata de uma versão logística de um modelo linear proposto por Rasch (1960), o mesmo é também conhecido como modelo Rasch. Sua representação é praticamente a mesma do

modelo de 2 parâmetros, com a diferença que agora o parâmetro *a* é o mesmo para todos os itens, geralmente este valor é fixado em um. Substituindo o valor de *a* por 1 na equação 3.3, temos a equação 3.7 que representa esse modelo:

$$P(\theta) = \frac{1}{1 + e^{-1D(\theta - b)}} \qquad (3.7)$$

O terceiro modelo logístico foi desenvolvido por Lord (1980), o qual é denominado de modelo logístico de 3 parâmetros – 3 LP. Neste modelo, temos a inclusão de um terceiro parâmetro na equação 3.6, que é o parâmetro *c*, que representa o acerto ao acaso (chute), ou seja, a probabilidade de um sujeito acertar ao item tendo uma habilidade muito baixa. A representação para este modelo é dada na equação 3.8:

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-aD(\theta - b)}} \qquad (3.8)$$

No gráfico 3.3 apresentamos a curva característica de um item com três parâmetros, na qual podemos observar o parâmetro *c* como sendo a assíntota inferior dessa curva.

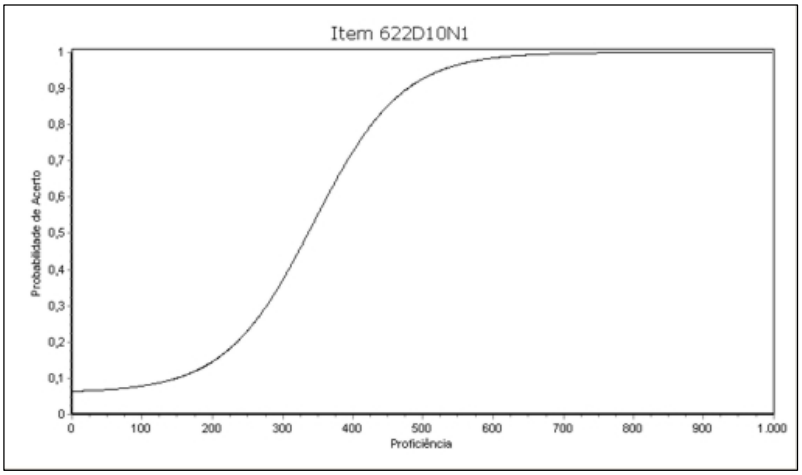

Gráfico 3.3: Curva característica do item em um modelo de três parâmetros. Fonte: CAEd

Dos três modelos apresentados, o modelo de 3 parâmetros – 3 LP, é o mais utilizado nas avaliações realizadas no Brasil e será, portanto, o modelo utilizado em nossas análises.

Assim, tendo por referência uma formulação mais completa para a equação 3.8, temos que a probabilidade de acertar um item, dada a aptidão teta do sujeito, a dificuldade do item, discriminação do item e o acerto aleatório do item é dada pela equação 3.9:

$$P(Xi = 1/\theta, bi, ai, ci) = ci + (1 - ci)\frac{1}{1 + e^{-aiD(\theta - bi)}} \qquad (3.9)$$

3.2 MÉTODOS DE ESTIMAÇÃO

Até o presente momento, apresentamos somente situações envolvendo probabilidades de acerto a um único item quando conhecemos a habilidade do sujeito e os parâmetros de itens. No entanto, na prática deparamo-nos com situações em que é preciso calcular proficiências de vários alunos que respondem a uma grande quantidade de itens em basicamente quatro situações distintas: i) estimar proficiências quando os parâmetros dos itens são conhecidos, ii) estimar parâmetros quando as proficiências são conhecidas, iii) estimar parâmetros e proficiências quando ambos são totalmente desconhecidos e iv) estimar parâmetros e proficiências quando as proficiências são desconhecidas e apenas parte dos itens têm parâmetros desconhecidos. As duas últimas situações são as mais complexas de todas e também as mais comumente encontrada nas avaliações educacionais, sendo que a quarta situação, além de envolver os métodos de estimação, já exige que os resultados estejam linkados a uma escala já existente mais estritamente relacionada aos métodos de *linkagens* apresentados no capítulo 4. Portanto, vamos nos concentrar nos métodos matemáticos utilizados no contexto das três primeiras situações, envolvendo apenas um único grupo de respondentes.

Uma vez selecionado um dos modelos do tópico anterior, em nossos estudos trabalharemos com o 3 LP, utilizaremos a base de dados com as respostas dadas pelos alunos ao conjunto de itens que compõem o teste, assim, se temos N alunos respondendo a n itens, obtemos uma matriz de respostas conforme apresentado no quadro 3.1.

Aluno	Item				
	1	2	3	...	n
1	1	0	0	...	0
2	1	1	0	...	0
3	1	1	1	...	0
⋮	⋮	⋮	⋮	...	0
N	1	1	1	...	1

Quadro 3.1 – Matriz de respostas de N alunos a n itens.

O conjunto de respostas dadas por cada aluno é denominado um vetor de respostas. Portanto, se n for igual a 10, podemos ter o seguinte vetor de respostas:

1111011001

Onde, 1 o aluno acertou o item e 0 o aluno errou o item

Para determinar qual a proficiência que melhor se ajusta para todos os padrões de respostas e/ou os parâmetros de discriminação, dificuldade e chute de cada um dos *n* itens do teste, podemos utilizar tanto métodos de máxima verossimilhança como métodos bayesianos. Ambos os métodos envolvem cálculos bastante complexos e recursos computacionais adequados. De fato, conforme Baker (2001) “até o desenvolvimento dos computadores e sua relativa facilidade de utilização, a partir da década de 1990, a TRI não era um método prático devido à demanda de recursos computacionais adequados, e consequentemente de *softwares* específicos”.

Atualmente, existe uma certa variedade de *softwares* disponíveis no mercado para se trabalhar com a TRI, sendo que no Brasil, o mais utilizado é o BILOG-MG produzido pela Scientific Software, Inc, Mislevy, R.J. e Bock, R. D (1990) , no qual encontramos diferentes possibilidades para os cálculos de parâmetros de itens e proficiências de alunos envolvendo variações dos métodos de máxima verossimilhança e de métodos bayesianos.

Descreveremos, agora, as principais características desses dois métodos e, em seguida, faremos uma descrição dos métodos e procedimentos utilizados no BILOG-MG nas avaliações nacionais e, consequentemente, em todos os estudos de casos apresentados neste trabalho.

3.2.1 Método de Máxima Verossimilhança - ML

Uma vez que temos os padrões de respostas, ou seja, sabemos se o aluno acertou ou errou o item, não faz sentido trabalharmos com probabilidades, e sim, com verossimilhança. Nesse caso, utilizamos a função de verossimilhança dada na equação 3.10, para relacionar padrões de respostas, habilidades de aluno e parâmetros de itens:

$$L(u_{1s}, u_{2s}, ..., u_{ns} \mid \theta) = \prod_{i-1}^{n} P_i(\theta_s)^{u_{si}} Q_i(\theta_s)^{1-u_{si}} \qquad (3.10)$$

Onde i =1, 2, ..., n itens

u_{is} = resposta do sujeito a cada item (1 = acertou, 0 = errou)

Considerando o modelo 3 LP, temos que:

$$Pi(\theta s) = ci + (1 - ci)\frac{1}{1 + e^{-aiD(\theta s - bi)}} \qquad (3.11)$$

$$Qi(\theta s) = (1 - ci) - (1 - ci)\frac{1}{1 + e^{-aiD(\theta s + bi)}} \qquad (3.12)$$

Temos, portanto, nesta equação que a verossimilhança de um dado padrão de respostas a um conjunto de n itens (U1s, Ums), levando-se em conta a habilidade dos alunos (O), consiste no produtório (II) das probabilidades de acerto (P) pelas probabilidades de erro (Q) de cada item individualmente.

No entanto, visando simplificar a expressão 3.10, utiliza-se o artifício de logaritimizá-la, transformando-a na função Log verossimilhança dada pela equação 3.13, a seguir:

$$LogL(u_{is} = 1 \mid \theta) = \sum_{i=1}^{n} u_{is} \log Pi(\theta) + (1 - u_{is}) \log Qi(\theta) \qquad (3.13)$$

Essa equação, denominada Log Máxima verossimilhança, pode ser usada para estimação das proficiências supondo conhecidos os parâmetros dos itens, ou para estimação dos parâmetros

dos itens supondo conhecidas as habilidades dos respondentes, ou estimação conjunta dos parâmetros dos itens e habilidades dos respondentes. Em todas essas situações são necessários métodos iterativos do tipo Newton-Raphson para a sua solução.

A seguir, veremos como o método funciona em cada uma dessas situações. Essa apresentação possui mais um caráter didático do que operacional, pois em algumas dessas situações o método ML é computacionalmente inviável. Entretanto o bom entendimento do método ML nas situações mencionadas é essencial para compreendermos os demais métodos adotados na prática. Portanto, os métodos serão apresentados segundo uma certa complexidade matemática o que implica em otimização de recursos computacionais visando uma solução ótima para a calibração de itens e/ou proficiências.

3.2.1.1 Aplicação do Método ML: Estimação de habilidades com o conhecimento dos parâmetros dos itens

Se já conhecermos os parâmetros dos itens, a solução para a equação Log Máxima verossimilhança (3.13), consiste em achar o valor de θ que corresponda ao maior valor da verossimilhança para cada respondente do teste. Esse valor, para cada vetor de resposta, ocorre quando a primeira derivada da função (3.13) for zero e a segunda derivada for negativa. Como esta segunda derivada é sempre negativa, a máxima verossimilhança ocorre quando a primeira derivada for zero, ou seja, ao acharmos a raiz da derivada da função Log Máxima verossimilhança:

$$f'(\theta) = \frac{\partial \ln L(x_j \mid \theta_j)}{\partial \theta_j} = 0 \qquad (3.14)$$

O método de máxima verossimilhança (*maximum likelihood*- ML) consiste na resolução da equação 3.12, como essa equação não pode ser resolvida diretamente, é necessário a utilização de métodos iterativos como por exemplo, o algorítimo de Newton-Raphson:

$$\theta_n + 1 = \theta_n - \frac{f'(\theta)}{f''(\theta)} \qquad (3.15)$$

Onde,

θ_n = valor inicial de θ

$\theta_n + 1$ = próximo valor de θ

$f'(\theta)$ = derivada 1ª da função (3.11)

$f''(\theta)$ = derivada 2ª da função (3.11)

O processo iterativo de Newton-Raphson segue os seguintes passos, conforme Pasquali (2004):

1°: Inicia-se o processo iterativo com θ =0;

2°: Calcular a 1ª e a 2ª derivadas da equação (3.13)

3°: Calcular a razão E entre 1ª derivada e 2ª derivada, ou seja, $E = \frac{1^a\, derivada}{2^a\, derivada}$;

4°: Calcular o próximo valor de θ através da fórmula (3.13)

5º: Voltar ao passo 1, com o novo valor de θ para iniciar o próximo ciclo

5°: Parar, quando o E tiver atingido o critério de convergência.

3.2.1.2 Aplicação do Método ML: Estimação dos parâmetros dos itens com o conhecimento da habilidade dos respondentes

Embora a estimação dos parâmetros dos itens com o conhecimento prévio da habilidade dos respondentes não seja normalmente encontrada na prática, seu entendimento é importante, uma vez que certos procedimentos envolvendo essa situação são utilizados pelo BILOG-MG no cálculo conjunto de habilidades e parâmetros de itens conforme apresentado no tópico 3.3.

Vejamos a seguinte situação apresentada por Baker (2001 p. 47):

> Para uma amostra de M alunos, que conhecemos a priori a habilidade, vamos verificar o percentual de acerto a um único item em função da habilidade dos alunos. Nesta situação os escores de habilidades desses alunos estarão distribuídos ao longo de todos os níveis da escala de habilidades. Estes alunos serão divididos em J grupos desta escala, sendo que dentro de um mesmo grupo, todos os alunos terão o mesmo nível de habilidade θ j. Assim, dentro de cada grupo j, onde j = 1, 2, 3...,J teremos mj alunos. Temos que dentro de cada grupo J, rj alunos responderam corretamente ao item. Portanto, para um determinado nível de habilidade, representado por θ j, a proporção de respostas corretas observadas será p(θ)= rj/mj, este valor é a estimativa da probabilidade de acerto do item, no nível considerado. Através de procedimentos análogos, podemos calcular a probabilidade de acerto ao item, em cada um dos J grupos considerados. Podemos então, construir um gráfico com a proporção observada de respostas corretas em função da habilidade dos alunos.

Desse modo, a tarefa, agora, consiste em encontrar a Curva Característica do Item – CCI – que melhor se ajusta à proporção observada de respostas corretas. Para tanto, basta aplicar o métodos ML para cada um dos J grupos considerados, seguindo os mesmos passos apresentados no tópico anterior, com a diferença de que nesse momento estaremos interessados em descobrir os valores dos parâmetros *a*, *b* e *c*, lembrando que em nossos estudos sempre trabalharemos com o modelo 3 LP, que maximizem a função 3.13. Dessa forma, após a convergência do método ML, obteremos o valor estimado da probabilidade de acerto, agora denominado P(θ), para cada nível de habilidade dos alunos.

Na figura 3.2, apresentamos a CCI estimada pelo modelo e os valores correspondentes à proporção observada de respostas corretas, assim, podemos verificar o quanto a estimação pelo modelo adotado se ajusta aos valores empíricos.

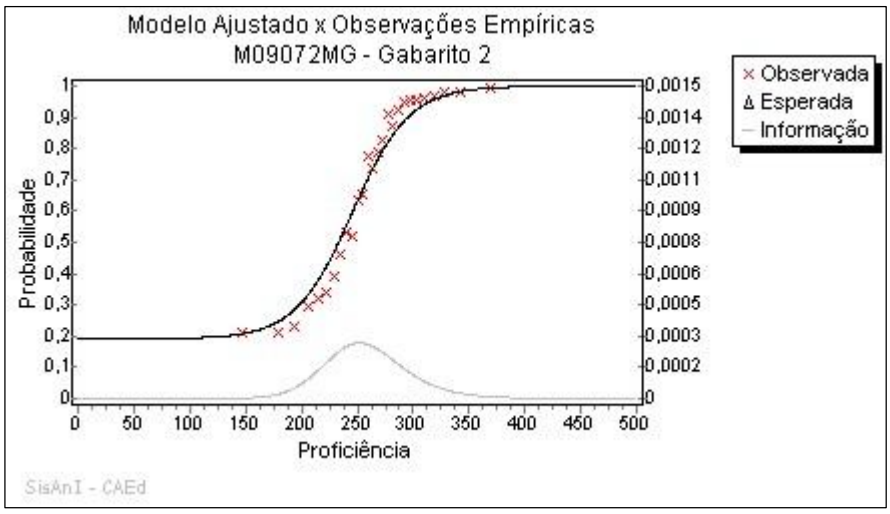

Gráfico 3.4: Ajuste da CCI com os valores empíricos. Fonte: CAEd.

Uma importante consideração nas análises da TRI é verificar o quanto os dados estão ajustados ao modelo adotado para a estimação dos parâmetros dos itens, o que pode ser obtido tanto pela observação do gráfico 3.4 ou através da utilização do índice de ajuste pela fórmula do qui-quadrado. Sugerimos a leitura de Baker (2001), para maiores detalhes relativos à aplicação desse último método.

3.2.1.3 Aplicação do método ML: Estimação de habilidades e parâmetros dos itens sem o conhecimento nem das habilidades nem dos parâmetros dos itens

Como em nossos estudos não conhecemos nem a habilidade nem os parâmetros dos itens, a utilização do método de máxima verossimilhança é ainda mais complexa, pois, ao utilizarmos um modelo de 3 LP, teremos que resolver, simultaneamente, um sistema de 4 equações, derivadas da equação de máxima verossimilhança (3.11): uma para a habilidades e três para os parâmetros dos itens. Para maiores detalhes sobre essas equações, indicamos a consulta de Pasquali,(2004. p. 96).

Esse tipo de estimação é também denominado de Máxima verossimilhança conjunta e para sua solução o processo de estimação é dividido em duas etapas. Na primeira etapa, supõem-

se conhecidas as proficiências dos alunos, normalmente utiliza-se o escore padronizado[3] como valores iniciais no algoritmo de Newton-Raphson e, na segunda, utilizam-se os valores dos parâmetros dos itens, obtidos na primeira etapa, como valores iniciais num processo de iteração para se descobrir a nova proficiência. Após a conclusão da segunda etapa, termina-se um ciclo e inicia-se um novo ciclo até a convergência do processo iterativo. Normalmente, o critério de parada do método iterativo, ocorre quando a diferença de duas verossimilhanças sucessivas, obtidas substituído os valores de parâmetros e proficiências dos respectivos ciclos na função de verossimilhança (3.11), for menor que um determinado valor pré-fixado.

Importante salientar que devido à inexistência dos parâmetros de itens e das proficiências de alunos deparamo-nos com um problema de indeterminação de escala, o qual é solucionado pela fixação da média e desvio padrão da população em cada ciclo, utilizando-se o valor zero para a média e 1 para o desvio padrão. Assim, ao final de cada ciclo, temos que:

- Fazer uma transformação linear na proficiência obtida no ciclo:

$$\theta_{jj} = (\theta_j - \theta_c)/SC \qquad (3.16)$$

θ_{jj} – escore do aluno padronizado na média 0 e desvio padrão 1

θ_j – escore do aluno no ciclo

θ_c – média dos escores no ciclo

Sc – Desvio padrão do ciclo

- Transformar os parâmetros dos itens para ficarem na mesma escala da proficiência:

$$bjj = (bc - \theta_c)/Sc \qquad (3.17)$$

bjj – parâmetro *b* do item i na escala padronizada

bc - parâmetro *b* do item i no ciclo

$$ajj = ac*Sc \qquad (3.18)$$

ajj - parâmetro *a* do item i na escala padronizada

ac - parâmetro *a* do item i no ciclo

[3] O escore padronizado de um respondente é o número de acertos no teste menos menos a média de acerto no teste dividido pelo desvio padrão dos acertos no teste. o número de acertos é também denominado de escore

O parâmetro *c* do item por ser uma proporção, não depende da escala e, portanto, não sofre nenhuma transformação:

$$cjj = cc \qquad (3.19)$$

cjj - parâmetro *c* do item i na escala padronizada
cc- parâmetro *c* do item i no ciclo

A não fixação da escala pelo procedimento visto anteriormente traria como consequência a não convergência do método iterativo. Cabe ressaltar que esse problema de indeterminação da escala não ocorre nas situações anteriores, pois o fato dos parâmetros dos itens estarem na mesma escala da proficiência se já conhecermos qualquer um dos dois, a estimativa do outro estará automaticamente na mesma escala.

A estimação conjunta de habilidades e parâmetros de itens e as transformações de escalas nos parâmetros de cada item e proficiência de cada aluno, tal qual apresentada, demanda uma solução mais complexa, fazendo com que esse método não seja utilizado na prática. Como alternativa, e ainda dentro da estrutura do algoritmo ML, considerado como uma variação do método de Máxima verossimilhança conjunta, normalmente utiliza-se o método de Máxima Verossimilhança Marginal (MML), o qual, para contornar a complexidade da estimação conjunta, usa o artifício de em uma primeira fase, supor conhecidas as habilidades, não de cada respondente, mas de uma distribuição de habilidades obtida de uma amostra da população de estudo e, então, calcular os parâmetros dos itens e, numa segunda fase, considerando verdadeiros os valores dos parâmetros dos itens obtidos da fase 1, calcular as proficiências de cada respondente. Outra vantagem desse método em relação ao anterior está no fato de, ao utilizar, na primeira fase, uma distribuição de habilidade da população, dita marginalizante, os cálculos de indeterminação de escala tornam-se mais simples, que foi o segundo problema apontado no método de máxima verossimilhança conjunta. Esta metodologia será mais detalhada na seção 3.2.3.2 ao descrevermos os métodos utilizados pelo BILOG-MG.

3.2.2 Método bayesiano

A estimação por métodos bayesianos consiste em combinar a função de verossimilhança com distribuições de probabilidade previamente definidas (denominadas de prioris), para os parâmetros dos itens e para as habilidades da população. Assim, utilizando uma aplicação do teorema de Bayes, as estimações não serão mais baseadas na função de verossimilhança e sim numa distribuição de probabilidade a posteriori, sendo essa distribuição proporcional ao produto da função de verossimilhança pelas distribuições a priori, a qual pode ser representada pela seguinte equação:

$$g(\theta,\tau,a,b,c,n\,|\,X) \propto L(X\,|\,\theta,a,b,c)\;\; g(\theta\,|\,\tau)g(\tau)g(a,b,c\,|\,n)\;\; g(n) \qquad (3.20)$$

Onde,

$g(\theta,\tau,a,b,c,n\,|\,X)$ – distribuição á posteriori

$L(X\,|\,\theta,a,b,c)$ – função de verossimilhança

$g(\theta\,|\,\tau)$ – Distribuição de probabilidade do parâmetro de habilidade θ, condicional aos parâmetros populacionais da distribuição de habilidade contida no vetor t.

$g(\tau)$ – distribuição de probabilidades dos hiperparâmetros da população

$g(a,b,c\,|\,n)$ – distribuição de probabilidade para os parâmetros dos itens, condicional aos parâmetros populacionais do vetor n.

$g(n)$ – distribuição de probabilidades dos hiperparâmetros dos itens.

A marginalização da distribuição a posteriori, pode-se dar de duas formas: i) A integração da função (3.20) em relação a θ nos permite calcular os parâmetros dos itens sem a necessidade do conhecimento das habilidades individuais, apenas definindo as prioris para esta população; e ii) A integração da função (3.20) em relação aos parâmetros dos itens nos permite calcular as habilidades sem o conhecimento dos parâmetros de cada item, mas especificando suas

priores. Assim, as estimativas dos parâmetros dos itens e das habilidades serão obtidas pela maximização das respectivas distribuições marginalizadas a posteriori.

Apresentaremos, no próximo tópico, como essas estimativas estão implementadas no BILOG-MG: o método de máxima distribuição marginal a posteriori – MMAP - para estimação de parâmetros de itens e os métodos máximo a posteriori ou modal de Bayes - MAP e o esperado a posteriori ou Bayes- EAP - para a estimação de habilidades.

3.2.3 Estimação através do BILOG-MG

O processo de estimação do BILOG-MG envolve 3 fases distintas, as quais são processadas ou usando o termo normalmente empregado pelos especialistas, rodadas[4], separadamente uma das outras, seguindo a ordem fase1, fase2 e fase 3: Na fase 1, são realizadas as análises clássicas dos itens. Na fase 2, são utilizadas informações da fase1, como default, para a calibração dos itens. Enquanto na fase 3, os parâmetros dos itens gerados na fase 2, são utilizados para gerar as habilidades dos respondentes.

Devemos salientar que a fase 2 somente será inicializada se não houver problemas na fase 1 como, por exemplo, erros na sintaxe, problemas na base e itens com bissseriais negativas. Da mesma forma para que seja possível as análises da fase 3, será necessário que se atinja a convergência na fase 2, caso contrário o programa indicará erro de processamento.

Constatamos, assim, que o BILOG-MG não calcula simultaneamente parâmetros de itens e habilidades de respondentes, pois, conforme descrito anteriormente, esse processo é de difícil resolução. Iremos descrever as principais características de cada uma dessas fases, considerando um testes com apenas um grupo de respondentes. A análise envolvendo múltiplos grupos será tratada no próximo capítulo.

[4] A palavra rodada está relacionada ao fato das diferentes fases serem processadas uma após a outra, e também ao número de ciclos iterativos da fase 2, dando origem às expressos: rodar o BILOG, no sentido de executar o programa e rodou em17 ciclos, significando que a convergência dos métodos iterativos da fase 2 foi atingida em 17 ciclos.

3.2.3.1 Fase 1: Análise clássica dos itens

Nessa fase são calculadas as estatísticas clássicas dos itens que compõem o teste. As principais medidas utilizadas são o percentual de acerto e a correlação bisserial, a qual é uma medida da correlação do resultado de um item em particular com o resultado do teste, sendo, portanto, uma medida da capacidade de discriminação do item em relação ao resultado do teste.

Uma análise importante nessa fase, consiste na verificação dos valores das correlações bisseriais de cada item. Pois, itens com bisseriais muito baixas (geralmente menor que 0.3) e negativas significam que são itens desajustados e que irão prejudicar o processo de calibração da fase 2, logo, são eliminados da análise e uma nova rodada é realizada sem os mesmos.

3.2.3.2 Fase 2: Calibração dos itens

Nessa fase são calculados os parâmetros dos itens. O BILOG-MG tem duas opções para a calibração: o método de máxima verossimilhança marginal - MML e o método de máxima distribuição marginal a posteriori- MMAP. O MML, tal qual implementado no BILOG-MG, conforme relatamos na seção 3.2.2, é uma variação do método ML, no sentido de que também realiza estimativas de máxima verossimilhança para os parâmetros dos itens, porém essas estimativas são obtidas a partir da maximização de uma função de verossimilhança marginal e não da função de verossimilhança. Essa função de verossimilhança marginal é obtida multiplicando-se a função de verossimilhança por uma distribuição a priori. Geralmente, utiliza-se como priori a distribuição normal padronizada (0,1)[5]

$$L(x_1, x_2, ..., x_n \mid a, b, c) = \prod_{j=1}^{n} \int P_j(x_j \mid \theta_j, a, b, c) g(\theta_j \mid \tau) d\theta \qquad (3.21)$$

[5] Distribuição normal padronizada: média zero e desvio padrão um

As estimativas dos parâmetros dos itens obtidas pela maximização da função (3.21), dependem apenas da priori g(θ) e como esta priori é a mesma para todos os θ s, ao final da estimação, os parâmetros dos itens estarão na mesma métrica dessa distribuição, ou seja, não temos o problema de indeterminação de escala.

A distribuição marginal g(θ) é assumida no BILOG-MG como uma amostra aleatória de uma população de habilidades a qual pode ser de três tipos: normal padronizada e mantida fixa durante os ciclos de iteração, empírica e mantida fixa durante os ciclos de iteração ou empírica variando junto com os parâmetros dos itens durante os ciclos de iteração. Em nossos estudos, utilizamos a função de distribuição normal padronizada fixa através do comando *fixed* na sintaxe do BILOG-MG, dividida em 40 pontos de quadratura.

A resolução da equação (3.21) pelo BILOG-MG é feita pela utilização dos algoritmos EM e de Newton-Gauss no processo de iteração. Segundo Valle (1999 p. 39), "o algoritmo EM é um procedimento iterativo para encontrar estimativas de máxima verossimilhança de parâmetros de modelos de probabilidade na presença de variáveis aleatórias não observáveis, chamadas variáveis latentes."

De acordo com Bock, (1997), a aplicação do método EM, *expectation-maximization* para a maximização da verossimilhança requer apenas cálculos das derivadas primeiras da função Log verossimilhança o que o torna fácil de implementar computacionalmente. Além do mais é um método numericamente robusto".

Durante o processamento da fase 2, o BILOG-MG fornece o fator de convergência de cada ciclo, sendo esse fato decrescente ao longo dos ciclos, até atingir um valor abaixo do especificado, em nossas análises trabalhamos com o valor de 0,01, quando então teremos a convergência do método. A observação desses fatores ao longo dos ciclos serve de controle para verificarmos a qualidade da estimação. Se ao invés de decrescer esses fatores forem aumentando ao longo dos ciclos, esse é um indicativo de que temos problemas ou na sintaxe ou na base de dados ou itens com bisseriais negativas. Se o número de ciclos forem muito superiores ao normalmente observado em estudos similares, também temos um indicativo de problema nos procedimentos. Normalmente em estudos de *linkagens* de estados com o SAEB o número de ciclos fica em torno de 20.

O outro método disponível no BILOG-MG para a estimação dos parâmetros dos itens é o de máxima distribuição marginal a posteriori-MMAP. Esse método segue basicamente a mesma

estrutura do método MML a diferença, é que o MMAP, por utilizar priores para os parâmetros dos itens se enquadra como um método bayesiano.

Não existe no BILOG-MG nenhum comando para a escolha entre um dos dois métodos. Essa seleção é feita automaticamente. Se não houver especificações de prioris para os parâmetros dos itens o BILOG-MG usará o método MML, sendo esse o *default* do programa, caso contrário, usará o método MMAP. Esse método pode ser usado em duas situações distintas: a primeira situação ocorre quando todos os itens não possuem parâmetros e nosso objetivo é estabelecer uma nova escala e, a segunda situação ocorre quando parte dos itens já possuem parâmetros relacionados a uma escala já construída, e desejamos calibrar os itens sem parâmetros, nessa escala já existente. Apresentamos, a seguir como utilizar o método para essas duas situações.

Tanto para a primeira como para a segunda situação apresentada acima, devemos especificar prioris normais para o parâmetro de dificuldade, prioris log-normais para os parâmetros de discriminação e prioris beta para o parâmetro de acerto ao acaso. Apresentamos a seguir como especificar essas prioris para essas duas situações.

A primeira situação envolve tanto avaliações com um ou mais grupos em que desejamos criar uma escala nova, e como todos os itens não possuem parâmetros, devemos especificar na sintaxe os valores padrões das prioris disponíveis no BILOG-MG. Já para a segunda situação, sua aplicação somente é possível para mais de um grupo, pois os itens com parâmetro conhecidos pertencem a um ou mais grupos relacionados a uma mesma escala já construída e os itens com parâmetros parcialmente conhecidos, os quais desejamos colocar na escala já existente, pertencem a um outro grupo ou diferentes grupos, e como veremos no capítulo 4, o BILOG-MG usa todos os grupos para a estimação dos itens novos.

O procedimento de calibração para essa situação consiste, segundo Klein (2009) em fixarmos os itens com parâmetros conhecidos, através de distribuições *a priori* com variância muito pequena e utilizarmos as prioris padrões para os itens novos. A fixação dos parâmetros dos itens com parâmetros conhecidos, conforme mencionado por Valle (1999), consiste em "definirmos prioris cujas médias são os próprios valores dos parâmetros que desejamos fixar e cujos desvios-padrão são tão pequenos que a distribuição torna-se praticamente degenerada naquele ponto."

Assim, ao rodarmos o BILOG-MG, todos os itens que já possuíam parâmetros terão seus valores reestimados em torno de seus valores originais e os itens novos terão seus parâmetros estimados na mesma escala já existente.

3.2.3.3 Fase 3: Cálculo das habilidades dos respondentes

O processo nessa fase é o de calcular as habilidades dos respondentes lendo os parâmetros obtidos na fase 2. O BILOG-MG possui 3 métodos para estimação das habilidades: um método de máxima verossimilhança que é o ML, apresentado na seção 3.2.1.1 e dois métodos bayesianos, que são os métodos máximo a posteriori ou modal de Bayes - MAP e o esperado a posteriori ou Bayes- EAP.

O método MAP consiste em achar o máximo da distribuição a posteriori marginal, esta função consiste na função de verossimilhança multiplicada por uma função a priori, assumida como sendo uma normal e tendo o seu máximo na moda distribuição, ou seja em θ = 0. O método MAP segue basicamente os mesmos passos do método ML, porém, ao introduzir as informações da distribuição a priori, no passo 3, conforme apresentado na seção 3.2.1.1, as estimativas das habilidades são mais exatas que as obtidas no método ML.

O método EAP, utilizado pelo SAEB e em todos os estudos de caso deste trabalho, é tão preciso quanto o MAP, porém por ser um método não-iterativo, possui solução mais simples e consiste em estimar a média da distribuição a posteriori de θ . Valle (1999) cita a técnica criada por Bock e Mislevy (1980), segundo a qual para calcular as habilidades dos respondentes, a distribuição a priori das habilidades de cada respondente são divididas em intervalos iguais, sendo que dentro de cada intervalo há o mesmo nível de θ . Esses intervalos são denominados nodos ou pontos de quadratura. Para cada ponto de quadratura são calculadas as suas probabilidades ou pesos, também denominados de densidade. Dessa forma, a estimativa da habilidade será feita levando-se em consideração os pontos de quadraturas e suas densidades. Nas sintaxes utilizadas pelo SAEB, trabalha-se com 20 pontos de quadratura e como priori, utiliza-se a opção no BILOG-MG relativa à distribuição normal. Existem ainda duas outras opções no

BILOG-MG: uma distribuição discreta arbitrária, ou uma distribuição discreta empírica fixando os pontos de quadratura e pesos gerados na fase 2 do BILOG-MG.

No próximo capítulo, ao trabalharmos com os métodos de *linkagem* aplicados na TRI, apresentaremos uma situação prática, no sentido de elucidar ainda mais os temas abordados, envolvendo a utilização dos métodos MMAP para calibração de parâmetro de itens e o método EAP para a estimação das habilidades dos respondentes.

4 MÉTODOS PARA *LINKAGEM* APLICADOS NA TRI

Definimos, no capítulo 1, a *linkagem* como métodos e procedimentos para equiparação ou equivalência de escores entre diferentes avaliações, o que até o presente momento foi o suficiente para o entendimento dos temas apresentados. Entretanto esta definição, tal qual apresentada, guarda implicitamente uma outra característica da TRI que é, também, colocar em uma mesma métrica os parâmetros de itens vindos de testes distintos. Portanto, para evoluirmos em nossos estudos, surge a necessidade de explicitarmos melhor a definição de *linkagem* de forma a torná-la mais abrangente e mais próxima dos temas a serem abordados. Para isso tomamos Valle (1999), segundo a qual "*linkagem* são métodos e procedimentos para se colocar em uma mesma métrica habilidades e parâmetros de itens relacionados a diferentes grupos e testes, isto é, numa escala comum, tornando os itens e/ou habilidades comparáveis."

Descreveremos, nas próximas seções, os métodos de *linkagens* adotados na TRI, os quais serão apresentados em dois agrupamentos: métodos lineares e não-lineares. Focaremos a aplicação dos diferentes métodos em situações envolvendo designs de grupos não equivalentes com itens comuns, em *linkagens* do tipo equalização vertical, pois essa é a situação, geralmente, encontrada nas avaliações nacionais e, consequentemente, em todos os estudos de caso desta dissertação.

Diferentes métodos de *linkagem* envolvendo os demais designs apresentados no capítulo 1, e também para escores baseados na Teoria Clássica do Teste podem ser encontrados em Kolen e Brennan (2004).

4.1 MÉTODOS DE *LINKAGEM* BASEADOS NA TRI: MÉTODOS LINEARES.

Os métodos lineares de *linkagem* são utilizados para se colocar em uma mesma métrica dois ou mais grupos em que as escalas foram obtidas separadamente. Nesse caso, a *linkagem* é realizada *a posteriori*. Para simplificar as apresentações dos principais métodos utilizados nesse contexto, iremos trabalhar apenas com dois grupos, destacamos, porém, que, para mais grupos as

análises são similares. Portanto, sejam os grupos i e j os quais possuem escalas I e J, respectivamente, obtidas através do modelo de 3 LP. As relações entre habilidades e parâmetros de itens entre as duas escalas são obtidas pelas fórmulas a seguir:

Equação de transformação da habilidade de indivíduos na escala I para uma escala J :

$$\theta_{Js} = A^* \theta_{Is} + B \qquad (4.1)$$

θ_{Js} – Habilidade do indivíduo s na escala J

θ_{Is} - Habilidade do indivíduo s na escala I

A e B – Constantes de transformação da equação linear

Equações de transformações de parâmetros de itens na escala I para uma escala J :

$$aJp = aIp/A \qquad (4.2)$$

$$bJp = A^*bIp + B \qquad (4.3)$$

$$cJp = cIp \qquad (4.4)$$

onde, aJp, bJp e cJp são os parâmetros do item p na escala J e aIp, bIp e cIp são os parâmetros do item p na escala I.

4.1.1 - Método de Regressão linear simples

A solução mais natural para a obtenção dos valores A e B de transformação entre duas escalas, no nosso caso, I e J é aplicar uma regressão linear simples entre os valores dos parâmetros *b* dos itens comuns entre essas duas escalas. No entanto, esse método não é muito

utilizado na prática por não ser simétrico, ou seja, a regressão linear de I por J não é a mesma de J por I.

Normalmente esse método é usado, para duas ou mais escalas distintas, em que os valores de habilidades e parâmetros de itens, foram obtidos de uma mesma população em que foram aplicados os mesmos itens, Nesse caso, valores idênticos das constantes de transformações A e B podem der obtidos via regressão linear entre os valores das habilidades ou entre os valores dos parâmetros *b* dos itens.

Essa situação é comum ao trabalharmos com o BILOG-MG, pois os arquivos de parâmetros de itens e proficiências obtidos nos processamentos estão sempre referenciados a uma distribuição normal (0,1) e via de regra temos que transformar esses valores para outra escala.

4.1.2 - Método Média/Média

Nesse método, as constantes de transformações A e b são obtidas através das seguintes equações:

$$A = \mu(aI) / \mu(aJ) \qquad (4.5)$$

$$B = \mu(bJ) - A^{*} \mu(bI) \qquad (4.6)$$

Os valores μ(aI) e μ(aJ) são as médias dos parâmetros *a* dos itens comuns às duas escalas.

4.1.3 Método Média/Sigma

Esse método utiliza a equação (4.6) para o cálculo da constante B, porém, para o cálculo da constante A são utilizados os desvios padrão dos parâmetros *b* dos itens comuns nas duas escalas conforme fórmula a seguir:

$$A = \sigma(bJ) / \sigma(bI) \qquad (4.7)$$

A vantagem desse método em relação ao método de regressão linear é a sua simetria, ou seja, a regressão linear de I por J é a mesma de J por I.

4.1.4 Métodos da Curva Característica

Nos dois últimos métodos apresentados nas seções anteriores, as constantes de transformações são obtidas sem levar em consideração os três parâmetros dos itens de forma simultânea. Esta característica gera problemas em situações em que os parâmetros *b* de um item em duas escalas são muito diferentes, porém suas curvas características são similares. Para contornar este problema temos os métodos de Haebara e Stocking e Lord, denominados métodos da curva característica. Esses dois métodos, através de procedimentos iterativos levam em consideração todos os parâmetros dos itens comuns às duas escalas simultaneamente para a estimação das constantes A e B. Maiores detalhes da utilização destes métodos, bem como comparações entre seus resultados com os resultados dos métodos de média/média e média/sigma podem ser obtidos em Kolem e Brennan, (2004).

A escolha de qual método utilizar para a *linkagem* é uma decisão difícil, pois não existe uma regra para auxiliar nesta tarefa. O que temos observado nas avaliações nacionais, é uma predominância do método média/sigma quando se utiliza *linkagem* que envolvem métodos lineares.

4.2 - MÉTODOS DE *LINKAGEM* BASEADOS NA TRI: MÉTODOS NÃO-LINERARES.

Diferente dos métodos lineares em que as *linkagens* dos diferentes grupos são realizadas *a posteriori*, nos métodos não-lineares a escala única entre diferentes grupos é obtida durante o processo de estimação dos parâmetros dos itens. Tendo como referência o BILOG-MG, apresentaremos no anexo 1- Métodos não-lineares de linkagem utilizados no BILOG-MG, os dois métodos característicos desse tipo de *linkagem*, que são os métodos de calibração simultânea e o de pré-fixação de parâmetros (FPIP)[6]. Em linhas gerais, o primeiro método é usado quando desejamos criar uma nova escala envolvendo diferentes grupos e o segundo, quando desejamos, também para diferentes grupos, manter uma escala já existente.

4.3 – CARACTERÍSTICAS DAS *LINKAGENS* NO BRASIL

Diante de um projeto de avaliação educacional, os especialistas, com a função de produzir uma escala de habilidades pela TRI, têm basicamente de definir qual o *software* a ser utilizado, o modelo logístico, quais os métodos de estimação de parâmetros e de escores e qual o método de *linkagem*. Essas decisões levarão em conta as características das populações envolvidas, os designs de coleta de dados e os designs dos testes. Não existe, na literatura, uma unanimidade com relação à melhor opção, o que observamos é uma grande divergência de opiniões vindas de diferentes especialistas.

Com relação às avaliações em larga escala realizadas no Brasil, tanto em âmbito nacional quanto estadual, o software utilizado é o BILOG-MG, o modelo logístico é o de 3 parâmetros (3 LP), para a estimação dos parâmetros dos itens utiliza-se o método MMAP, para a estimação dos escores o método EAP e para a *linkagem*, que é do tipo equalização vertical, utiliza-se o método FPIP. Apresentamos no diagrama da figura 4.1 um esquema com essas características.

---

[6] *Fixed Precalibrated Item Parameter*

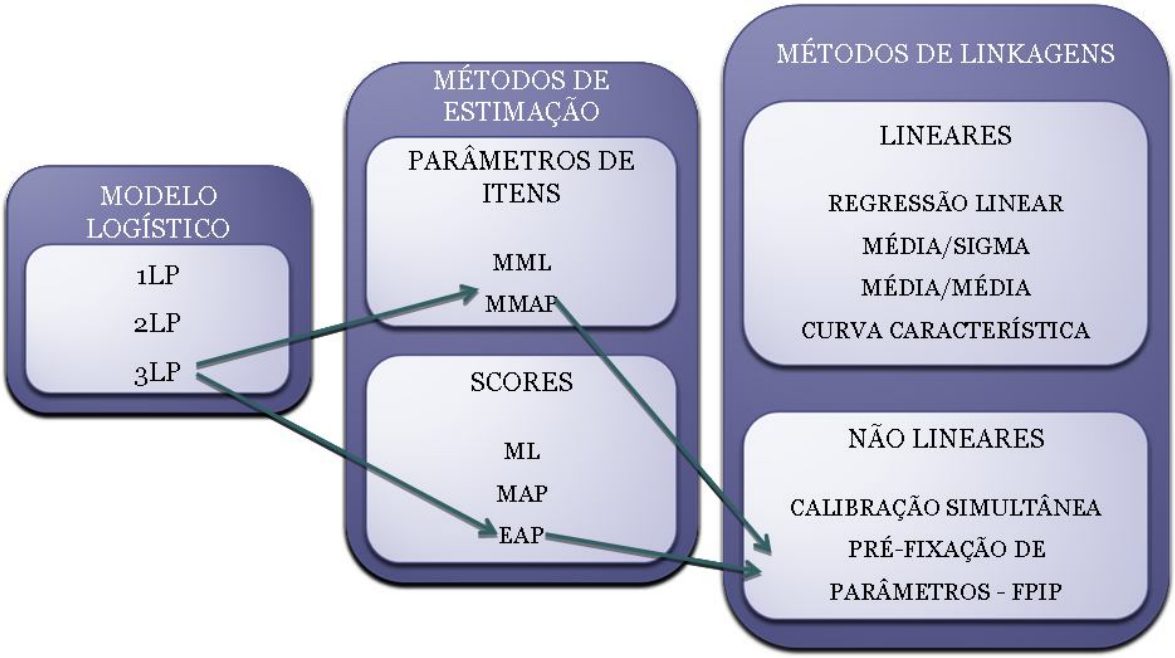

Figura 4.1: Diagrama das avaliações em larga escala no Brasil.

Nos processos de *linkagem* pelo método FPIP, além da exigência de se ter itens comuns entre os diferentes grupos, há a necessidade de se utilizar as mesmas bases de dados com as opções de respostas utilizadas para a calibração desses itens comuns. Assim, o procedimento utilizado pelo SAEB, ao longo de suas diferentes versões, foi de utilizar em uma determinada versão, itens comuns e populações da versão anterior, ou seja, no SAEB de 1999 foram utilizados itens comuns com o SAEB de 1997 e, nos processos de *linkagem* foram utilizadas as mesmas populações de 1997 com as novas populações de 1999, e, seguindo os procedimentos apresentados no anexo 1, os parâmetros de todos os itens de 1997, assim como as proficiências dos alunos desse ano foram reestimadas e os parâmetros dos itens e proficiências de 1999 foram calculados na mesma escala do SAEB. Esta mesma dinâmica, ou seja, utilizar a avaliação anterior para estimar a avaliação seguinte foi adotada nas análises seguintes, ou seja, SAEB 2001 com SAEB 1999, SAEB 2003 com SAEB 2001, SAEB 2005 com SAEB 2003, SAEB 2007 com SAEB 2005 e SAEB 2009 com SAEB 2007.

Este procedimento apresenta o inconveniente de limitar a utilização de itens comuns apenas com a avaliação anterior, ou seja, embora se tenha uma grande quantidade de itens calibrados

desde 1995, estes não podem ser utilizados, pois, para mesclar itens de diferentes versões do SAEB através do método de FPIP, também seria necessário colocar as mesmas populações empregadas para a calibração desses itens, o que tornaria os procedimentos de *linkagens* extensos e complicados, o que o tornaria inviável.

O fato de se manter essas características em todas as avaliações educacionais realizadas no Brasil desde 1995 é um forte indicativo de que os resultados dessas avaliações podem ser comparados entre si. O que veremos a seguir, serão variações em designs de testes e variações em características das populações e os efeitos na comparabilidade de resultados, porém, mantendo inalteradas as características apresentadas na figura 4.3, ou seja, todos os estudos de casos apresentados nesse trabalho seguem essa mesma estrutura.

5 ESTUDOS PRÁTICOS SOBRE A EFICÁCIA DAS *LINKAGENS*

No Brasil, observamos uma convicção entre as diversas instâncias que lidam com os resultados das avaliações educacionais tanto no âmbito nacional quanto no estadual e no municipal que os resultados alcançados nas avaliações dos diferentes entes federados são totalmente comparáveis entre si e entre diferentes anos.

Entretanto, certas condições técnicas, necessárias para uma boa comparabilidade de resultados, nem sempre são seguidas, o que pode gerar conclusões indevidas. Um exemplo clássico dessa situação é o caso denominado *NAEP reading anomaly* ocorrido nos EUA, no ano de 1986, conforme nos apontam Kolen e Brennan(2004). Observou-se, naquele ano de 1986, uma queda muito acentuada nos resultados do teste de leitura apresentados pelos alunos de 17 anos de idade ao ser comparado com os resultados do ano de 1984. Analistas dos setores envolvidos não concordavam que, num período de 2 anos, durante o qual nenhuma alteração no sistema educacional havia sido implementada, pudessem ocorrer variações tão significativas nos resultados dos alunos.

Após diversas análises, chegou-se à conclusão de que a queda nos resultados estava relacionada não à problemas de aprendizado dos alunos, mas sim a alterações nas características dos testes, como diferentes disciplinas avaliadas entre os dois anos, diferentes tempos disponíveis para os alunos responderem aos itens e diferentes posicionamentos dos itens nos testes.

Esse caso revela o quão séria é a utilização dos resultados de uma avaliação concebida com o propósito de estar em uma mesma escala com outra avaliação. É neste cenário que abordaremos neste capítulo, os principais fatores que influenciam as comparabilidades entres resultados de avaliações e, através de exemplos reais utilizando dados das avaliações do SAEB e de alguns estados brasileiros, estudar as influências de certas alterações nestas avaliações e suas conseqüências na geração de proficiências de alunos e parâmetros de itens e o quanto essas alterações impactam na comparabilidade dos resultados, de forma a podermos garantir se estamos realmente comparando resultados de forma confiável.

5.1 FATORES QUE INFLUENCIAM A *LINKAGEM*

A confiabilidade dos resultados obtidos nos processos de *linkagens* pode sofrer influência de quatro fatores, os quais serão abordados a seguir.

5.1.1 Conteúdo do teste

Diferentes conteúdos, medidos através de diferentes matrizes de referência, podem afetar a qualidade da *linkagem* por vários motivos. O principal é a unidimensionalidade do teste, hipótese que deve ser verificada sempre que diferentes testes são construídos a partir de diferentes conteúdos. Evidentemente, modelos multiníveis poderão vir a ser utilizados para lidar com essa situação. No entanto, há dificuldades técnicas para se utilizar esse tipo de modelo.

Testes com diferentes conteúdos podem medir diferentes performances entre os grupos avaliados. Por exemplo, estudantes com problemas de aprendizagem em álgebra ou que ainda não estudaram esse conteúdo terão um desempenho baixo em um teste de Matemática que tenha focado essa área. Porém, esses mesmos estudantes poderão ter um desempenho alto em testes que estejam focados em outras disciplinas. Assim, segundo Feuer et al (1999), quando as diferenças de conteúdos entre testes são significativas, qualquer tentativa de *linkagem* entre os mesmos fornecerão pouco significado prático e poderão gerar falsas interpretações em algumas utilizações.

5.1.2 Formato do teste

Os efeitos de diferentes formatos e tipos de aplicação de testes no processo de *linkagem* não são previsíveis, mas podem ser grandes. Podemos destacar, como diferenças entre testes, os seguintes aspectos:

i) Testes de tamanhos diferentes. Por exemplo, na 4ª série do Ensino Fundamental da Prova Brasil, nos anos de 2005 e 2007, o número de itens por caderno aumentou de 40 para 44.

ii) Testes que avaliam diferentes disciplinas. Por exemplo, a *linkagem* de Língua Portuguesa da avaliação do estado do Rio de Janeiro no ano de 2004 com o SAEB 2003. Na avaliação do Rio de Janeiro, em um mesmo caderno de teste, foram avaliadas as disciplinas Língua Portuguesa, Matemática, Ciências Humanas e Ciências da Natureza e, no SAEB 2003, apenas uma disciplina. Um outro exemplo dessa situação seria a *linkagem* de Língua Portuguesa e Matemática entre o SAEB 2005 com o SAEB 2007, pois, no SAEB 2005 havia um teste para cada disciplina e no SAEB 2007 as duas disciplinas estavam no mesmo teste.

iii) Testes com itens fechados e abertos. Linn et al. (1992) demonstraram diferenças percentuais de alunos nos níveis do NAEP, ao considerar seus desempenhos em itens abertos e fechados.

iv) Testes com itens apenas lidos pelos alunos linkados com testes que apresentam itens lidos pelo aluno, lidos parcialmente e/ou totalmente lidos pelo aplicador. Por exemplo, a *linkagem* entre a 2ª e 3ª série do Ensino Fundamental do Programa de Avaliação da Alfabetização de Minas Gerais – PROALFA – no qual no 2º ano os itens são lidos para os alunos e no 3º ano os alunos lêem o itens.

v) Testes aplicados por um aplicador externo, não pertencente à escola e testes aplicados por um professor da escola. Por exemplo, a *linkagem* entre as avaliações do Programa de Avaliação do Ensino Básico de Minas Gerais - PROEB – com o SAEB. No primeiro caso, o aplicador de testes é o professor da própria escola (de uma turma diferente da qual ele leciona), enquanto no SAEB há um aplicador externo.

vi) Ordem das disciplinas nos testes: ao se avaliar duas disciplinas, o que é a situação mais comum nas avaliações realizadas no Brasil, em que, geralmente, se avalia Língua Portuguesa e Matemática, podemos encontrar quatro diferentes tipos de montagem de cadernos: (a) testes com apenas uma disciplina aplicados em dias diferentes; (b) testes em que todos os cadernos têm a primeira metade dos itens de uma disciplina e a segunda metade de outra disciplina, (c) testes em que a primeira metade dos cadernos começa com uma disciplina e a segunda metade com outra disciplina; (d) testes em que há uma mistura de blocos das duas disciplina de forma alternada, por exemplo: metade dos cadernos contendo o primeiro bloco com a disciplina A, o segundo com a disciplina B, o terceiro com a disciplina A e o último com a

disciplina B; enquanto a outra metade dos cadernos começa com a disciplina B e termina com a disciplina A. Mostraremos, na seção 5.2, um exemplo de como a comparabilidade de resultados é afetada pela utilização de diferentes designs de testes.

5.1.3 Usos e conseqüências dos resultados das avaliações

A estabilidade do processo de *linkagem* é afetada quando sanções e premiações são adotadas em um grupo e não no outro. Isto ocorre ao *linkar*mos determinados estados que adotam políticas de premiações para escolas e alunos, bonificação de professores e ranqueamento de escolas ao SAEB que não possui esse tipo de política.

Quando as consequências das avaliações são pequenas, os alunos têm pouco incentivo em fazer os testes da melhor forma possível. Se há razão para se preocupar com os resultados, então o empenho é maior. Com relação aos professores, também observa-se um maior empenho quando as consequências em relação aos resultados são altos. Nesse caso, costuma-se observar estratégias focando conhecimentos e habilidades específicas que serão abordadas nos testes, visando um melhor desempenho dos alunos nos testes.

Conforme assevera Feuer et al (1999, p. 89),

> [...] quando testes relacionados com sanções e premiações são linkados com testes que não possuem esta característica, a dificuldade relativa entre os mesmo é alterada, isto é, o teste parecerá mais fácil para avaliações inseridas no primeiro contexto e isto pode afetar a estabilidade da *linkagem* ao longo do tempo.

5.1.4 Erro do método estatístico

Os diferentes procedimentos de equalização exigem uma série de pressupostos que podem não ser verificados na prática como, por exemplo, a quantidade e qualidade de informação trazidas pelos itens comuns com o grupo ao qual se deseja realizar a *linkagem* (isso em design de

itens comuns). A literatura tem estudos extensivos para avaliar esses erros, sejam analíticos, sejam obtidos por simulação.

Esses diversos fatores apresentados anteriormente, se considerados isoladamente, poderão não ter grandes efeitos nas *linkagem*, entretanto o problema maior ocorre quando há diferentes conjugações dos mesmos, o que acarretará grandes influências na comparação de resultados. Apresentaremos a seguir alguns estudos práticos envolvendo diferentes situações de *linkagens* em que a não consideração dos fatores mencionados podem levar a resultados enganosos.

5.2 ESTUDOS EMPÍRICOS ENVOLVENDO A CONFIABILIDADE DE *LINKAGENS*

Conforme relatado anteriormente, as *linkagens* realizadas pelo SAEB e pelos estados brasileiros são do tipo equalização vertical e, conforme apontado por Kolem e Brennan (2004), os resultados desse tipo de *linkagem* dependem fortemente das características das populações avaliadas, do sistema de coleta de dados e dos métodos estatísticos empregados. Em nossas análises, envolvendo diferentes avaliações, manteremos: i) o mesmo modelo logístico (3 –LP), ii) os mesmos métodos estatísticos para estimação dos parâmetros dos itens (MMAP) e proficiências (EAP) e iii) o mesmo método de *linkagem* (FPIP), ou seja: os mesmos procedimentos adotados pelo SAEB conforme apresentado no capítulo 4. Assim, mantendo-se inalterados esses três conjuntos de fatores, analisaremos, para algumas situações de *linkagem*, ocorridas no Brasil, apenas os efeitos de diferentes designs de testes e de diferentes populações envolvidas no processo, com o intuito de detectarmos, até que ponto esses efeitos, isoladamente, influenciam os resultados das avaliações de maneira a ainda podermos ter comparabilidade com a escala do SAEB, uma vez que, em praticamente todas as avaliações estaduais realizadas, o objetivo é que as mesmas estejam na escala nacional. Na prática, esses efeitos, além de ocorrerem juntos, podem ainda estar alinhados com outros, dificultando a detecção de suas influências nos resultados. Entretanto, em neste trabalho, utilizaremos uma técnica de comparabilidade entre situações nas quais apenas uma característica será alterada, isso nos permitirá mensurar o nível de seus efeitos na comparabilidade de resultados.

Temos observado, ao longo dos anos, que alguns estados brasileiros e até mesmo as avaliações nacionais tiveram variações em suas características originais, o que nos faz questionar se estas alterações não tornaram inviáveis as comparações ao longo dos anos, e até que ponto podemos comparar os resultados das avaliações nacionais entre diferentes estados que realizaram suas próprias avaliações e entre estes próprios estados.

Podemos destacar variações como a de mudança de design ocorrida no SAEB em 2007, estados que avaliam mais de duas disciplinas em um mesmo teste, testes com diferentes números de itens e ou diferentes matrizes.

Através de dados reais de algumas avaliações estaduais e nacionais abordaremos diferentes situações de *linkagens*, com o objetivo de diagnosticarmos até que ponto podemos garantir comparabilidade nos resultados das avaliações, envolvendo proficiências de alunos e parâmetros de itens quando são alteradas as características dos designs dos testes e das populações envolvidas nas *linkagens*.

5.2.1 Efeito do design dos testes na proficiência da população

Faremos duas análises, através de diferentes projetos, sobre o efeito de se colocar em um mesmo caderno de teste, duas disciplinas diferentes, em designs em que uma disciplina sempre aparece no início ou fim do caderno e em designs onde as disciplinas são alternadas nos cadernos, ou seja, ora no início, ora no fim do caderno, de forma balanceada.

5.2.1.1 Nova Escola nos anos de 2005 e 2006

No projeto de avaliação educacional Nova Escola, realizado no Estado do Rio de Janeiro, nos anos de 2005 e 2006, os cadernos de testes eram compostos por Língua Portuguesa e Matemática (cadernos ímpares iniciados com Língua Portuguesa e cadernos pares, com Matemática), segundo o mesmo design aplicado na Prova Brasil 2005. Segundo esse design,

observamos uma diferença significativa nos cálculos de proficiência, ao se comparar os resultados dos grupos formados pela ordem das disciplinas nos cadernos, ou seja: alunos que fizeram os testes começando com Língua Portuguesa tiveram, nessa disciplina, um valor de proficiência maior que os alunos que fizeram essa disciplina no final do caderno. Para o caso inverso, ou seja, alunos que fizeram os testes começando por Matemática, também obtiveram um resultado maior que os alunos que fizeram essa disciplina no final do caderno. O fato de se ter duas disciplinas em um mesmo caderno de teste provoca um cansaço e, consequentemente, uma queda na proficiência na disciplina que está no final do caderno. Os valores relativos a estas diferenças são apresentados no Quadro 5.1

SÉRIE	ORDEM DAS DISCIPLINAS	2005		2006	
		Língua Portuguesa	Matemática	Língua Portuguesa	Matemática
4 EF	LP/ MAT	175	177	182	186
	MAT/ LP	169	181	173	193
	DIFERENÇA	6	4	9,5	7
8 EF	LP/ MAT	227	224	229	225
	MAT /LP	215	230	216	233
	DIFERENÇA	12	6	13,3	8
3 EM	LP /MAT	253	257	245	250
	MAT /LP	241	263	228	257
	DIFERENÇA	12	6	16,6	7

Quadro 5.1 Influência da ordem das disciplinas, Língua Portuguesa e Matemática, na proficiência dos alunos no Projeto Nova Escola. Fonte: CAEd

Observamos, nesse quadro, que as diferenças de proficiências em Língua Portuguesa foram mais significativas do que em Matemática, ao se considerar a ordem das disciplinas nos cadernos. Também podemos verificar que, em Língua Portuguesa, essa diferença, na 8ª série EF e 3º ano EM, é mais significativa do que na 4ª série EF e que, em Matemática, as diferenças entre estes três anos de escolaridade é praticamente a mesma, ou seja, o efeito na proficiência dos alunos, provocado por esse design é mais forte na 8º série EF e 3º ano EM de Língua Portuguesa.

Fica evidente a existência de um efeito devido ao cansaço, provocando uma queda no desempenho dos alunos, ao fazer uma disciplina no final do caderno.

Ao analisarmos a distribuição dos percentis dos alunos da 4ª série de Língua Portuguesa, no ano de 2006 (gráfico 5.1), levando em consideração a posição dessa disciplina no teste, ou seja, no início ou no final, constatamos que as maiores diferenças ocorrem no meio da curva e que, nas

extremidades da curva, essas diferenças são menores, evidenciando que, para alunos com proficiência muito baixa ou muito alta, o fato de mudar a posição da disciplina no caderno de teste produz pouco efeito na proficiência. Esse efeito é maior para os alunos medianos. Essa característica foi a mesma observada nas demais séries para as duas disciplinas nos dois anos analisados.

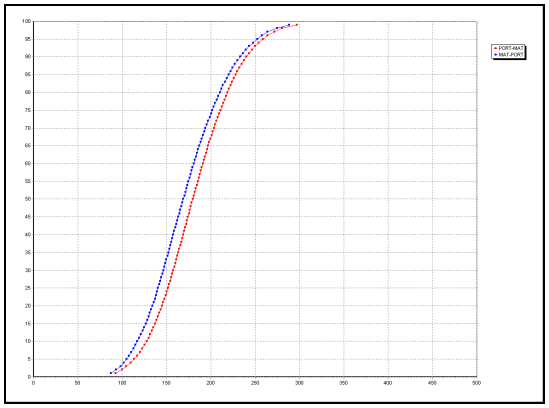

Gráfico 5.1: Percentis dos alunos na avaliação de Língua Portuguesa na 4ª série do EF no ano de 2006 no programa Nova Escola para testes na ordem Língua Portuguesa/Matemática e na ordem Matemática/Língua Portuguesa. Fonte: CAEd

5.2.1.2 SAEB nos anos 2005 e 2007

No quadro 5.2, a seguir, observamos que, no SAEB de 2005, ano em que os alunos responderam a apenas uma disciplina no teste, as diferenças de proficiências entre alunos que responderam a cadernos pares e os alunos que responderam a cadernos ímpares é muito baixa e deve-se exclusivamente a variações aleatórias de dificuldades entre os cadernos. Já, no SAEB de 2007, ano em que as disciplinas de Língua Portuguesa e de Matemática foram aplicadas no mesmo caderno, verificamos diferenças significativas e não mais aleatórias nas proficiências, pois há uma tendência em superestimar as proficiências quando a disciplina se encontra no primeiro bloco do caderno e subestimar quando a mesma está no segundo bloco do caderno.

SÉRIE	SAEB 2005			SAEB 2007		
	ORDEM DAS DISCIPLINAS	Língua Portuguesa	Matemática	ORDEM DAS DISCIPLINAS	Língua Portuguesa	Matemática
5 EF	ÍMPAR	171.4	182.3	ÍMPAR (LP/MAT/LP/MAT)	175.9	191.0
	PAR	172.6	182.7	PAR (MAT/LP/MAT/LP)	173.2	194.4
	DIFERENÇA	-1.2	-0.4	DIFERENÇA	2.7	-3.4
9 EF	ÍMPAR	230.7	239.1	ÍMPAR (LP/MAT/LP/MAT)	235.5	242.7
	PAR	232.4	240.1	PAR (MAT/LP/MAT/LP)	229.3	248.6
	DIFERENÇA	-1.6	-1.0	DIFERENÇA	6.1	-5.9
3 EM	IMPAR	256.7	270.7	IMPAR (LP/MAT/LP/MAT)	264.6	269.7
	PAR	257.7	271.8	PAR (MAT/LP/MAT/LP)	257.8	276.3
	DIFERENÇA	-1.0	-1.2	DIFERENÇA	6.8	-6.6

Quadro 5.2 Influência da ordem das disciplinas, Língua Portuguesa e Matemática, no caderno de teste e a proficiência dos alunos no SAEB. Fonte: Arquivos SAEB

5.2.1.2 .1 Comparação entre os designs nova escola 2005/2006 e saeb 2007

Comparando o programa Nova Escola 2005/2006 e o SAEB 2007, verificamos que, ao alternar as disciplinas no caderno de teste, como foi o caso do SAEB 2007, e não concentrando as mesmas no início ou final do caderno, como no caso do Nova Escola 2005/2006, as diferenças de proficiências entre as disciplinas ficaram menores no SAEB 2007 aquelas encontradas no Nova Escola. Isto é, o efeito cansaço provocado pela inclusão das duas disciplinas em um mesmo caderno é minimizado.

Entretanto, observamos que, tanto no âmbito nacional quanto no estadual, as comparações de resultados são realizadas sem se levar em consideração os fatos apresentados.

Fica evidente que teremos problemas na comparabilidade entre avaliações que adotem diferentes designs de testes. Por exemplo:

- Avaliações do SAEB até 2005 com as avaliações do SAEB a partir de 2007
- Avaliações nacionais e avaliações estaduais com diferentes designs.
- Entre avaliações estaduais com diferentes designs.

Portanto, assim como devemos ter cuidado ao compararmos os resultados do Nova Escola 2005/2006 com os resultados da Prova Brasil 2007 para o Rio de Janeiro, esta mesma atenção deveria ser adotada sempre que tivermos comparações que se enquadrem nos casos apresentados neste estudo.

5.2.2 Efeito da população na geração dos parâmetros e proficiências

Uma das principais características da TRI é a invariância dos parâmetros dos itens em relação aos grupos, ou seja, os parâmetros dos itens não dependem do nível de habilidades dos respondentes, estes, são uma característica exclusiva dos itens e não das diferentes populações que os responderam.

Entretanto, conforme relatado por Kolen, Brennan (2004), quanto maior a diferença de habilidades entre grupos de indivíduos, fica mais difícil para os métodos estatísticos de *linkagens* separar as diferenças dos grupos e das formas dos testes, ou seja, nenhum método oferecerá uma boa comparabilidade quando os grupos forem muito diferentes.

Surge, então, um questionamento: Até que ponto podemos *linkar* grupos diferentes e ainda assim obtermos resultados confiáveis?

Esta questão é extremamente relevante no cenário das avaliações brasileiras, onde temos diversas situações de *linkagens* envolvendo grupos, formados por estados, com características bem diferentes,. Diante disso, colocamos, novamente, a questão: será que, ao *linkar*mos estados como Rio Grande do Norte e Minas Gerais com uma mesma base do SAEB, estaremos realmente gerando resultados comparáveis entre si?

Numa tentativa de elucidar e quantificar esta questão, apresentamos, seguindo uma ordem de complexidade de fatores que, a princípio, dificultam as comparabilidades das avaliações, os efeitos nos cálculos dos parâmetros dos itens e proficiências dos alunos, em duas situações:

- Sub-grupos de uma mesma população submetidos aos mesmos itens, porém com diferentes designs de testes.
- Populações diferentes submetidas ao mesmo design de teste e mesmos itens.

Há de se esperar que as comparações sejam mais eficientes na primeira situação e menos eficientes na segunda situação. Sendo que a segunda situação é o nosso grande foco, pois é o caso mais usualmente encontrado nas avaliações estaduais brasileiras, onde realizamos as calibrações dos itens dentro de um determinado estado, geralmente com escolas públicas, e não em uma amostra nacional, contendo escolas públicas e privadas, conforme procedimento adotado pelo SAEB. Essa situação leva-nos à uma questão final: será que ao calibrarmos os itens dentro de um

determinado estado garantimos comparabilidade de seus resultados com os resultados oficiais do SAEB?

Nosso objetivo, portanto, é mapear, através de estudos empíricos, se a comparabilidade de resultados, envolvendo as proficiências dos alunos entre diferentes avaliações, é mantida, e também, se os parâmetros dos itens não se alteram significativamente, sendo que este último fato é extremamente relevante quando levamos em consideração a criação de bancos de itens em uma mesma escala com o SAEB.

5.2.2.1 Sub-grupos de uma mesma população submetida aos mesmos itens, com diferentes designs de testes

Tendo como referência o programa Nova Escola de 2006, simulamos o efeito na geração dos parâmetros dos itens e proficiência dos alunos quando realizamos *linkagens* em sub-grupos de uma mesma população. Para tanto, dividimos as bases de dados do Nova Escola em dois grupos: alunos que fizeram uma determinada disciplina na primeira parte do caderno e alunos que a fizeram no final do caderno. Com este procedimento, obtivemos de uma mesma população, dois grupos com médias de proficiências abaixo e acima de uma média padrão, que no caso em específico, é a média com a base completa.

Dividimos as bases de dados de cada uma das disciplinas em dois arquivos: um arquivo com os alunos que fizeram a disciplina na primeira metade do caderno e um outro arquivo com os alunos que fizeram a disciplina no final do caderno, desta forma, foram gerados 4 arquivos:

- Arquivo 1: alunos que responderam a cadernos ímpares
- Arquivo 2: Alunos que responderam a cadernos pares

E para Matemática:

- Arquivo 3: alunos que responderam a cadernos pares
- Arquivo 4: Alunos que responderam a cadernos ímpares

No caso de Língua Portuguesa, após a separação dos arquivos, calibramos os itens, com cada um dos dois arquivos gerados, e em seguida calculamos dois valores de proficiência, desta vez utilizando a base completa: um primeiro valor, lendo o arquivo de parâmetros gerado com o

arquivo 1 e em seguida, um segundo valor lendo o arquivo de parâmetros gerado com o arquivo 2. Para Matemática fizemos o mesmo procedimento, utilizando os arquivos 3 e 4. Estes valores, assim como as proficiências oficiais do programa, obtidas com a leitura do arquivo de parâmetros com a base completa, estão apresentados nos quadros 5.3e 5.4

SÉRIE	PROFICIÊNCIA			DESVIO PADRÃO		
	LENDO PARÂMETROS GERADOS COM A BASE COMPLETA	LENDO PARÂMETROS GERADOS COM O ARQUIVO 1	LENDO PARÂMETROS GERADOS COM O ARQUIVO 2	LENDO PARÂMETROS GERADOS COM A BASE COMPLETA	LENDO PARÂMETROS GERADOS COM O ARQUIVO 1	LENDO PARÂMETROS GERADOS COM O ARQUIVO 2
4 EF	177.42	177.73	177.06	44.38	44.23	44.34
8 EF	222.88	222.89	222.87	45.88	45.88	45.58
3 EM	236.37	236.52	236.34	51.41	52.08	50.61

Quadro 5.3 Proficiência em Língua Portuguesa por série no programa Nova Escola 2006 em 3 situações de leitura de parâmetros de itens: Parâmetros gerados com a base completa, com a base de cadernos ímpares (arquivo 1) e com a base de cadernos pares (arquivo 2).

SÉRIE	PROFICIÊNCIA			DESVIO PADRÃO		
	LENDO PARÂMETROS GERADOS COM A BASE COMPLETA	LENDO PARÂMETROS GERADOS COM O ARQUIVO 3	LENDO PARÂMETROS GERADOS COM O ARQUIVO 4	LENDO PARÂMETROS GERADOS COM A BASE COMPLETA	LENDO PARÂMETROS GERADOS COM O ARQUIVO 3	LENDO PARÂMETROS GERADOS COM O ARQUIVO 4
4 EF	189.92	189.86	190.13	45.60	45.57	45.54
8 EF	229.29	228.95	229.74	46.18	46.37	45.93
3 EM	253.43	252.81	254.54	47.99	48.01	47.84

Quadro 5.4 Proficiência em Matemática, por série no programa Nova Escola 2006 em 3 situações de leitura de parâmetros de itens: Parâmetros gerados com a base completa, com a base de cadernos pares (arquivo 3) e com a base de cadernos ímpares (arquivo 4).

Conforme pode ser observado, o fato de se usar para a *linkagem* com o SAEB a base completa do Nova Escola ou bases separadas, os valores das proficiências são muito próximos, tanto para Língua Portuguesa quanto para Matemática.

Faremos agora, uma análise sobre os valores dos parâmetros dos itens nos diferentes tipos de *linkagens*. Para tanto, apresentamos no quadro 5.5 os quantitativos de itens que são comuns com o SAEB e itens que são apenas do Nova Escola:

SÉRIE	QUANTIDADE DE ITENS		
	COMUM COM SAEB	APENAS SAERJ	TOTAL
4 EF	39	26	65
8 EF	38	40	78
3 EM	38	40	78
TOTAL	115	106	221

Quadro 5.5 Quantitativos de itens de Língua Portuguesa e Matemática no Nova Escola 2006 – itens apenas do Nova Escola e itens comuns com o SAEB 2003

Constatamos, nesse quadro, uma incidência muito alta de itens comuns com o SAEB, 60% na 4ª série e algo em torno de 49 % nas duas outras séries. Este fato favorece a qualidade da comparação dos resultados entre o SAEB e o Nova Escola, uma vez que a literatura recomenda, segundo Kolen e Brennan (2004), no mínimo 20% de itens comuns entre as avaliações nas quais se pretende colocar na mesma escala e com isso terem seus resultados comparáveis.

Verificaremos a seguir os valores dos parâmetros dos itens obtidos nos diferentes tipos de *linkagens*. Apresentamos, nas planilhas do Anexo 2, os valores dos parâmetros *b* dos itens, calibrados nas três situações descritas anteriormente, para cada disciplina, ou seja, lendo os parâmetros obtidos com a base completa, e lendo os parâmetros gerados de acordo com os 4 arquivos descritos anteriormente. Cabe ressaltar que, como os itens comuns com o SAEB tiveram seus parâmetros fixados com os mesmos valores nestes 3 processos de *linkagens*, os mesmos permaneceram inalterados e, por esta razão, não foram apresentados nas referidas planilhas. Portanto, os valores dos parâmetros apresentados, são apenas dos itens exclusivos do programa Nova Escola e que foram deixados livres nos três tipos de *linkagens*, para cada uma das disciplinas consideradas.

Nestas planilhas apresentamos nas primeiras colunas os valores dos parâmetros *b* para cada uma das *linkagens* estudadas e nas duas últimas colunas as diferenças entre os valores dos parâmetros b obtidos ao se utilizar as bases dos arquivos separados e a base com todos os alunos.

Concentraremos nossas análises no comportamento do parâmetro *b* dos itens nos diferentes tipos de *linkagens*. Assim, utilizamos gráficos do tipo *scatter*, com o intuito de diagnosticarmos os quantitativos de itens em situações de grandes e poucas alterações de valores de seus parâmetros.

Nos gráficos do tipo *scatter*, teremos no eixo x os valores dos parâmetros *b* com a base completa e no eixo y, teremos as diferenças entre os valores dos parâmetros *b* com as bases separadas e com a base completa, ou seja, arquivos 1 e 2 para Língua Portuguesa e arquivos3 e 4 para Matemática:

Dif_ b_ÍMPAR_JUNTO = b_ÍMPAR - b_JUNTO
Dif_b_PAR_JUNTO = b_PAR - b_JUNTO

Dessa forma, obtivemos 6 gráficos para cada disciplina os quais apresentamos nas figuras 5.1 e 5.2. Nessas figuras, os gráficos foram posicionados de forma que os três gráficos da esquerda representam a situação de maior proficiência e os da esquerda a situação de menor proficiência. Assim, há uma alternância entre as ordens nos gráficos de Língua Portuguesa e Matemática para que seja respeitada essa condição.

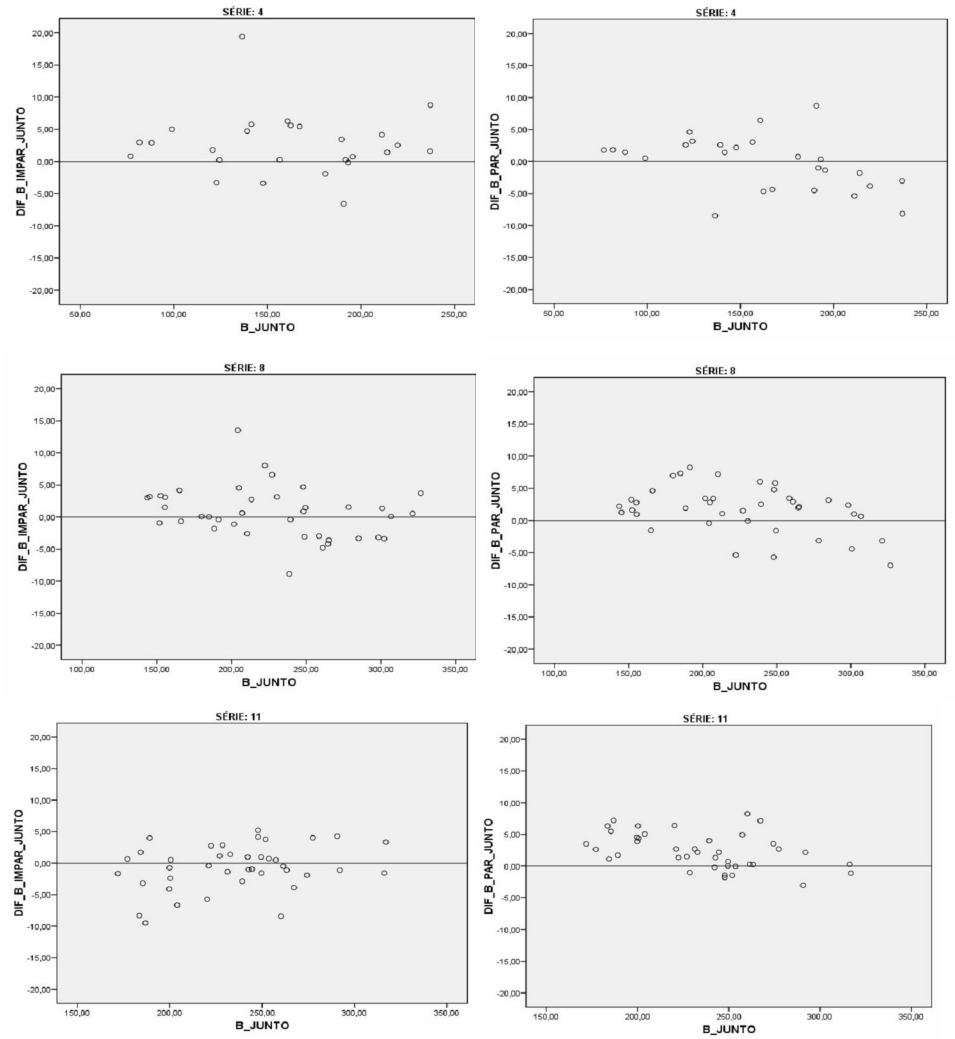

Figura 5.1: *Scatter plots* para o parâmetro *b* em Língua Portuguesa: Parâmetro *b*_junto por Parâmetro *b*_arquivos separados - *b*_junto

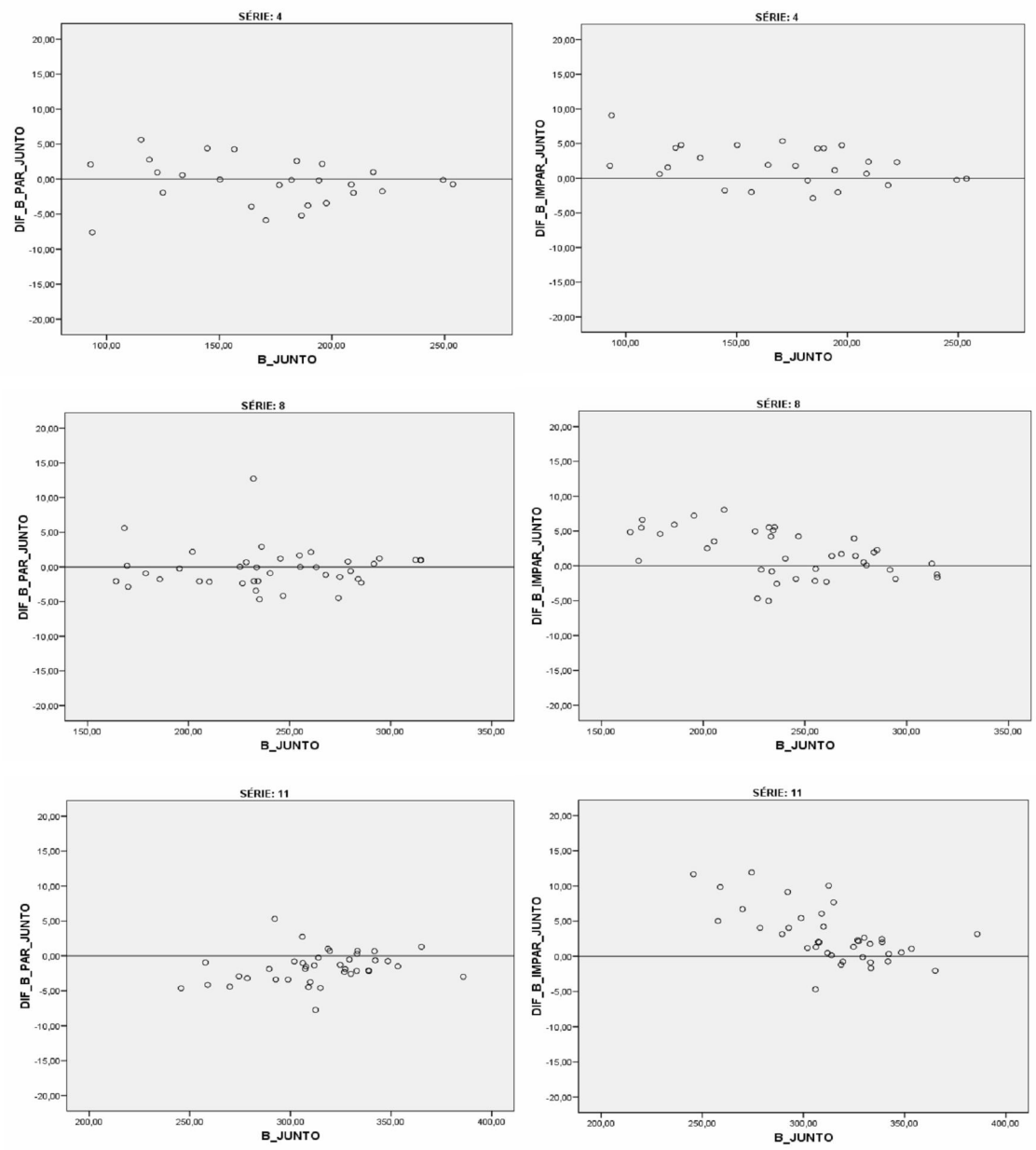

Figura 5.2: *Scatter plots* para o parâmetro *b* em Matemática: Parâmetro *b*_arquivos separados – b_junto

A análise das figuras 5.1 e 5.2, indica-nos que: i) as variações nas diferenças dos parâmetros b dos itens estão bem concentradas no intervalo de -10 a +10; ii) para o 3º ano EM, período de escolarização em que tivemos as maiores divergências de proficiências, conforme apresentado no quadro 5.1, existe uma leve tendência a termos uma superestimação de itens fáceis e subestimação de itens difíceis nos grupos de menores proficiências (gráficos da direita), e uma situação inversa nos grupos de maiores proficiências, ou seja uma subestimação de itens fáceis e uma superestimação de itens difíceis (gráficos da esquerda). Essa característica ficará

mais evidente, no próximo tópico, ao estudarmos situações em que as diferenças de proficiências entre as populações são maiores.

Entendemos que o efeito cansaço, como demonstrado na seção 5.5.1, provocado pelo posicionamento dos itens no final do caderno, traz variações nos parâmetros dos itens, porém, essas variações não alteram significativamente as proficiências dos alunos. Isto pode estar relacionado a dois fatores: a quantidade de itens comuns com o SAEB os quais permaneceram com seus valores fixos, "segura" a proficiência dos alunos e/ou como as variações nas proficiências dos grupos utilizados para calibrar os parâmetros, ou seja, considerando apenas cadernos ímpares ou cadernos pares são pequenas, as variações nos parâmetros dos itens observadas nos gráficos anteriores não são suficientes para alterar significativamente os valores das proficiências.

Analisamos, a seguir, uma situação um pouco mais desfavorável em termos de *linkagem*, pois, embora os testes sejam exatamente os mesmos nas diferentes abordagens, as populações possuem características diferentes, com variações nas médias de proficiências superiores aos casos até então abordados.

5.2.2.2 Populações diferentes submetidas ao mesmo design de teste e mesmos itens

No ano de 2007, tivemos em âmbito nacional duas avaliações: SAEB e Prova Brasil, embora tenhamos retratado as características dessas avaliações no capítulo 2 em termos de designs de testes, falta apresentar as características dos procedimentos de calibração de itens e geração de proficiências utilizadas nessas avaliações. O entendimento desses procedimentos é essencial para os estudos utilizados nesse tópico.

Devido ao fato de os itens da Prova Brasil 2007 serem os mesmos do SAEB 2007 a calibração dos mesmos foi realizada através de um procedimento de equalização vertical entre as bases do SAEB 2005 e SAEB 2007. Esse procedimento garante duas características importantes: i) manter as mesmas características das *linkagens* realizadas pelo SAEB ao longo dos anos, no que se refere a sintaxe e base de dados e ii) uma vez calibrados os parâmetros dos itens, através de uma base de dados que a partir de agora denominaremos base amostra Brasil, as proficiências

da Prova Brasil, para cada Unidade da Federação – UF[7], foi calculada apenas lendo esses parâmetros através do BILOGMG.

Tendo como referência as bases da PROVA BRASIL 2007, analisaremos os parâmetros dos itens e proficiências de alunos, desta avaliação, para a 4ª e 8ª séries do EF, através de dois processos de *linkagem*: Calibração via amostra Brasil versus calibração dentro da Unidade da federação - UF. Cabe ressaltar que a calibração dos itens via amostra nacional, envolve escolas públicas e privadas e são os resultados oficiais do SAEB e a calibração dentro de cada UF foi calculada por nós, tendo como referência as bases da Prova Brasil, portanto, não são os resultados oficiais. Nosso propósito neste tópico é verificar possíveis divergências nos parâmetros e proficiências, ao compararmos os parâmetros dos itens calibrados dentro da própria UF com alunos apenas da rede pública de ensino, que é a característica da Prova Brasil, com os calculados pelo SAEB.

Para a realização deste estudo, selecionamos 6 (seis) UFs com diferentes características em termos de médias e desvios padrão, de forma a termos: i) UFs com médias inferiores à média do Brasil, ii) UF com média próxima à média do Brasil e iii) UF com médias superiores à média do Brasil. Ao todo serão 12 (doze) análises, tendo em vista as disciplinas de Língua Portuguesa e de Matemática.

Para essa seleção, comparamos os resultados do Brasil gerados pelo SAEB com os resultados de cada UF fornecidos pela Prova Brasil. Assim, no quadro 5.6, temos os resultados gerais para Língua portuguesa e Matemática, e no quadro 5.7 os resultados de proficiência da Prova Brasil 2007 para cada UF em Língua Portuguesa e em Matemática.

SÉRIE	PROFICIÊNCIA			
	LÍNGUA PORTUGUESA		MATEMÁTICA	
4 EF	174.6	42.9	192.6	44.9
8 EF	232.5	47.2	245.5	48.1
3 EM	261.4	51.8	272.9	54.9

Quadro 5.6: Proficiência em Língua Portuguesa e Matemática – SAEB 2007. Fonte: SAEB

[7] Em virtude da inclusão do Distrito Federal nas análises, faz-se necessário substituir a palavra estado por unidade da federação - UF

ESTADO	LÍNGUA PORTUGUESA						MATEMÁTICA					
	SÉRIE 4EF			SÉRIE 8EF			SÉRIE 4EF			SÉRIE 8EF		
	MÉDIA	D. PAD	ALUNOS	MÉDIA	D. PAD	ALUNOS	MÉDIA	D. PAD	ALUNOS	MÉDIA	D. PAD	ALUNOS
RN	151.0	35.3	33731	218.3	41.6	22663	168.9	36.7	33703	230.2	40.6	22648
AL	155.0	33.6	37162	210.6	39.8	27645	172.0	34.8	37155	221.9	38.1	27632
MA	157.6	36.4	73476	216.6	41.2	53513	174.6	38.1	73460	223.4	39.0	53495
PE	157.8	36.4	80616	211.7	41.4	69287	174.1	37.2	80557	221.9	39.4	69166
CE	159.4	37.7	87384	217.3	41.8	75811	174.6	38.1	87371	226.6	39.9	75795
AP	160.1	35.8	11281	220.0	39.9	7168	173.7	34.9	11274	226.0	35.8	7167
PA	160.4	34.6	89425	222.7	39.4	49731	175.1	35.1	89442	230.7	37.5	49743
PB	161.2	36.5	37968	217.0	41.2	31415	178.5	37.9	37948	226.9	38.8	31396
SE	161.3	34.7	19715	218.2	41.0	13443	177.8	35.8	19712	230.3	39.6	13437
BA	162.1	35.8	124686	217.5	42.3	95026	177.2	36.4	124675	227.2	39.6	95008
PI	162.6	36.0	31320	218.3	41.6	22068	178.0	37.2	31323	231.6	41.0	22062
AM	164.9	36.9	50804	226.4	41.2	35161	179.8	37.7	50779	232.8	41.5	35125
TO	166.4	37.6	21716	223.2	42.1	16142	181.1	38.8	21725	231.6	40.2	16143
RO	168.5	37.3	21379	226.3	40.5	14016	184.5	37.8	21379	238.9	39.2	14008
RR	170.6	36.8	6512	224.9	40.8	4255	185.3	37.0	6515	235.6	39.7	4256
GO	170.7	38.5	68147	225.9	42.0	59641	186.5	39.5	68121	237.8	40.4	59578
AC	171.1	37.4	10224	224.3	40.5	6310	182.8	36.6	10217	233.0	37.3	6310
MT	172.9	37.9	39046	226.2	41.9	29678	189.4	39.7	39028	239.2	42.0	29688
RJ	176.6	40.3	168648	230.2	45.2	109753	192.8	41.6	168506	238.1	44.4	109641
MS	178.2	37.9	35948	238.5	40.7	21777	195.8	40.3	35927	252.2	41.3	21769
ES	178.2	39.5	44014	231.3	42.9	32493	195.2	41.2	44008	245.3	43.3	32464
MG	179.9	44.0	261594	237.2	44.8	228890	199.6	45.8	261512	252.6	45.7	228869
RS	179.9	39.6	128387	239.1	42.2	90105	197.8	40.7	128344	251.4	42.5	90071
SP	180.5	44.1	579266	232.3	45.5	494293	198.8	46.5	579049	243.3	43.7	494281
SC	181.1	39.6	77841	235.5	42.2	61943	199.8	42.4	77831	251.6	42.0	61921
PR	184.6	39.4	134508	235.7	42.1	105431	205.2	42.8	134505	252.2	41.8	105430
DF	191.2	39.5	31256	238.0	44.4	20133	208.8	41.1	31249	252.2	43.8	20133

Quadro 5.7: Proficiência Prova Brasil por Unidade da Federação - UF. Fonte: SAEB

Selecionamos, considerando o quadro 5.7, o qual está ordenado segundo a proficiência em Língua Portuguesa na 4ª série EF, 6 (seis) UFs: Rio Grande do Norte, Ceará, Espírito Santo, Mato Grosso do Sul, Minas Gerais, do Sul e Distrito Federal. Em seguida, agrupamos essas UFs segundo os valores de suas proficiências estarem abaixo, próxima ou acima da média nacional, para cada série e disciplina avaliada na Prova Brasil, formando assim os grupos 1, 2 e 3 respectivamente.

Nos quadros 5.8 e 5.9, apresentamos as médias e desvios-padrão das proficiências e respectivos grupos para as UFs selecionadas

UF	4ª SÉRIE EF				GRUPO
	LÍNGUA PORTUGUESA		MATEMÁTICA		
	MÉDIA	DES. PAD	MÉDIA	DES. PAD	
RN	151	35	169	37	1
CE	159	38	174	38	1
ES	178	39	195	41	2
MS	178	38	196	40	2
MG	179	44	199	46	2
DF	191	39	208	41	3

Quadro 5.8: Proficiência Prova Brasil por UF, agregadas por grupo. Fonte: SAEB

UF	8ª SÉRIE EF				GRUPO
	LÍNGUA PORTUGUESA		MATEMÁTICA		
	MÉDIA	DES. PAD	MÉDIA	DES. PAD	
RN	218	42	230	41	1
CE	217	42	227	40	1
ES	231	43	245	43	2
MS	238	41	252	41	3
MG	237	44	252	46	3
DF	238	44	252	43	3

Quadro 5.9: Proficiência Prova Brasil por UF, agregadas por grupo. Fonte: SAEB

Temos, portanto, para a 4ª série EF, RN e CE, no grupo, ES, MS e MG, no grupo 2, e DF, no grupo 3. Para a 8ª série EF, RN e CE continuam no grupo 1, ES,no grupo 2, e MS, MG e DF, no grupo 3.

Realizamos, então, procedimentos de *linkagens* entre o SAEB 2005 e as bases da Prova Brasil 2007 dentro de cada UF selecionada. Utilizamos como referência a sintaxe oficial do BILOGMG utilizada pelo INEP para *linkar* o SAEB 2005 com SAEB 2007. Porém, como a Prova Brasil não avalia o 3º EM, fizemos alterações nessa sintaxe de forma a ajustá-la para nossos estudos. Estas alterações consistiram basicamente em eliminar das sintaxes as formas e grupos referentes a este ano de escolaridade, as demais informações como priores de 2005, métodos de calibração de itens e geração de escores permaneceram inalterados.

Nos quadros 5.10 e 5.11, apresentamos, para cada UF selecionada, os valores médios e desvios-padrão das proficiências, apresentados oficialmente pela Prova Brasil, e os novos valores obtidos pelas *linkagens* dentro das UFs:

LÍNGUA PORTUGUESA							
UNIDADE DA FERERAÇÃO	SÉRIE	PROFICIÊNCIA					
		MÉDIA			DESV. PAD.		
		REF. 1	REF. 2	DIF	REF. 1	REF. 2	DIF
RIO GRANDE DO NORTE	4 EF	151	149.6	-1.4	35.3	35.1	-0.2
	8EF	218.3	218.2	-0.1	41.6	40.6	-1
CEARÁ	4 EF	159.4	158.7	-0.7	37.7	37.5	-0.2
	8EF	217.3	217.3	0	41.8	40.7	-1.1
ESPÍRITO SANTO	4 EF	178.2	177.5	-0.7	39.5	39.4	-0.1
	8EF	231.3	231.3	0	42.9	42.9	0
MATO GROSSO DO SUL	4 EF	178.2	177	-1.2	37.9	38.2	0.3
	8EF	238.5	237.6	-0.9	40.7	40.5	-0.2
MINAS GERAIS	4 EF	179.9	181.6	1.7	44	42.6	-1.4
	8EF	237.2	238	0.8	44.8	44.1	-0.7
DISTRITO FEDERAL	4 EF	191.2	195.7	4.5	39.5	38.8	-0.7
	8EF	238	242.3	4.3	44.4	44.2	-0.2

Quadro 5.10: Comparação das proficiências, em Língua Portuguesa, entre os resultados oficiais da Prova Brasil, (REF.1), os obtidos através de *linkagens* dentro das UFs, (REF.2) e a diferença ente REF.2 e REF.1 (DIF)

MATEMÁTICA							
UNIDADE DA FERERAÇÃO	SÉRIE	PROFICIÊNCIA					
		MÉDIA			DESV. PAD.		
		REF. 1	REF. 2	DIF	REF. 1	REF. 2	DIF
RIO GRANDE DO NORTE	4 EF	168.9	164.3	-4.6	36.7	36.1	-0.6
	8EF	230.2	227	-3.2	40.6	40	-0.6
CEARÁ	4 EF	174.6	173.1	-1.5	38.1	37.6	-0.5
	8EF	226.6	224.8	-1.8	39.9	39.1	-0.8
ESPÍRITO SANTO	4 EF	195.2	193.4	-1.8	41.2	41.3	0.1
	8EF	245.3	244	-1.3	43.3	43	-0.3
MATO GROSSO DO SUL	4 EF	195.8	196.1	0.3	40.3	39.9	-0.4
	8EF	252.2	250.8	-1.4	41.3	40.8	-0.5
MINAS GERAIS	4 EF	199.6	201.3	1.7	45.8	44.8	-1
	8EF	252.6	252.8	0.2	45.7	45.5	-0.2
DISTRITO FEDERAL	4 EF	208.8	210.4	1.6	41.1	40.08	-1.02
	8EF	252.2	252.8	0.6	43.8	43.5	-0.3

Quadro 5.11: Comparação das proficiências, em matemática, entre os resultados oficiais da Prova Brasil, (REF.1), os obtidos através de *linkagens* dentro das UFs, (REF.2) e a diferença ente REF.2 e REF.1 (DIF)

Analisando esses quadros, observamos uma tendência a termos a nova estimativa da proficiência dada pela variável REF.2 menor que a proficiência oficial do SAEB, dada pela variável REF.1, quando a UF tem proficiência inferior à média nacional e para as UFs com médias superiores à média nacional ocorre uma situação inversa, ou seja, a variável REF.2 é maior que a variável REF1. O extremo dessas variações é observado em Língua Portuguesa, no DF, e, em Matemática, no RN, que são as UFs com a maior e menor proficiência, respectivamente.

Para analisarmos os efeitos nas estimativas dos parâmetros dos itens em função dos dois tipos de *linkagens* mencionados, utilizamos, assim como no tópico anterior, gráficos do tipo *scatter* com o intuito de diagnosticarmos os quantitativos de itens em situações de grandes e poucas alterações de valores de seus parâmetros.

Concentraremos nossas análises no comportamento do parâmetro *b* dos itens nos diferentes tipos de *linkagens*. Assim, no anexo 3, apresentamos as tabelas com os valores dos parâmetros *b* dos itens nesses dois tipos de *linkagens* para as duas disciplinas nas 6 (seis) UFs consideradas, bem como as diferenças entre os valores dos parâmetros *b* obtidos na *linkagem* dentro da UF com os valores oficiais do SAEB, ou seja:

$$Dif\_b\_UF = (b\_UF - b\_SAEB) \qquad (5.1)$$

Esses valores foram trabalhados em gráficos do tipo *scatter plot*, onde, no eixo x, temos os valores dos parâmetros *b* oficiais do SAEB e, no eixo y, as diferenças calculadas pela equação 5.1.

Dividiremos nosso estudo em 4 etapas, em função das disciplinas e séries avaliadas, desta forma teremos:

Etapa 1: Língua Portuguesa 4ª série para as UFs: RN, CE, ES, MS, MG e DF
Etapa 2: Matemática 4ª série para as UFs: RN, CE, ES, MS, MG e DF
Etapa 3: Língua Portuguesa 8ª série para as UFs: RN, CE, ES, MS, MG e DF
Etapa 4: Matemática 8ª série para as UFs: RN, CE, ES, MS, MG e DF

Para cada etapa apresentaremos as seguintes análises :
1 – Variações dos parâmetros b dos itens em cada UF: através de gráficos do tipo *scatter* entre os valores dos parâmetros b do SAEB e suas respectivas diferenças entre a UF e o SAEB. Nestes gráficos, itens com diferenças nos parâmetros b, dada pela equação 5.1, maiores que 40 pontos, serão assinalados para análise.
2 - Itens com valores discrepantes do parâmetro *b*: gráficos com as curvas características dos itens com valores discrepantes do parâmetro *b*, conforme mencionado anteriormente.

Etapa 1: Língua Portuguesa 4ª série

Análise1_etapa1: Variações dos parâmetros b dos itens em cada UF

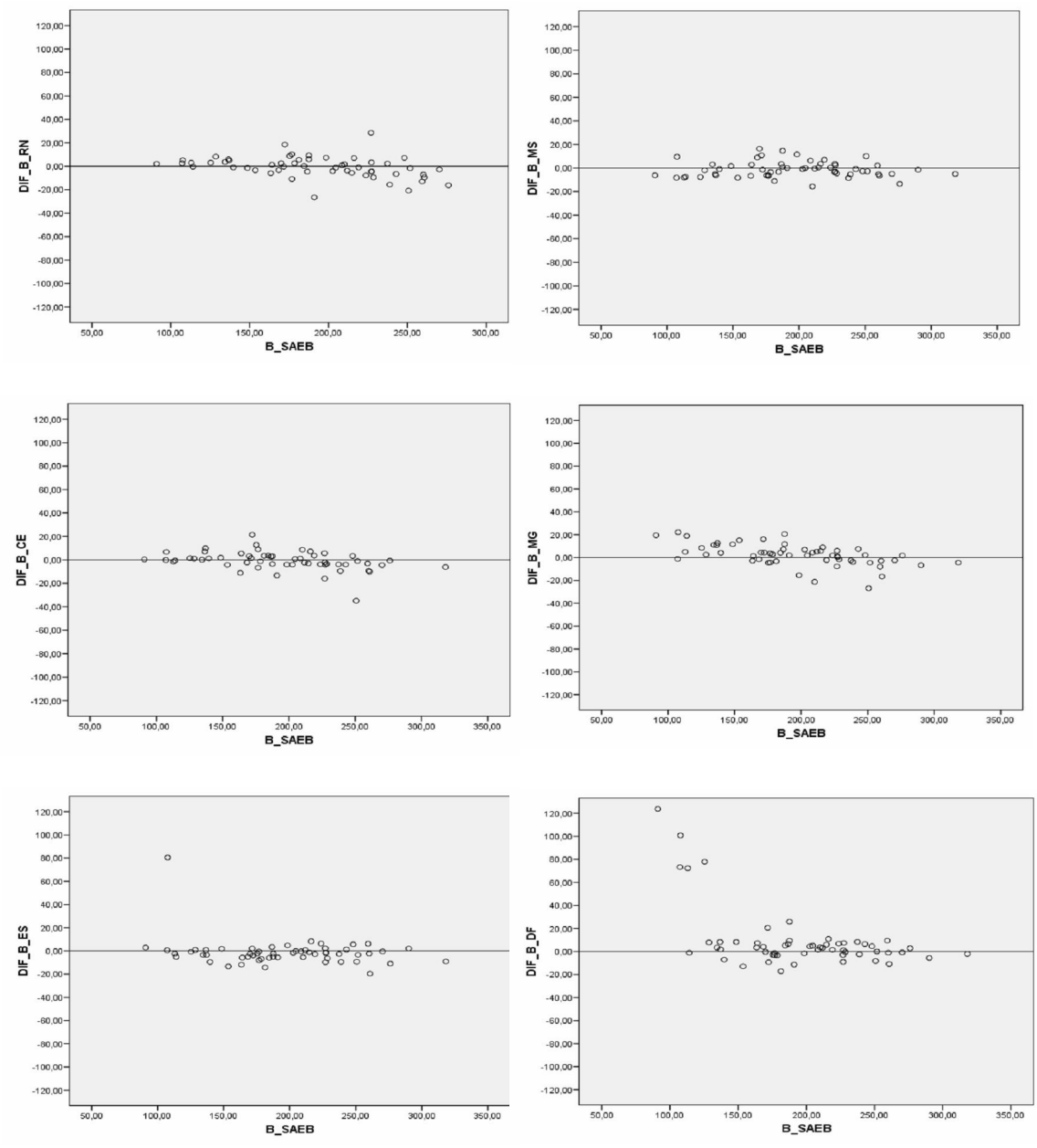

Figura 5.3: Gráficos do tipo *scatter* entre os valores dos parâmetros b do SAEB e suas respectivas diferenças entre a UF e o SAEB. Língua Portuesa 4ª série.

Análise 2_etapa1: Itens com valores discrepantes do parâmetro *b*

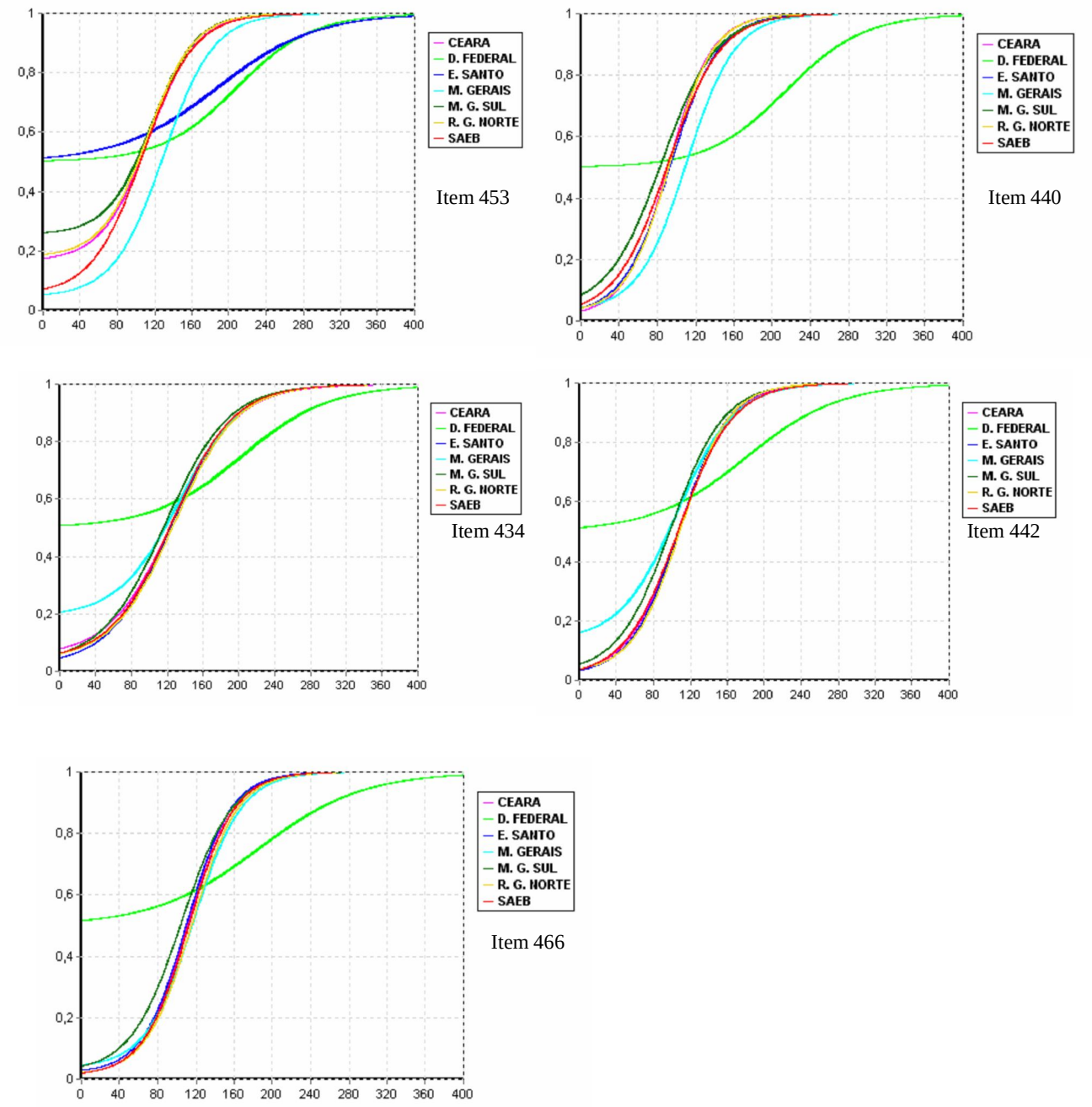

Figura 5.4: Gráficos com as curvas características dos itens com valores discrepantes do parâmetro *b*. Língua Portuguesa 4ª série.

Etapa 2: Matemática 4ª série

Análise1_etapa2: Variações dos parâmetros b dos itens em cada UF

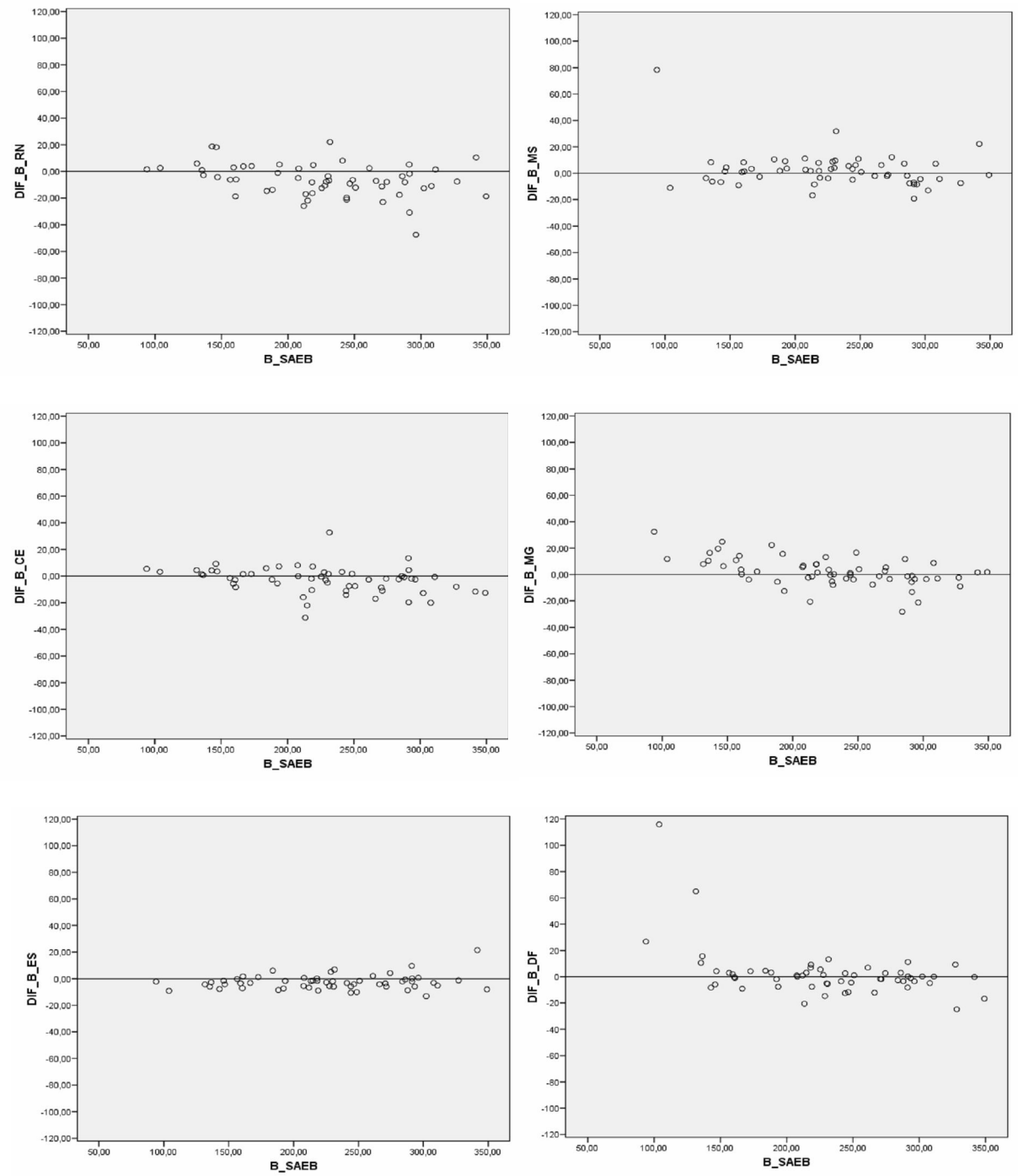

Figura 5.5: Gráficos do tipo *scatter* entre os valores dos parâmetros b do SAEB e suas respectivas diferenças entre a UF e o SAEB. Matemática4ª série

Análise 2_etapa2: Itens com valores discrepantes do parâmetro *b*

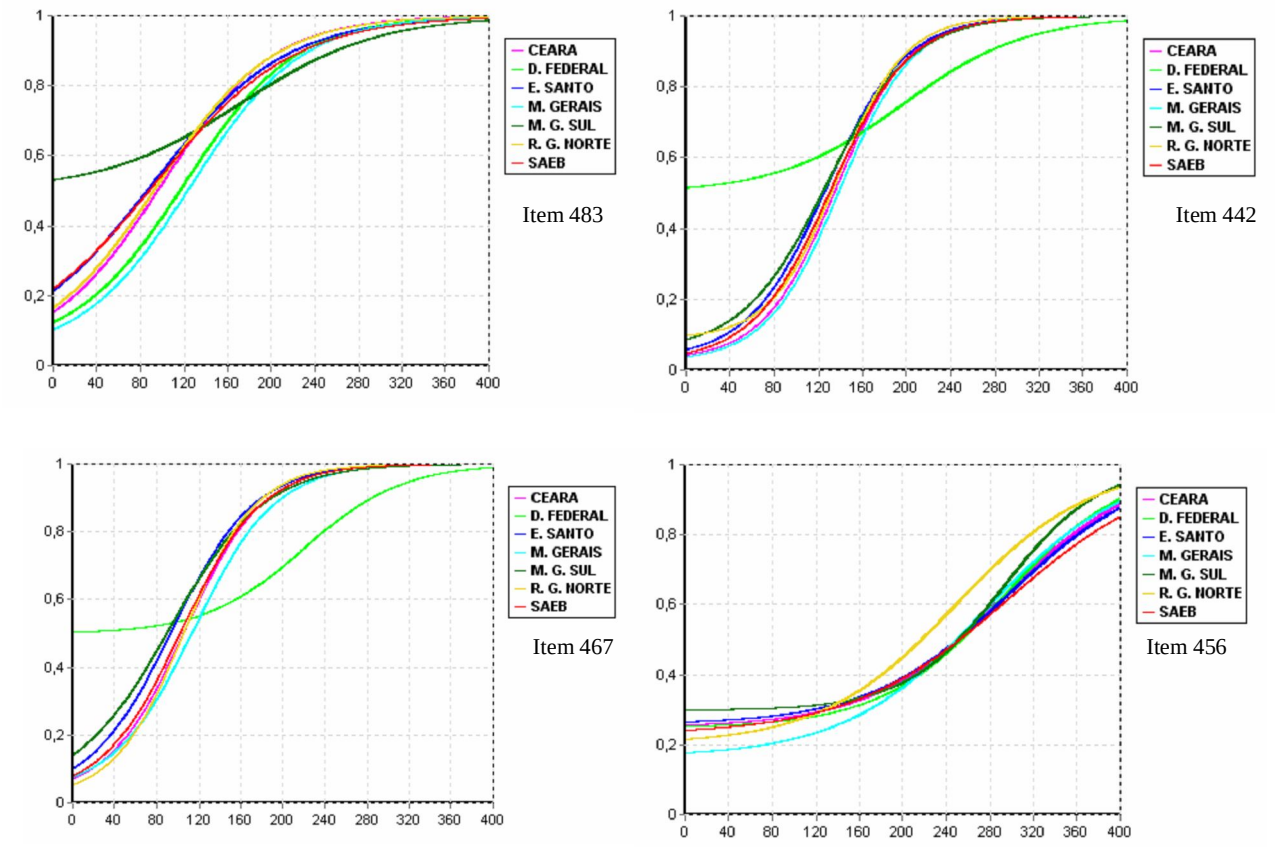

Figura 5.6: Gráficos com as curvas características dos itens com valores discrepantes do parâmetro *b*. Matemática 4ª série.

Etapa 3: Língua Portuguesa 8ª série

Análise1_Etapa3: Variações dos parâmetros b dos itens em cada UF

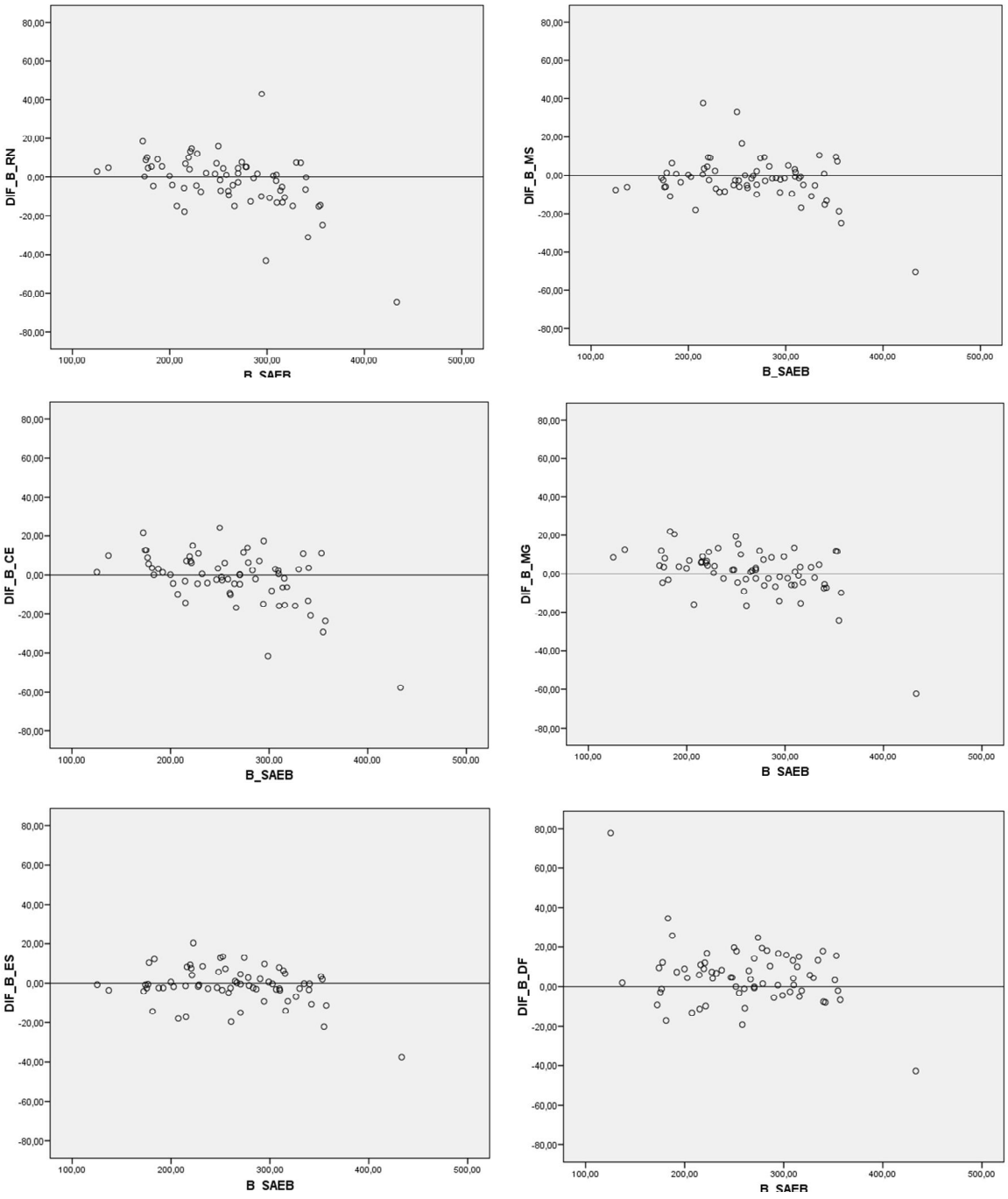

Figura 5.7: Gráficos do tipo *scatter* entre os valores dos parâmetros b do SAEB e suas respectivas diferenças entre a UF e o SAEB. Língua Portuguesa 8ª série.

Análise 2_etapa3: Itens com valores discrepantes do parâmetro *b*

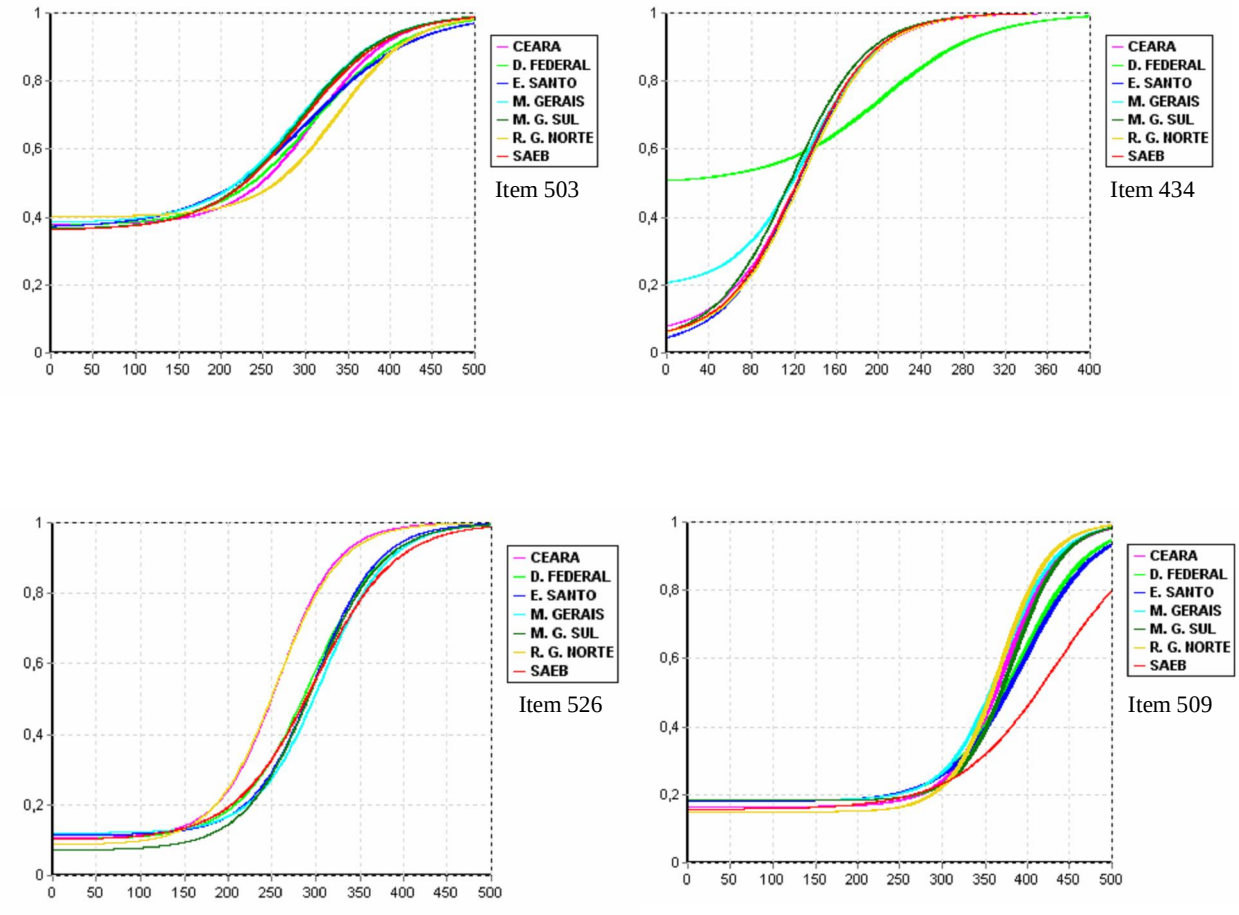

Figura 5.8: Gráficos com as curvas características dos itens com valores discrepantes do parâmetro *b*. Língua Portuguesa 8ª série.

Etapa 4: Matemática 8ª série

Análise1_etapa4: Variações dos parâmetros b dos itens em cada UF

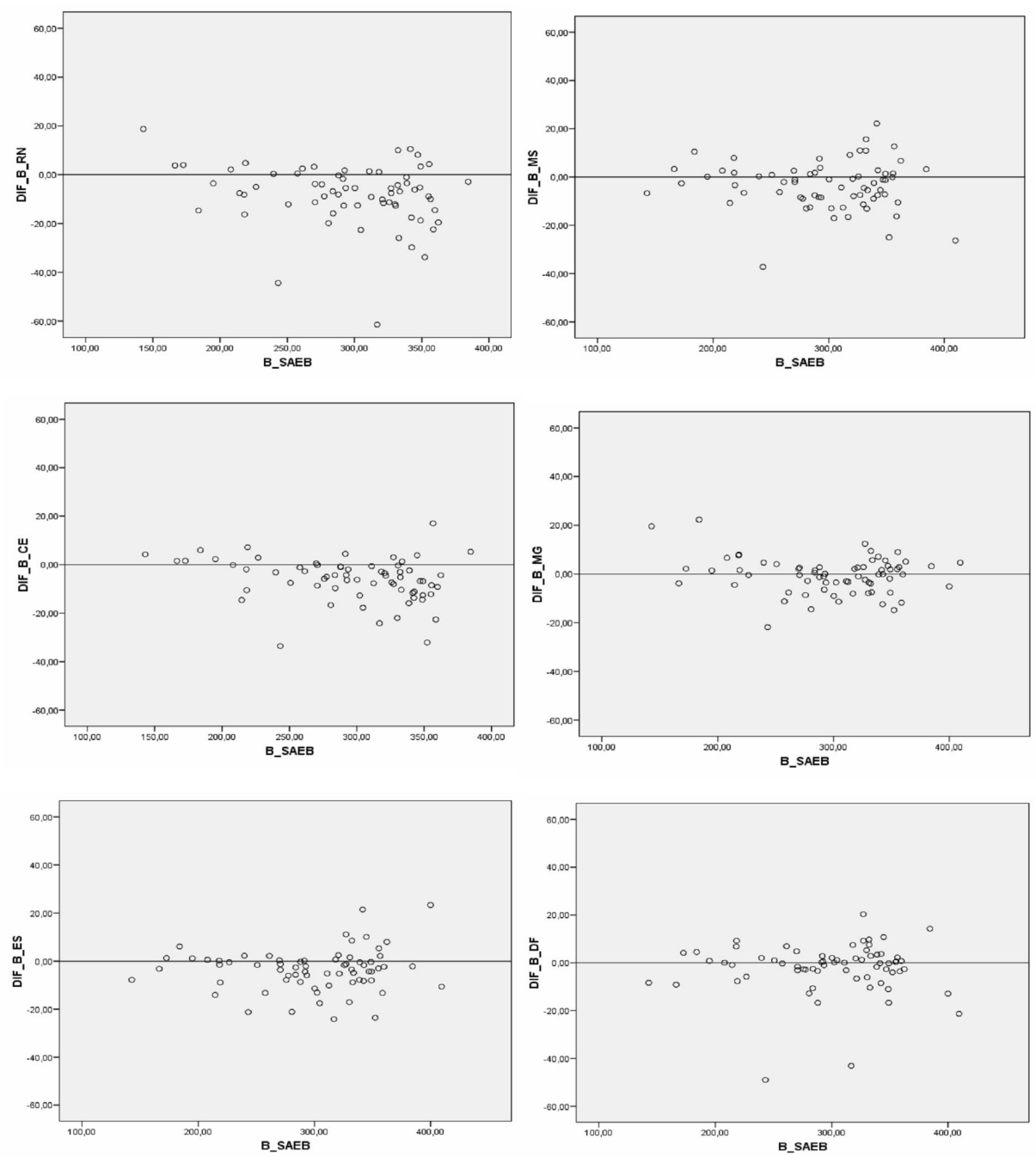

Figura 5.9: Gráficos do tipo *scatter* entre os valores dos parâmetros b do SAEB e suas respectivas diferenças entre a UF e o SAEB. Matemática 8ª série.

Análise2_etapa4: valores dos parâmetros b dos itens em cada UF

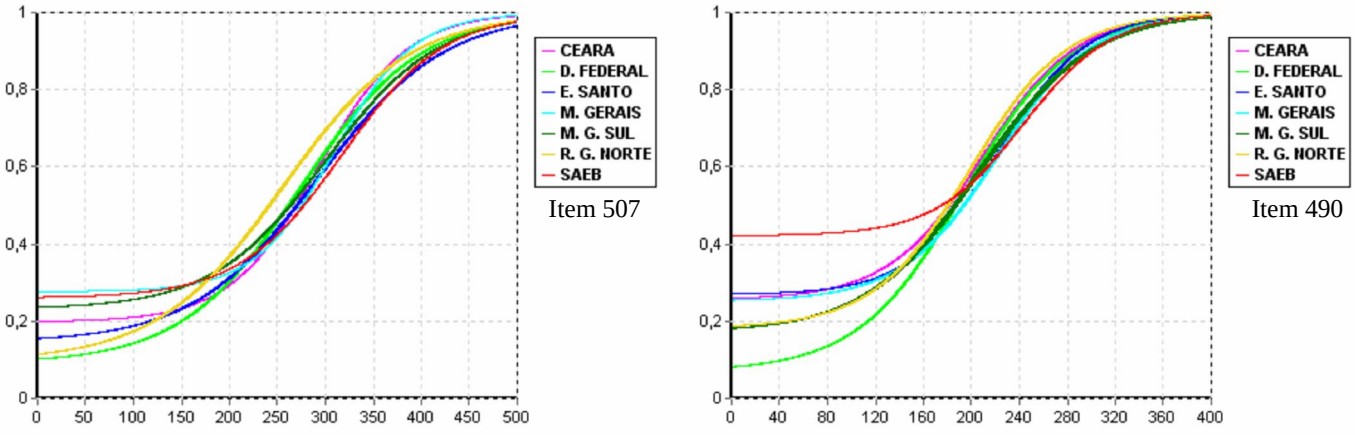

Figura 5.10: Gráficos com as curvas características dos itens com valores discrepantes do parâmetro *b*. Matemática 8ª série.

Apresentamos, nos quadros 5.12 e 5.13, os quantitativos de itens eliminados por bisserial negativa, bem como o quantitativo de itens com diferença entre o parâmetro *b* calibrado na UF e seu respectivo valor no SAEB maior que 40 pontos, ou seja, valores discrepantes do parâmetro *b*.

UF	LÍNGUA PORTUGUESA					
	4ª SÉRIE			8ª SÉRIE		
	BISSERIAL NEGATIVA	Dif_b_UF >40	TOTAL	BISSERIAL NEGATIVA	Dif_b_UF >40	TOTAL
RN	3	0	3	4	3	7
CE	2	0	2	4	2	6
ES	1	1	2	3	1	4
MS	1	0	1	3	1	4
MG	1	0	1	3	1	4
DF	1	5	6	3	2	5

Quadro 5.12: Número de itens com bisserial negativa e valores discrepantes do parâmetro *b* em Língua Portuguesa, por série e UF.

UF	MATEMÁTICA					
	4ª SÉRIE			8ª SÉRIE		
	BISSERIAL NEGATIVA	Dif_b_UF >40	TOTAL	BISSERIAL NEGATIVA	Dif_b_UF >40	TOTAL
RN	3	1	4	4	2	6
CE	2	0	2	4	0	4
ES	2	0	2	2	0	2
MS	2	1	3	3	0	3
MG	1	0	1	2	0	2
DF	1	2	3	2	2	4

Quadro 5.13: Número de itens com bisserial negativa e valores discrepantes do parâmetro *b* em Matemática, por série e UF.

Pela análise desses dois últimos quadros, observamos que os maiores quantitativos de itens com bisseriais negativas e discrepâncias nos valores dos parâmetros b, ocorrem nas UFs do Rio Grande do Norte e Distrito Federal, que correspondem às UFs com a menor e maior proficiência, respectivamente. Já para as UFs com proficiências em torno da média Brasil, as incidências de itens com as duas características mencionadas são menores. Esta constatação está coerente com fato de que as melhores *linkagens* ocorrem quando as populações são equivalentes.

Como nosso objetivo é verificar o quanto as diferenças entre as populações interferem nos resultados das *linkagens*, refizemos os cálculos das proficiências das UFs eliminando os itens com discrepâncias no parâmetro *b* maiores que 40 pontos. No quadro 5.14 apresentamos os novos valores encontrados para Língua Portuguesa e no quadro 5.15 os novos valores para Matemática:

LÍNGUA PORTUGUESA							
UNIDADE DA FERERAÇÃO	SÉRIE	PROFICIÊNCIA					
		MÉDIA			DESV. PAD.		
		REF. 1	REF. 2	DIF	REF. 1	REF. 2	DIF
RIO GRANDE DO NORTE	4 EF	151	149,8	-1,2	35,3	35,1	-0,2
	8EF	218,3	218,3	0	41,6	40,6	-1
CEARÁ	4 EF	159,4	158,8	-0,6	37,7	37,6	-0,1
	8EF	217,3	217,5	0,2	41,8	40,7	-1,1
ESPÍRITO SANTO	4 EF	178,2	176,7	-1,5	39,5	39,5	0
	8EF	231,3	231,1	-0,2	42,9	43	0,1
MATO GROSSO DO SUL	4 EF	178,2	177	-1,2	37,9	38,2	0,3
	8EF	238,5	237,7	-0,8	40,7	40,5	-0,2
MINAS GERAIS	4 EF	179,9	181,6	1,7	44	42,5	-1,5
	8EF	237,2	237,8	0,6	44,8	44	-0,8
DISTRITO FEDERAL	4 EF	191,2	191,6	0,4	39,5	38,9	-0,6
	8EF	238	240,3	2,3	44,4	44,9	0,5

Quadro 5.14: Comparação das proficiências, em Língua Portuguesa, entre os resultados oficiais da Prova Brasil, (REF.1), e os obtidos através de *linkagens* dentro das UFs, (REF.2) após eliminação de itens com discrepâncias no parâmetro *b* maior que 40 pontos.

MATEMÁTICA							
UNIDADE DA FERERAÇÃO	SÉRIE	PROFICIÊNCIA					
		MÉDIA			DESV. PAD.		
		REF. 1	REF. 2	DIF	REF. 1	REF. 2	DIF
RIO GRANDE DO NORTE	4 EF	168,9	164,6	-4,3	36,7	36,2	-0,5
	8EF	230,2	227,2	-3,0	40,6	39,7	-0,9
CEARÁ	4 EF	174.6	173.1	-1,5	38,1	37.6	-0,5
	8EF	226.6	224.8	-1,8	39,9	39.1	-0,8
ESPÍRITO SANTO	4 EF	195.2	193.4	-1,8	41,2	41.3	0,1
	8EF	245.3	244	-1,3	43,3	43	-0,3
MATO GROSSO DO SUL	4 EF	195,8	195,8	0,0	40,3	40	-0,3
	8EF	252,2	250,8	-1,4	41,3	40,8	-0,5
MINAS GERAIS	4 EF	199,6	201.3	1,7	45,8	44.8	-1,0
	8EF	252,6	252.8	0,2	45,7	45.5	-0,2
DISTRITO FEDERAL	4 EF	208,8	208,4	-0,4	41,1	40,4	-0,7
	8EF	252,2	252,1	-0,1	43,8	43,7	-0,1

Quadro 5.15: Comparação das proficiências, em Matemática, entre os resultados oficiais da Prova Brasil, (REF.1), e os obtidos através de *linkagens* dentro das UFs, (REF.2) após eliminação de itens com discrepâncias no parâmetro *b* maior que 40 pontos.

Apresentamos no quadro 5.16, a seguir, as diferenças entre os valores de proficiência, média e desvio padrão, considerando os itens com valores discrepantes de b, variável DIF1 e eliminando esses itens. Variável DIF2.

UNIDADE DA FERERAÇÃO	SÉRIE	LÍNGUA PORTUGUESA				MATEMÁTICA			
		MÉDIA		DES. PAD1		MÉDIA		DES. PAD1	
		DIF1	DIF2	DIF1	DIF2	DIF1	DIF2	DIF1	DIF2
RIO GRANDE DO NORTE	4 EF	-1,4	-1,2	-0,2	-0,2	-4,6	-4,3	-0,6	-0,5
	8EF	-0,1	0	-1	-1	-3,2	-3	-0,6	-0,9
CEARÁ	4 EF	-0,7	-0,6	-0,2	-0,1	-1,5	-1,5	-0,5	-0,5
	8EF	0	0,2	-1,1	-1,1	-1,8	-1,8	-0,8	-0,8
ESPÍRITO SANTO	4 EF	-0,7	-1,5	-0,1	0	-1,8	-1,8	0,1	0,1
	8EF	0	-0,2	0	0,1	-1,3	-1,3	-0,3	-0,3
MATO GROSSO DO SUL	4 EF	-1,2	-1,2	0,3	0,3	0,3	0	-0,4	-0,3
	8EF	-0,9	-0,8	-0,2	-0,2	-1,4	-1,4	-0,5	-0,5
MINAS GERAIS	4 EF	1,7	1,7	-1,4	-1,5	1,7	1,7	-1	-1
	8EF	0,8	0,6	-0,7	-0,8	0,2	0,2	-0,2	-0,2
DISTRITO FEDERAL	4 EF	4,5	0,4	-0,7	-0,6	1,6	-0,4	-1,02	-0,7
	8EF	4,3	2,3	-0,2	0,5	0,6	-0,1	-0,3	-0,1

Quadro 5.16: Comparativo entre médias e desvios padrão em Língua Portuguesa e Matemática ao considerarmos itens com valores discrepantes de b maior que 40 pontos (DIF1) e eliminando estes itens (DIF2).

Verificamos, nesse quadro, que de uma forma geral os resultados de proficiência ficam mais próximos dos valores oficiais do SAEB ao eliminarmos os itens com valore discrepantes de b maiores que 40 pontos. Ao analisarmos os resultados do RN e DF, onde tínhamos as maiores divergências de resultados, observamos que no DF a queda na diferença entre os valores de proficiências foi altamente significativa, ou seja, indicando uma melhor aproximação dos valores oficiais. Já para o RN, embora tenha ocorrido a queda na diferença de proficiência entre as situações abordadas, a nova diferença ainda indica uma divergência entre os valores oficiais, ou seja, apenas eliminando os itens com valores discrepantes de b maior que 40 pontos, não foi o suficiente para uma boa convergência de resultados para esta UF.

Vejamos como ficou a nova distribuição dos parâmetros b, ao rodarmos novamente o BILOGMG sem os itens com valores discrepantes de b. Utilizamos as mesmas análises apresentadas anteriormente, considerando apenas a etapa 1, porém, para facilitar a visualização das variações das diferenças dos valores dos parâmetros b, acrescentamos nos gráficos do tipo *scatter*, intervalos referentes aos quartis dos parâmetros b dos itens na equalização oficial do SAEB. Dessa forma, os itens fáceis estão concentrados no 1º quartil, os itens medianos no 2º e 3º quartil e os itens difíceis no 4 quartil, e uma curva cúbica modelando os valores das diferenças dos parâmetros b conforme os gráficos a seguir.

Etapa 1: Língua Portuguesa 4ª série

Análise1_etapa1: Variações dos parâmetros b dos itens em cada UF, após eliminação de itens com valores discrepantes de b maior que 40 pontos.

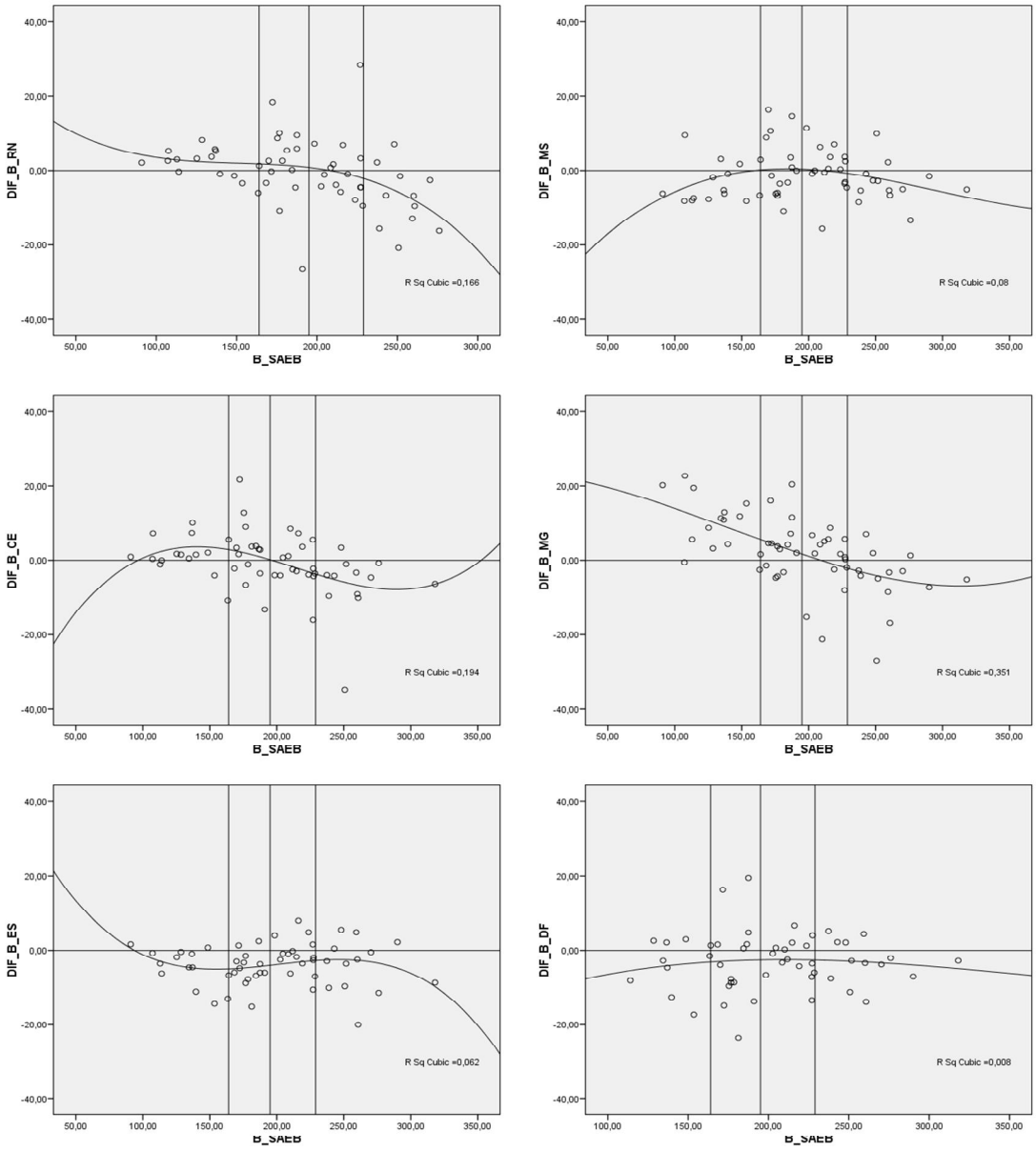

Figura 5.11: Gráficos do tipo *scatter* entre os valores dos parâmetros b do SAEB e suas respectivas diferenças entre a UF e o SAEB.

Etapa 2: Matemática 4ª série

Análise1_etapa2: Variações dos parâmetros b dos itens em cada UF, após eliminação de itens com valores discrepantes de b maior que 40 pontos

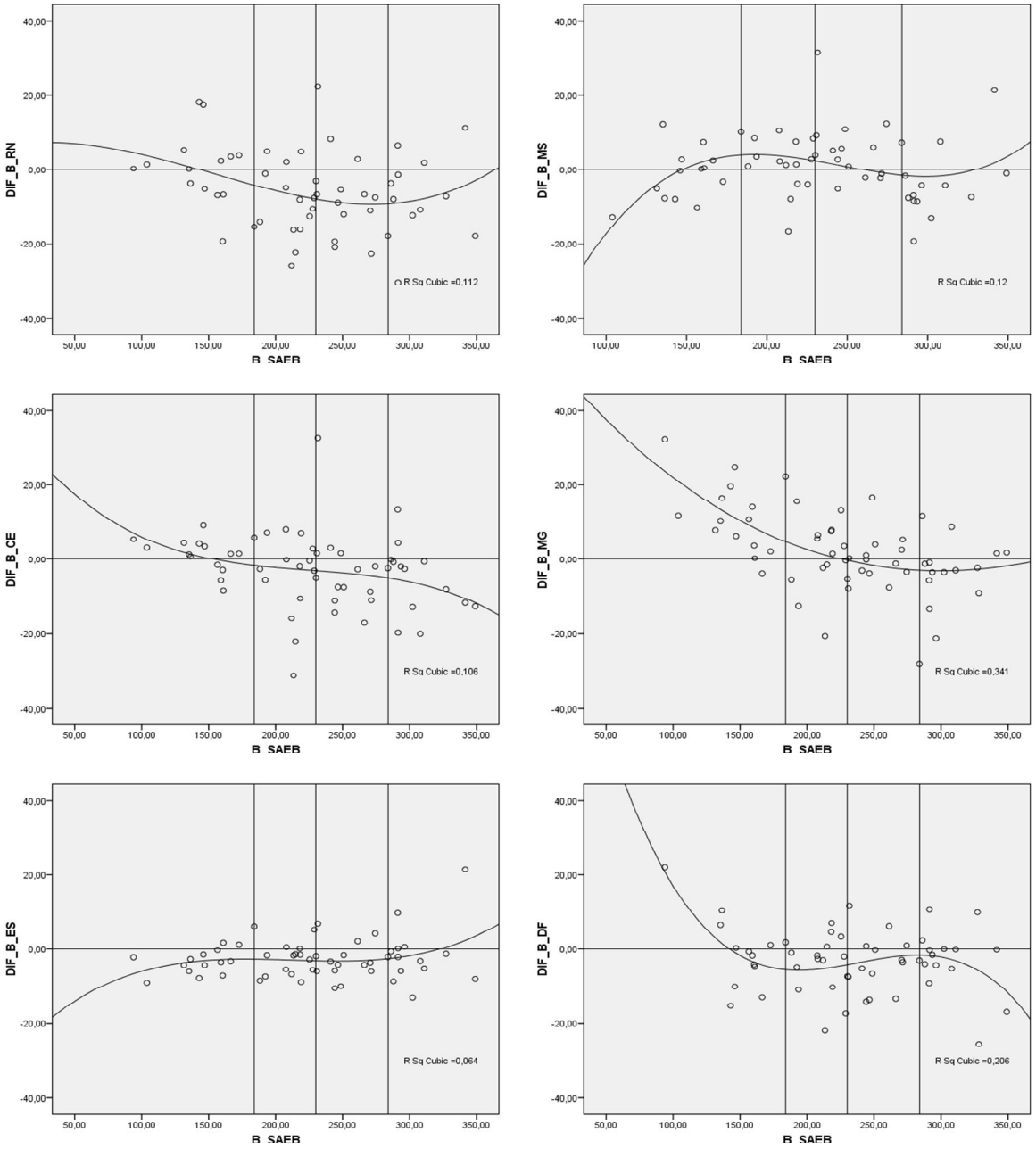

Figura 5.12: Gráficos do tipo *scatter* entre os valores dos parâmetros b do SAEB e suas respectivas diferenças entre a UF e o SAEB.

Etapa 3: Língua Portuguesa 8ª série

Análise1_etapa3: Variações dos parâmetros b dos itens em cada UF, após eliminação de itens com valores discrepantes de b maior que 40 pontos

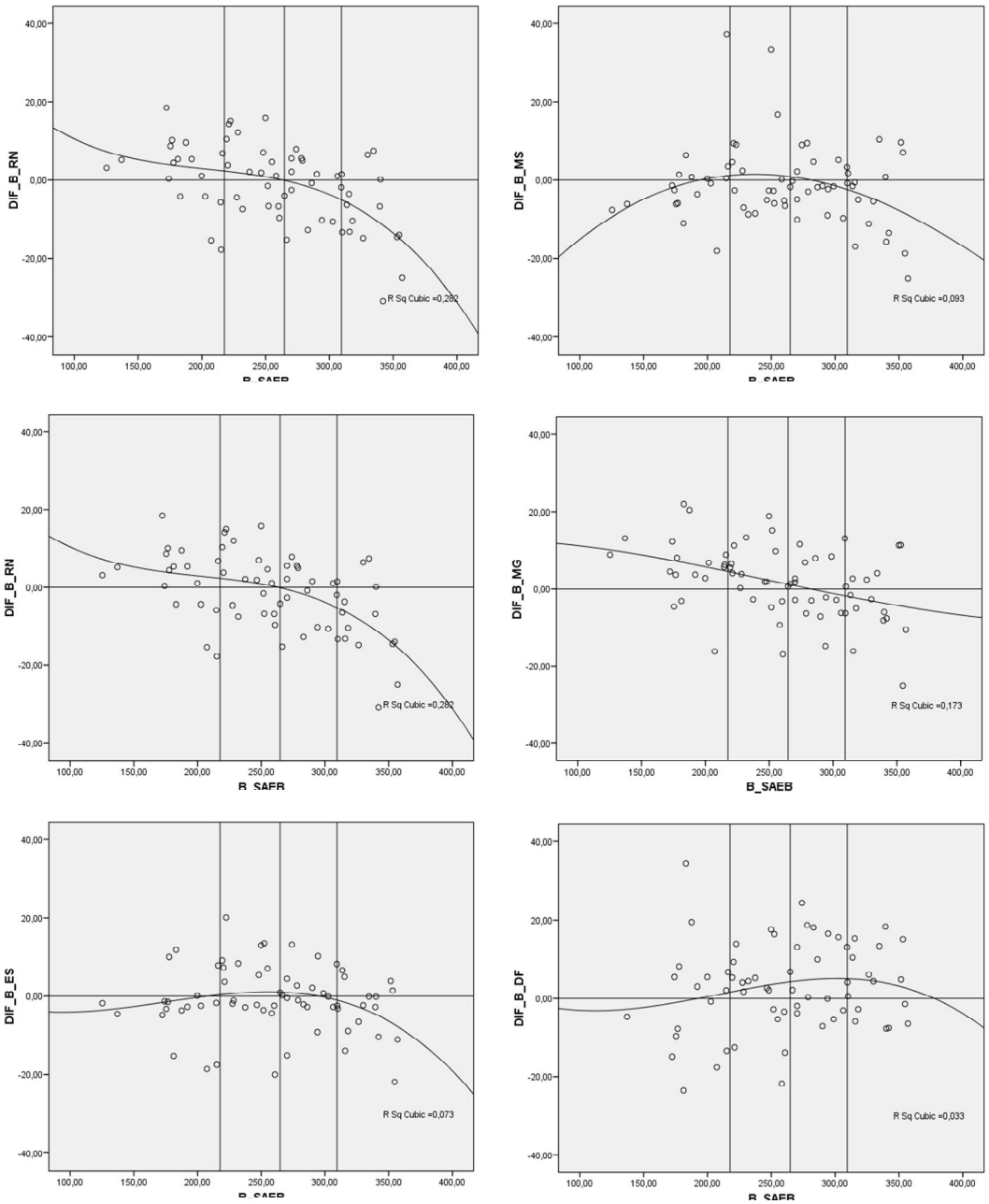

Figura 5.13: Gráficos do tipo *scatter* entre os valores dos parâmetros b do SAEB e suas respectivas diferenças entre a UF e o SAEB.

Etapa 4: Matemática 8ª série

Análise1_etapa3: Variações dos parâmetros b dos itens em cada UF, após eliminação de itens com valores discrepantes de b maior que 40 pontos

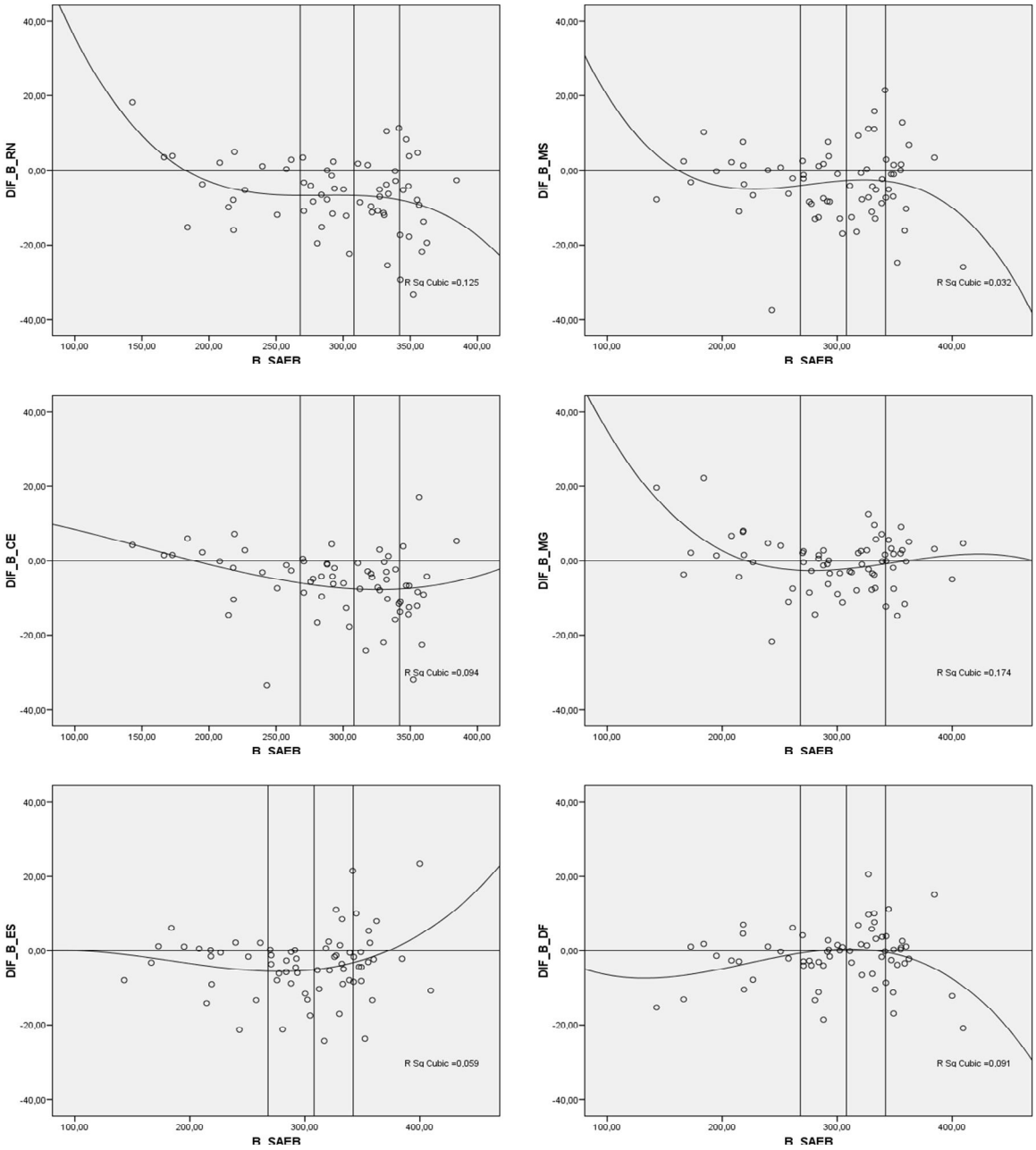

Figura 5.14: Gráficos do tipo *scatter* entre os valores dos parâmetros b do SAEB e suas respectivas diferenças entre a UF e o SAEB.

5.2.2.3 Conclusões

Analisando a concentração de pontos e o formato das curvas cúbicas nos quartis dos gráficos apresentados, podemos concluir que:

1- A calibração de itens em UFs com proficiências abaixo da média nacional tende a superestimar os itens fáceis e subestimar os itens difíceis. Situação inversa ocorre na calibração de itens em UFs com proficiências acima da média nacional, ou seja, os itens fáceis são subestimados e os difíceis são superestimados.

A primeira parte dessa constatação é facilmente verificada ao analisarmos os gráficos referentes a RN e CE como representantes de UFs com proficiências abaixo da média nacional, definidas, nesse estudo como grupo 1. Nesses gráficos, no primeiro quartil que agrupa os itens fáceis, as variações do parâmetro *b* estão mais concentradas na região positiva, indicando a superestimação dos parâmetros b e no último quartil, que agrupa os itens difíceis, as variações do parâmetro *b* estão mais concentradas na região negativa, ou seja, subestimação dos parâmetros b. A curva cúbicas utilizada para modelar as variações do parâmetro *b*, representa bem esta característica, pois começa positiva no 1º quartil, fica próxima de zero, no segundo e terceiro quartil, e, no último quartil, tende a ficar negativa.

Com relação à segunda parte da constatação, utilizaremos as UFs do grupo 3 que para a 4ª série EF é o DF e para a 8ª série, tomaremos MS e DF. MG, embora pertença a este grupo apresentou um comportamento diferenciado e será retratada mais adiante.

Cumpre ressaltar que, na situação anterior, as concentrações no primeiro e último quartil eram bem delimitadas pela linha de referência na posição zero do gráfico, o que não é tão delimitado nessa situação. Entretanto, nos gráficos dessas UFs, ao compararmos o primeiro e o último quartil, verificamos que as variações dos parâmetros b no primeiro quartil estão, geralmente, mais agrupadas abaixo das variações dos parâmetros b no último quartil, o que, conforme salientamos, nem sempre significa que estejam mais concentradas abaixo ou acima da referência zero, respectivamente. No entanto, esse é um indicativo da subestimação de itens fáceis e superestimação de itens difíceis. As curvas cúbicas, utilizadas para modelar as variações do parâmetro *b*, representam bem essa característica, pois, estão mais baixas no primeiro quartil, mais altas no último quartil e tendem a se aproximar de zero nos quartis intermediários.

2 - Na análise anterior, focamos as variações das UFs em relação ao Brasil, porém se fizermos uma comparação entre as UFs, por exemplo, RN e DF as evidências mencionadas ficam ainda mais fortes, pois, no primeiro quartil, as concentrações de itens no DF são sempre inferiores ao RN e no último quartil a concentração de itens no DF são sempre superiores ao RN. Evidenciando que, quanto maiores as diferenças de proficiências entre os grupos, mais evidente são as subestimação e superestimação dos parâmetros dos itens.

3 - Para as UFs do grupo 2, representadas por ES na 4ª série EF e ES e MS na 8ª série, observamos um meio termo com relação às características mencionadas para os grupos 1 e 3. MG, também alocada no grupo 2, apresentou um comportamento diferente e será analisada posteriormente, Para esse grupo, as considerações são similares, embora não tão evidentes, às apresentadas anteriormente. Assim, se compararmos as estimativas dos parâmetros b nestas UFs com as estimativas realizadas nas UFs do grupo 1, teremos uma subestimação dos itens fáceis e uma superestimação dos itens difíceis, e o inverso ocorrerá se compararmos essas UFs com as UFs do grupo 3.

4 - Com relação a MG, que, na 4ª série EF, se encontra no grupo 2, e, na 8ª série EF, no grupo 3, o comportamento dos valores dos parâmetros b, no último quartil, é coerente com o relatado para as outras UFs desses grupos. Entretanto, no primeiro quartil, em todas as situações estudadas, encontramos uma superestimação do parâmetro *b*, sendo esta superestimação maior inclusive que os valores encontrados para o RN. Talvez essa característica esteja relacionada com alto desvio padrão encontrado para a proficiência nessa UF. Estudos adicionais envolvendo os resultados da Prova Brasil 2009 seriam importantes para verificar se esta é realmente uma característica de MG.

5 - Ao analisarmos as variações das diferenças dos valores dos parâmetros b nos gráficos anteriores, verificamos que a quase totalidade das variações estão no intervalo de -20 a 20 pontos, o que poderia acarretar diferenças nos novos valores de proficiências calculados dentro das UFs em relação aos resultados oficiais. No entanto, o fato dessas variações médias nos quartis, assim como em toda a distribuição serem muito baixas, conforme pode ser verificado nos quadros 5.17 e 5.18, tal situação não acontece.

4ª SÉRIE EF											
UF	GRUPO	LÍNGUA PORTUGUESA					MATEMÁTICA				
		QUARTIS				TOTAL	QUARTIS				TOTAL
		1	2	3	4		1	2	3	4	
RN	1	1.9	1.3	0.4	-7.6	-0.6	-0.2	-9.0	-7.2	-6.5	-5.6
CE	1	1.2	2.7	-1.1	-7.0	-0.8	1.5	-4.9	-3.3	-5.6	-2.9
ES	2	-4.6	-5.1	-1.4	-4.4	-3.8	-2.9	-3.3	-3.1	-1.2	-2.6
MS	2	-3.7	1.3	0.5	-3.4	-1.2	-0.8	1.1	5.7	-4.0	0.7
MG	2	10.3	4.4	-0.5	-5.6	2.2	12.0	1.4	-2.0	-2.8	2.2
DF	3	-4.4	-3.0	-2.2	-3.2	-3.1	-0.6	-4.3	-3.9	-3.2	-3.0

Quadro 5.17: Variações médias das diferenças dos parâmetros b por quartil, total e UF na 4ª série EF

8ª SÉRIE EF											
UF	GRUPO	LÍNGUA PORTUGUESA					MATEMÁTICA				
		QUARTIS				TOTAL	QUARTIS				TOTAL
		1	2	3	4		1	2	3	4	
RN	1	1,9	3,1	-1,9	-9,1	-1,3	-2,2	-7,5	-5,3	-11,0	-6,5
CE	1	3,4	2,9	0,8	-8,0	0,0	-3,1	-5,7	-7,0	-8,6	-6,1
ES	2	-2,9	2,9	0,6	-4,2	-0,9	-3,7	-6,6	-2,5	-2,0	-3,6
MS	3	-0,8	1,2	-0,7	-5,3	-1,3	-3,4	-5,0	-1,4	-4,3	-3,5
MG	3	6,0	3,2	0,5	-3,6	1,6	1,8	-3,1	0,2	-1,0	-0,5
DF	3	-0,5	1,3	6,9	3,8	2,9	-2,3	-3,3	2,5	-2,7	-1,4

Quadro 5.18: Variações médias das diferenças dos parâmetros b por quartil, total e UF na 8 série EF

6 - Fica ainda mais evidente, através dos dois quadros anteriores, o fato de MG apresentar as maiores estimações do parâmetro *b* no primeiro quartil, principalmente na 4ª série EF, conforme já relatado no item 4. Esta característica faz com que MG apresente as maiores estimativas de proficiências recalculadas dentro das 6 UFs selecionadas, conforme apresentado no quadros 5.14 e 5.15, principalmente na 4ª série EF.

7– Procuramos analisar duas situações extremas, através de duas simulações. Na primeira, calculamos proficiência do RN, lendo o arquivo de parâmetros dos itens calibrado no DF, enquanto, na segunda, calculamos a proficiência do DF, lendo o arquivo de parâmetros de itens calibrado no RN. Os valores dessas simulações estão apresentados nos quadros 5.19 e 5.20.

RIO GRANDE DO NORTE												
SÉRIE	PROFICIÊNCIA LÍNGUA PORTUGUESA						PROFICIÊNCIA MATEMÁTICA					
	OFICIAL (1)		LENDO PARÂMETROS DO DF (2)		DIF (2)-(1)		OFICIAL (1)		LENDO PARÂMETROS DO DF (2)		DIF (2)-(1)	
	MÉDIA	DP	MÉDIA	DP	MÉDIA	DP	MÉDIA	DP	MÉDIA	DP	MÉDIA	DP
4	151.0	35.3	154.0	32.7	3.0	-2.6	168.9	36.7	170.8	35.1	1.9	-1.6
8	218.3	41.6	220.5	42.1	2.2	0.5	230.2	40.6	230.1	40.5	-0.1	-0.1

Quadro 5.19: Valores de proficiência no RN lendo o arquivo de parâmetros dos itens calibrado no DF.

DISTRITO FEDERAL												
SÉRIE	PROFICIÊNCIA LÍNGUA PORTUGUESA						PROFICIÊNCIA MATEMÁTICA					
	OFICIAL (1)		LENDO PARÂMETROS DO RN (2)		DIF (2)-(1)		OFICIAL (1)		LENDO PARÂMETROS DO RN (2)		DIF (2)-(1)	
	MÉDIA	DP	MÉDIA	DP	MÉDIA	DP	MÉDIA	DP	MÉDIA	DP	MÉDIA	DP
4	191,2	39,5	189,4	39,1	-1,80	-0,40	208,8	40,1	203,9	41,0	-4,90	0,92
8	238,0	44,4	237,1	43,0	-0,90	-1,40	252,2	43,5	248,3	42,5	-3,90	-1,03

Quadro 5.20: Valores de proficiência no DF lendo o arquivo de parâmetros dos itens calibrado no RN.

Por meio dos quadros 5.19 e 5.20, podemos afirmar que as variações nas médias de proficiências são inferiores a 5 pontos, o que indica uma variação muito pequena.

Importante observar, nos casos apresentados, a relação existente entre a geração da proficiência e a origem do arquivo de parâmetros. Constatamos um aumento na proficiência no RN,ao ler os parâmetros dos itens calibrados no DF, e uma diminuição na proficiência do DF, ao ler os parâmetro dos itens calibrados no RN, ou seja, podemos generalizar que, se a proficiência de uma população X for gerada com itens já calibrados em uma população Y, por exemplo através da utilização de um banco de itens, o valor real da proficiência da população X poderá ser subestimado ou superestimado, se a população Y possuir proficiência menor ou maior, respectivamente que a população X.

6 CONSIDERAÇÕES FINAIS

A utilização do termo *linkagem* e *linkar*, adequação lingüística da palavra original na língua inglesa *linkage*, norteia todo esse trabalho, embora seja uma palavra não existente na Língua Portuguesa ela é facilmente entendida como ligação o que é extremante apropriada no contexto da equiparação de escores. Nossa intenção em usar uma palavra nova foi de certa forma enfatizar a necessidade de se utilizar corretamente os termos técnicos inerentes à TRI. Foi nesse sentido que iniciamos esse trabalho fazendo uma revisão de conceitos e procedimentos da TRI alinhados com os existentes na literatura internacional.

Ao apresentarmos as características da TRI e os fatores que interferem na qualidade de seus resultados, esperamos conscientizar pesquisadores, gestores de políticas públicas e demais profissionais envolvidos em projetos de avaliação educacional o quão complexo é gerar resultados comparáveis entre avaliações. O simples fato de estarmos utilizando o mesmo modelo logístico, os mesmos métodos de estimação e os mesmos procedimentos de *linkagem* entre diferentes avaliações, por si só, não garante que teremos resultados devidamente comparáveis.

Diante da questão levantada nessa dissertação sobre se realmente estamos comparando os resultados de nossas avaliações de forma confiável através de uma mesma escala, temos que novamente enfatizar que as análises e simulações realizadas, não englobam todas as nuances possíveis que envolvem essa questão.

Ao concentrarmos nossos estudos nas variações dos designs de testes e nos efeitos de diferentes populações nas calibrações de itens e geração de proficiências, não contemplamos vários outros fatores como, por exemplo, os efeitos relativos às características dos testes no que se referem à qualidade dos itens, distribuição dos descritores por níveis de dificuldades, quantidade e qualidade de itens comuns entre as avaliações. Também não analisamos os efeitos relativos ao modo de como os testes são aplicados como, por exemplo, se é o professor da escola ou um aplicador externo, se os alunos têm um tempo fixo para responder ao teste como um todo ou se o tempo de resposta é por blocos de itens. Esses são apenas alguns fatores não contemplados nas nossas análises e temos convicção de que, ao analisarmos determinado projeto de avaliação, encontraremos outras particularidades que poderão trazer como consequência problemas nas comparabilidades dos resultados.

Portanto, nossa resposta à questão sobre a eficácia da comparabilidade nas avaliações educacionais brasileiras, deverá ser considerada com certa cautela, pois embora considere alguns dos principais fatores que podem afetar a confiabilidade das comparabilidades, esses não são os únicos e devemos também ter em mente que alguns fatores irrelevantes em certas avaliações podem ser importantes em outras realidades.

Diante dessas considerações e no contexto dos estudos de caso utilizados nesse trabalho podemos concluir que é extremamente relevante que avaliações cujo objetivo seja estar na mesma escala nacional sigam o mesmo design atualmente adotado pelo SAEB, pois, conforme observamos em nossos estudos no capítulo 5, mudanças de designs estão geralmente relacionadas ao efeito cansaço nos alunos e suas consequências inviabilizam a comparabilidade de resultados, mesmo mantendo os demais fatores que influenciam as *linkagens* (conforme apresentado no capítulo 5), rigorosamente alinhados com o SAEB.

Destacamos, ainda que a calibração dos itens de uma avaliação estadual dentro do próprio estado, garante praticamente os mesmos valores de proficiência que as obtidas pelo SAEB, onde os itens foram calibrados em uma outra população com a característica de ser amostral e representativa das dependências administrativas[8] nacionais. Portanto, esse procedimento, prática comum nas avaliações estaduais e municipais brasileiras, garante a comparabilidade com a escala SAEB, desde que, voltamos a reiterar, os demais fatores que afetam a *linkagem* sejam devidamente considerados.

Constatamos, através das simulações apresentadas no capítulo 5, que, nos processos de calibração dentro dos estados, alguns itens apresentaram bisseriais negativas e valores discrepantes do parâmetro *b*. Normalmente esses itens estão relacionados a itens fáceis e difíceis em populações com proficiências altas e baixas, respectivamente. A eliminação desses, dentro dos estados e séries em que ocorreram os problemas, melhorou as estimativas de proficiência em relação ao SAEB.

Essa constatação merece uma atenção especial nos projetos de avaliação estadual que têm como característica a produção de resultados na escala nacional, pois, o INEP, ao disponibilizar uma certa quantidade de itens de uma determinada versão do SAEB, (até o presente momento trabalha-se com itens do SAEB 2007) para serem comuns nos procedimentos de *linkagem*, via de regra, fornece os mesmos itens para todos os estados que queiram realizar suas próprias

[8] Escolas públicas e privadas

avaliações. É comum, nessas análises, ter que eliminar itens fornecidos pelo SAEB, com problemas de bisseriais negativas, diminuindo assim, a quantidade de itens comuns entre as avaliações e, consequentemente, uma perda de qualidade nos processos de equalização vertical. Um procedimento simples, como sugestão ao INEP, para contornar esse problema, seria rodar o BILOG-MG para o estado que necessita de itens comuns com o SAEB e selecionar apenas os itens com bisseriais boas e sem valores discrepantes do parâmetro *b*, conforme metodologia empregada nessa dissertação.

O grande número de avaliações estaduais e municipais realizadas na concepção de estarem na mesma escala do SAEB, remete-nos à possibilidade de elaboração de um banco de itens, composto por itens calibrados nessas avaliações. No entanto, conforme constatamos no capítulo 5, os valores dos parâmetros dos itens são fortemente influenciados pelas características das populações envolvidas no processo de calibração, principalmente os itens fáceis e os difíceis. Assim, para a estruturação de um banco de itens, torna-se relevante considerar a inclusão das características das populações, como a média e o desvio padrão da proficiência do estado em que o item foi calibrado. Isso possibilitaria a seleção de itens mais alinhados com a realidade da avaliação a ser implementada.

Entendemos ser importante a realização de estudos futuros envolvendo as demais UFs avaliadas em 2007 pela Prova Brasil, assim como conduzir esses mesmos estudos com as bases da Prova Brasil 2009, pois, quanto mais informações tivermos mais aptos estaremos para melhorar a qualidade dos processos avaliativos.

Destacamos, ainda, que a avaliação educacional no Brasil, oferece uma grande variedade de dados e situações para estudos. Nossa realidade nesse campo é ímpar e a definição dos melhores procedimentos visando à garantia da qualidade e confiabilidade dos resultados terão que ser estabelecidos através nossos próprios esforços e adequações às novas necessidades que se apresentam. Esse é um processo dinâmico e nossa contribuição hoje poderá estar desatualizada em um futuro próximo. Assim, esperamos que este trabalho possa contribuir para novas pesquisas de forma a caminharmos nesse processo evolutivo, visando a manutenção de uma escala de proficiência única para todas as avaliações realizadas no Brasil.

REFERÊNCIAS

BAKER, F. B. **The Basics of Item Response Theory.** 2ª ed. EUA: 2001. 172 p.

BIRNBAUM, A. Some Latent Traits Models and Their Use in Inferring na Examinee's Ability. In: LORD, F. M. & NOVICK, M. R. **Statistical Theories of Mental Test Scores**. Reading, USA: Addison-Wesley, 1968. p. 397-472.

BOCK, R. D. (1997) **A brief history of item response theory**. Educational Messurement: Issues and Practise, 16(4), 21 – 32.

CRAMER, H. **Mathematical Methods of Statistics**. Princeton, USA: Princeton Univ. Press, 1946.

FEUER, M. J. et al. **Uncommon measures: equivalence and linkage among educational tests.** Washington: National Academy of Sciences, 1999. 119 p.

FRANCO, C. e ALVES, MTG. " A pesquisa em Eficácia Escolar no Brasil: Evidências sobre o efeito das escolas e fatores associados à eficácia escolar". In: BROOKE, N. Comentários. In: BROOKE, N.; SOARES, J. F. (Orgs). **Pesquisa em eficácia escolar: origem e trajetória**. Belo Horizonte: Editora UFMG, 2008. p. 482-500.

HAMBLETON, R. K.; SWANINATHAN, H. & ROGERS, H. **Fundamentals of Item Response Theory**. Newbury Park, USA: Sage Publications Inc., 1991.

KLEIN, R. Utilização da teoria da resposta ao item no sistema nacional de avaliação da educação básica (SAEB). **Revista Ensaio**, n.40, v.11, p.283-296, jul./set. 2003.

KOLEN, M. J.; BRENNAN, R. L. **Test Equating, Scaling, and Linking: Methods and Practices.** 2ª ed. New York: Springer, 2004. 548 p.

LINN, R .L; L. SHEOARD, and E. HARTKA The relative standing of states in the 1990 trial state assessment: The influence of choice of content, statistics, and subpopulation breakdowns in Studies for the Evaluateion of the National Assessment of Educational Progress Trial State Assessment. Stantford, CA: National Academy of Education, 1992.

LORD, F.M. (1980). **Applications of item response theory to practical testing problems**. Hillsdale: Lawrence Erlbaum, New York.

PASQUALI, L. **Teoria da Resposta ao Item – TRI – Manual para Iniciantes**. Brasília: Editora X, Laboratório de Pesquisa em Avaliação e Medida – LabPAM, 2004.230p.

SECRETARIA DO ESTADO DA EDUCAÇÃO DE MINAS GERAIS. **PROEB 2001** - Boletim Pedagógico. Ciências Humanas. Competências e habilidades investigadas pelo SIMAVE para a 4ª e 8ª séries do Ensino Fundamental e 3ª série do Ensino Médio. Juiz de Fora: UFJF/CAEd, 2002.

SOARES, T. M. Influência do Professor e do Ambiente em Sala de Aula Sobre a Proficiência Alcançada Pelos Alunos Avaliados no SIMAVE-2002. **Estudos em Avaliação Educacional:** Fundação Carlos Chagas, [São Paulo], v. 28, 2003.

THISSEN D.; STEINBERG L. & WAINER H. Detection of Differential Item Functioning Using the Parameters of Item Response Models. In: HOLLAND, P. W. & WAINER, H. (Eds.). **Differential Item Functioning**. Hillsdale, USA: Lawrence Erlbaum, 1993.

THISSEN, D. & WAINER H. **Test Scoring**. Mahwah, USA: Lawrence Erlbaum Associates Pub, 2001.

TOIT, M. IRT from SSI: **BILOG-MG, MULTILOG, PARSCALE, TESTFACT**. Copyright 2003 by Scientific Software Internacional, Inc.

VALLE, RAQUEL. **Teoria da Resposta ao Item**. SP: USP 1999.

WAINER, H. Model-Based Standardized Measurement of an Item's Differential Impact. In: HOLLAND, P. W. & WAINER, H. (Eds.). **Differential Item Functioning**. Hillsdale, USA: Lawrence Erlbaum, 1993.

WONG, K. & RUTLEDGE, S. (org.) **System-wide Efforts to Improve Student Achievement**. Charlotte, USA: Information Age Publishing Inc., 2006.

ZIMOWSKI, M.F., MURAKI, E., MISLEVY, R.D. (1996). **BILOG-MG: Multiple-group IRT Anaysis and Test Maintenance for Binary Itens.** Scientific Software International, Inc.

ANEXO 1

Métodos não - lineares de linkagem utilizados no BILOG-MG

Apresentaremos os procedimentos para linkagem tal qual implementado no BILOG-MG, envolvendo dois métodos: i) método de calibração simultânea e ii) método de pré-fixação de parâmetros.

1 - Método da Calibração Simultânea.

Conforme relatado anteriormente, esse método é utilizado quando queremos construir uma nova escala envolvendo mais de um grupo de respondentes. Vamos, então, trabalhar na seguinte situação de *Linkagem*:

Duas avaliações em Língua Portuguesa foram realizadas nos anos de 2008 e 2010 em um mesmo estado envolvendo alunos da 4ª série EF. Desejamos construir uma escala de habilidades única, entre essas avaliações. Em cada uma das avaliações, para as suas respectivas populações, foram aplicados 2 cadernos de testes compostos de 16 itens cada. Os itens foram do tipo dicotômico com 4 opções de respostas. Dentro de uma mesma população, os cadernos foram diferentes e possuiam itens comuns entre si. Da mesma forma, existiam itens comuns entre as populações. Os dois modelos de cadernos dentro das diferentes populações foram elaborados de forma a contemplar uma mesma matriz de habilidades. Essa é uma característica importante a ser estabelecida em avaliações que envolvem *linkagem* do tipo equalização vertical, caso contrário, a qualidade das comparabilidades de resultados entre as populações ficará comprometida.

Vamos agora transformar essa situação em bases de dados e sintaxes de forma a podermos utilizar o BILOG-MG em nossa tarefa de construir uma escala de habilidades única para as duas populações. Esse procedimento, envolve 4 etapas: i) construção da base de dados; ii) construção da sintaxe; iii) processamento ou rodada e iv) construção da escala de habilidades. Apresentamos a seguir, as características de cada uma dessas etapas:

1.1 – Construção da base de dados

A base de dados é um arquivo em formato *txt*, constituído por *m* linhas e *n* colunas, sendo m o número de alunos que participaram da avaliação e n as informações relativas a esses alunos. Assim, para cada aluno, teremos qual a sua população, o seu caderno de teste e suas opções de respostas dadas aos itens do teste. No quadro 4.1, apresentamos como fica a estrutura da base de dados para a nossa situação de trabalho. Nesse arquivo, cada aluno recebe um número sequencial de identificação através da variável sequencial (SEQ); as duas populações estão representadas na variável GRUPO, através dos valores 1 ou 2; os 4 modelos de cadernos de testes são indicados na variável FORMA; e as opções de respostas de cada caderno estão indicadas nas variáveis de P1 a P16.

SEQ	GRUPO	FORMA	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
1	1	1	A	B	B	C	C	C	A	A	D	D	A	A	B	B	C	C
2	1	1	A	B	A	C	C	B	B	A	D	D	B	A	C	B	C	C
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5001	1	2	A	B	B	B	C	B	A	A	D	C	A	B	B	A	B	D
5002	1	2	B	B	A	C	C	B	A	A	B	A	A	C	B	A	D	D
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
10001	2	3	B	B	C	A	A	D	A	B	D	C	A	C	D	A	D	D
10002	2	3	B	B	A	A	B	D	A	C	A	C	A	C	A	B	B	C
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15001	2	4	B	B	C	A	A	C	A	B	A	A	C	C	C	D	D	A
15002	2	4	B	B	A	A	C	C	A	D	A	A	B	B	C	D	D	B
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Quadro 4.1: Base de dados no formato para processamento no BILOG-MG.

Para que o BILOG-MG possa identificar se o aluno acertou ou errou o item, devemos fornecer, nas primeiras linhas do arquivo, representado no quadro 4.1, o gabarito de cada item em sua respectiva forma e grupo. Ao elaborar a sintaxe, informaremos que a chave de correção se encontra dentro do mesmo arquivo com as opções de respostas dos alunos. O BILOG-MG também dá a opção de que a chave de correção seja fornecida em arquivo separado, sendo que essa opção também deverá ser devidamente informada na elaboração da sintaxe.

1.2 – Construção da sintaxe

Após a construção da base de dados, o próximo procedimento é a construção da sintaxe. Descreveremos através dos passos de 1 a 6 as principais características a serem fornecidas para que o BILOG-MG consiga realizar a calibração simultânea:

1 – Seleção de qual o modelo da TRI a ser utilizado: existe a opção para os modelos logísticos de 1, 2 ou 3 parâmetros,

2 - Como ler a base de dados: Se por exemplo utilizarmos o seguinte código:

(10A1, I1, I1, 16A1)

Estaremos informando ao programa que os 10 primeiros campos são referentes à identificação do aluno, o próximo campo de tamanho 1 é referente ao grupo, o campo seguinte é referente à forma e os últimos 16 campos são referentes às opções de respostas. Dessa forma o programa terá como relacionar a resposta de cada aluno ao respectivo gabarito e assim poder dar início ao processamento.

3-Como identificar os itens comuns: Para que o BILOG-MG efetue a *linkagem* entre as diferentes formas e grupos, devemos fornecer um mesmo número de identificação para os itens comuns nas diferentes formas e grupos em que os mesmos estiverem presentes. Dessa maneira o programa "entenderá" que esses itens terão os seus parâmetros como elementos de ligação na construção da escala.

Para esse nosso caso de estudo, apresentamos na figura 1, como estão posicionados os itens comuns em suas respectivas formas e grupos. Assim, temos que os 8 primeiros itens da forma 1 são comuns com os 8 primeiros itens da forma 2 (esta relação está indicada com a cor verde), os 8 primeiros itens da forma 3 são comuns com os 8 primeiros itens da forma 4 (esta relação está indicada com a cor azul) e, os 8 últimos itens da forma 2 são comuns com os 8 últimos itens da forma 4 (esta relação está indicada com a cor amarela). Dessa maneira, garantimos a ligação entre todas as formas e grupos.

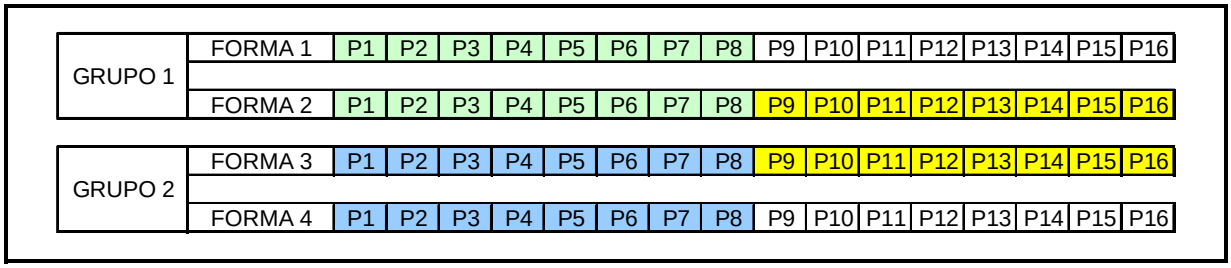

Figura.1: Posição dos itens comuns em suas formas e grupos.

Devemos, nesse momento, informar ao BILOG-MG como identificar os itens comuns. Essa tarefa é obtida criando um número de identificação do item. Através do diagrama da figura 2, a seguir, indicamos como isso é realizado. Note-se que o número de identificação do item é posicionado de forma a respeitar a posição do item no teste e ao mesmo tempo fornecer a ligação entre as diferentes formas e grupos. Assim, os itens comuns entre as diferentes formas recebem o mesmo número de identificação.

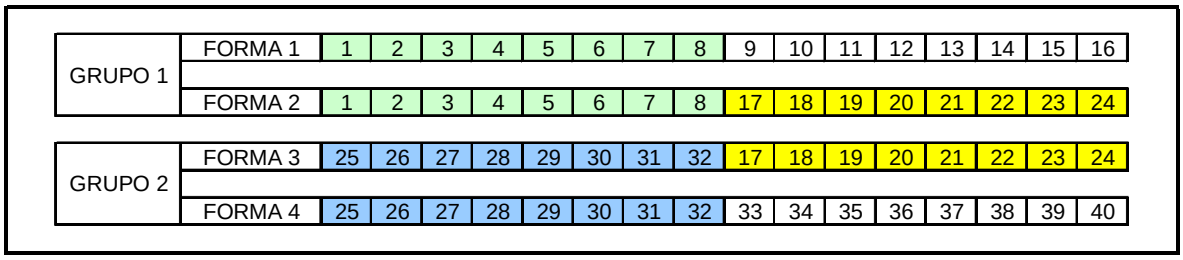

Figura.2: número de identificação dos itens em suas respectivas formas e grupos.

Da maneira como foi apresentado, se tomarmos como exemplo o décimo item da forma 3 (P10), veremos que ele é comum com o décimo item da forma 2 (P10) e ambos possuem o mesmo número de identificação, 18.

4 – Informar qual o método de calibração. Através do comando CALIB, informamos as características necessárias para a utilização do método MML ou MMAP. Caso o método selecionado seja o MML, devemos informar: número de ciclos EM e de Newton, número de pontos de quadratura, critério de convergência, tipo de distribuição a priori e o grupo de referência . Caso a opção seja o método MMAP, além das informações já citadas para o método MML, devemos especificar valores para as prioris dos parâmetros dos itens através do comando PRIORS. Essas informações já foram retratadas no capítulo 3.

5– Informar qual o método de geração de escores e suas características através do comando SCORE .

6 – Salvar os arquivos de parâmetros dos itens e scores de alunos

1.3 – Processamento ou rodada

Uma vez construída a base de dados e a sintaxe, a próxima etapa é rodar o BILOG-MG. Como estamos utilizando o método de calibração simultânea, em uma única rodada, teremos os arquivos de parâmetros dos itens e os de escores dos alunos na mesma métrica. Após as devidas verificações para que seja garantida a qualidade dos resultados obtidos, como por exemplo, se o critério de convergência foi atingido, se os dos dados estão bem ajustados ao modelo e se não há itens com comportamento diferencial. Sugerimos, para mais detalhes sobre critérios para verificação da qualidade das *linkagens*, consultar Pasquali (2004).

1.4 – Construção da escala

Como os escores dos alunos são fornecidos em uma escala variando de -3 a +3, normalmente, para eliminar valores negativos e facilitar a interpretação dos resultados, fazemos um escalonamento dos escores originais, através da aplicação de um determinado fator de escala. Esse fator de escala consiste em multiplicar e somar cada escore original por valores pré-estabelecidos. Normalmente, para a produção do escore final que, em nosso caso, denominamos de proficiência, utiliza-se como fator multiplicativo o valor 50 e o fator de soma o valor 250:

$$\text{ESCORfinal} = \text{ESCORE ORIGINAL} * 50 + 250 \qquad (1)$$

Para que os parâmetros dos itens estejam nessa mesma métrica ,teremos que efetuar as seguintes operações:

$$Afinal = Aorig/50 \quad (2)$$

$$Bfinal = borig*50 + 250 \quad (3)$$

$$Cfinal = corig \quad (4)$$

Através das etapas apresentadas, teremos atingido nosso objetivo de *linkagem*, que foi gerar uma nova escala única para as duas populações. Veremos a seguir uma situação um pouco mais complexa onde o objeto é *linkar* uma nova avaliação a uma escala já estabelecida.

2- Método da Pré-fixação dos Parâmetros dos Itens – FPIP.

Em uma situação em que desejamos *linkar* uma avaliação a uma escala já estabelecida, o método de calibração simultânea, apresentado anteriormente, não funciona. Nesse caso, teremos que empregar o método de Pré-fixação dos Parâmetros do Itens – FPIP.

Para entendermos como esse método funciona, vamos considerar que para o caso apresentado anteriormente, quando foi realizada a primeira avaliação no ano de 2008 foi construída uma escala de habilidades, e que ao se aplicar uma nova avaliação no ano de 2010, o nosso trabalho é *linkar* os resultados dessa última avaliação à escala estabelecida para a avaliação de 2008.

Neste trabalho, utilizaremos o método FPIP, tomando, como exemplo, as mesmas populações e testes do caso anterior. A solução pelo método FPIP segue as seguintes etapas: i) construção da base de dados; ii) construção da sintaxe 1; iii) processamento1 ou rodada 1; iv) construção da sintaxe 2; v) processamento 2 ou rodada 2 e vi) construção da escala habilidades. Apresentamos a seguir, as características de cada uma dessas etapas e suas particularidades e semelhanças com relação ao apresentado para a calibração simultânea.

2.1 – Construção da base de dados

Não se altera com relação ao apresentado no tópico 1.1

2.2 – Construção da sintaxe 1

Na construção da sintaxe 1, utilizamos os mesmos procedimentos apresentados no tópico 1.2, até o passo 3. A partir do passo 4, os métodos possuem características distintas conforme apresentamos a seguir:

4 – Informar qual o método de calibração. No presente caso, em que estamos utilizando o FPIP, a única opção de calibração é pelo método MMAP. Assim, utilizaremos todas as informações já citadas para o método MML e acrescentaremos o comando PRIORS para fixar os parâmetros dos itens já calibrados. Dessa maneira, o novo arquivo de parâmetros a ser gerado, manterá os mesmos parâmetros dos itens já calibrados na população 1 (itens de 1 a 24) e a estimação dos itens exclusivos da população 2 (itens de 25 a 40), serão estimados na mesma escala já estabelecida inicialmente. Os itens comuns às duas populações (itens de 17 a 24) é que serão os responsáveis pela *linkagens* entre as populações.

5– Informar qual o método de geração de escores e suas características através do comando SCORE. Deveremos também informar os fatores de escalas (SCA e LOC) para a transformação dos arquivos dos parâmetros na escala desejada. Os cálculos desses fatores são baseados em transformações lineares a partir do arquivo de parâmetros gerado na fase 2 do BILOG-MG,obtido quando esse arquivo foi gerado na criação da escala original, tendo como referência apenas a população 1. Essas constantes de transformações são obtidas a partir da média e desvio-pardrão, que o grupo adotado como referência nessa nova situação, possuía quando foi gerada a escala original. Como nesse nosso caso, a média e desvio padrão da população 1, quando da geração da escala original, é 0 e 1 respectivamente, os valores dos parâmetros dos parâmetros dos itens a serem fixados, permanecem os mesmos.

6 – Salvar o arquivo de parâmetros dos itens.

2.3 – Processamento 1 ou rodada 1

Uma vez construída a base de dados e a sintaxe 1, a próxima etapa é rodar o BILOG-MG. Como estamos utilizando o método FPIP, não é possível obter os parâmetros dos itens e os escores em uma mesma escala em uma única rodada. Assim, essa tarefa será dividida em duas rodadas. Portanto, nessa primeira rodada, obteremos apenas os parâmetros dos itens na escala já estabelecida. Devemos ressaltar que na fixação dos parâmetros dos itens já calibrados (itens de 1 a 24), devemos utilizar os parâmetros na escala original.

Os mesmos critérios para a garantia da qualidade dos resultados, conforme relatado no tópico 1.3, deverão ser seguidos.

2.4 – Construção da sintaxe 2

Da mesma maneira que na construção da sintaxe 1, na construção da sintaxe 2, utilizamos os mesmos procedimentos apresentados no tópico 4.2.1.2, até o passo 3. A partir do passo 4 os métodos possuem características distintas conforme apresentamos a seguir:

4 – Informar qual o método de calibração. Como os parâmetros dos itens já foram calculados na rodada1, devemos informar que esse comando não será executado (SELECT=0 no comando CALIB), e que os parâmetros deverão ser lidos em arquivo específico, que foi salvo na rodada1. o comando para essa operação é IFNAME=’nome do arquivo de parâmetros.PAR’.

5– Informar qual o método de geração de escores e suas características através do comando SCORE.

6 – Salvar o arquivo de escores.

2.5 – Construção da escala

A construção da escala segue os mesmos procedimentos abordados no tópico1.3.

Em resumo o método FPIP, utilizado para a *linkagem* de uma população a uma escala já existente, exige que o BILOG-MG seja rodado 2 vezes. Em uma primeira rodada são obtidos os parâmetros dos itens na escala desejada e, em uma segunda rodada, a proficiência dos alunos nessa mesma escala original. Esse método também exige que a mesma população utilizada para a calibração inicial os itens seja utilizada para a *linkagem* com outra população. Caso não seja utilizada a mesma população poderão ocorrer erros nos processos de estimação de parâmetros de itens e escores de alunos, colocando em risco a qualidade da comparabilidade de resultados entre as populações envolvidas.

ANEXO 2

Comportamento do parâmetro b quando sub-grupos de uma mesma população é submetida aos mesmos itens, com diferentes designs de testes

Quadros com os valores dos parâmetros *b* dos itens no programa Nova Escola 2006 obtidos em três procedimentos de linkagem de acordo com a base de dados utilizada para calibração: i) base completa cadernos pares e ímpares juntos ii) base separada utilizando apenas os cadernos ímpares e iii) base separada utilizando apenas os cadernos pares.

LÍNGUA PORTUGUESA						1/2
ITEM	SÉRIE	B_JUNTO	B_IMPAR	B_PAR	DIF_B_IMPAR_JUNTO	DIF_B_PAR_JUNTO
448	4	88,1	91,0	89,5	2,9	1,4
449	4	139,2	143,9	141,8	4,7	2,6
450	4	120,7	122,4	123,3	1,8	2,6
451	4	162,4	168,0	157,7	5,6	-4,7
452	4	191,8	192,1	190,8	0,3	-1,0
453	4	214,0	215,4	212,2	1,4	-1,8
454	4	193,1	193,0	193,5	-0,2	0,4
455	4	190,7	184,2	199,5	-6,6	8,7
456	4	76,8	77,6	78,6	0,8	1,8
457	4	167,2	172,7	162,8	5,4	-4,4
458	4	180,9	179,0	181,7	-1,9	0,7
459	4	147,7	144,3	149,9	-3,4	2,2
460	4	189,7	193,1	185,1	3,4	-4,5
461	4	136,5	155,9	128,0	19,4	-8,5
462	4	160,7	167,0	167,1	6,3	6,4
463	4	81,6	84,6	83,4	3,0	1,8
464	4	236,9	238,5	233,8	1,6	-3,0
465	4	237,1	245,8	228,9	8,8	-8,2
466	4	122,7	119,4	127,3	-3,3	4,6
467	4	219,6	222,1	215,7	2,5	-3,8
468	4	195,5	196,3	194,2	0,7	-1,4
469	4	98,9	104,0	99,5	5,0	0,5
470	4	211,2	215,3	205,8	4,2	-5,4
471	4	124,3	124,5	127,5	0,2	3,2
472	4	156,5	156,8	159,6	0,3	3,0
473	4	141,5	147,3	142,9	5,8	1,4
474	8	300,8	302,1	296,4	1,3	-4,4
475	8	180,0	180,0	186,9	0,1	7,0
476	8	184,9	184,9	192,2	0,0	7,3
477	8	230,4	233,5	230,3	3,1	-0,1
478	8	302,1	298,8	303,1	-3,4	1,0
479	8	298,4	295,2	300,7	-3,2	2,4
480	8	264,6	260,4	266,6	-4,2	2,0
481	8	248,2	249,1	253,0	0,9	4,8
482	8	152,5	155,8	154,1	3,3	1,6
483	8	285,0	281,7	288,2	-3,3	3,2
484	8	213,3	216,0	214,4	2,7	1,1
485	8	306,8	306,9	307,5	0,1	0,6
486	8	143,7	146,7	145,9	3,0	2,2
487	8	188,4	186,6	190,3	-1,8	1,9
489	8	155,3	156,7	158,0	1,5	2,8
490	8	204,9	209,4	207,7	4,5	2,8
491	8	222,3	230,3	216,9	8,0	-5,4
492	8	239,5	239,1	242,0	-0,4	2,5
493	8	265,1	261,5	267,2	-3,6	2,1
494	8	258,5	255,6	262,0	-3,0	3,4
495	8	321,1	321,7	318,0	0,5	-3,2
496	8	248,9	245,8	254,7	-3,1	5,8
497	8	326,7	330,5	319,8	3,7	-7,0
498	8	238,7	229,8	244,7	-8,9	6,0
499	8	207,1	207,7	210,5	0,6	3,4
500	8	151,8	150,8	155,0	-0,9	3,2
501	8	261,0	256,2	263,9	-4,8	2,9
502	8	145,2	148,4	146,4	3,2	1,2
503	8	248,0	252,6	242,3	4,7	-5,7
504	8	210,4	207,8	217,6	-2,6	7,2
505	8	227,2	233,7	228,7	6,6	1,5
506	8	191,4	191,0	199,6	-0,4	8,2
507	8	166,1	165,5	170,7	-0,7	4,6
508	8	249,5	250,9	247,9	1,4	-1,6
509	8	278,3	279,9	275,2	1,5	-3,1
510	8	155,5	158,5	156,4	3,1	1,0
511	8	201,6	200,5	205,0	-1,1	3,4
512	8	204,2	217,7	203,8	13,5	-0,4
513	8	165,1	169,2	163,6	4,1	-1,5
514	11	220,3	214,6	226,7	-5,7	6,4

LÍNGUA PORTUGUESA 2/2						
ITEM	SÉRIE	B_JUNTO	B_IMPAR	B_PAR	DIF_B_IMPAR_JUNTO	DIF_B_PAR_JUNTO
515	11	247,8	252,0	246,4	4,2	-1,4
516	11	200,5	201,0	204,9	0,5	4,4
517	11	171,9	170,2	175,4	-1,6	3,5
518	11	177,1	177,7	179,7	0,6	2,6
519	11	184,3	186,0	185,4	1,7	1,1
520	11	204,0	197,3	209,0	-6,7	5,1
521	11	186,8	177,3	194,0	-9,5	7,2
522	11	183,6	175,3	189,9	-8,3	6,3
523	11	232,7	234,1	234,9	1,4	2,2
524	11	247,8	253,0	246,0	5,2	-1,8
525	11	185,4	182,2	190,9	-3,2	5,5
526	11	189,1	193,1	190,8	4,0	1,7
527	11	227,0	228,2	228,5	1,1	1,5
528	11	199,8	199,1	203,7	-0,8	3,8
529	11	316,1	314,6	316,4	-1,6	0,3
530	11	316,9	320,3	315,8	3,3	-1,1
531	11	257,4	257,9	262,3	0,5	4,9
532	11	228,6	231,4	227,6	2,8	-1,0
533	11	290,7	294,9	287,6	4,3	-3,0
534	11	222,4	225,2	223,8	2,7	1,3
535	11	242,6	241,6	244,0	-1,0	1,3
536	11	251,8	255,6	250,3	3,8	-1,5
537	11	231,3	229,9	234,0	-1,4	2,7
538	11	261,5	261,0	261,8	-0,5	0,3
539	11	263,3	262,2	263,5	-1,1	0,2
540	11	242,1	243,1	241,9	1,0	-0,2
541	11	249,4	250,4	249,5	1,0	0,0
542	11	244,5	243,5	246,7	-0,9	2,2
543	11	249,6	248,0	250,3	-1,6	0,7
544	11	253,6	254,3	253,6	0,7	0,0
545	11	199,7	195,6	204,2	-4,1	4,6
546	11	292,0	290,9	294,2	-1,1	2,2
547	11	267,3	263,4	274,4	-3,9	7,1
548	11	260,2	251,8	268,4	-8,4	8,3
549	11	277,4	281,4	280,0	4,0	2,7
550	11	200,2	197,8	206,6	-2,4	6,3
551	11	274,3	272,4	277,8	-1,9	3,5
552	11	239,2	236,3	243,2	-2,9	4,0
553	11	221,1	220,7	223,8	-0,4	2,7

MATEMÁTICA						1/2
ITEM	SÉRIE	B_JUNTO	B_PAR	B_IMPAR	DIF_B_IMPAR_JUNTO	DIF_B_PAR_JUNTO
440	4	182,0	181,9	181,7	-0,3	-0,1
441	4	164,2	160,3	166,1	1,9	-3,9
442	4	170,7	164,8	176,1	5,4	-5,9
443	4	253,7	253,0	253,7	0,0	-0,8
444	4	208,6	207,8	209,3	0,7	-0,8
445	4	92,9	95,0	94,7	1,8	2,1
446	4	186,5	181,3	190,8	4,3	-5,2
447	4	125,0	123,1	129,8	4,8	-1,9
448	4	133,6	134,2	136,5	2,9	0,6
449	4	218,4	219,4	217,4	-1,0	1,0
450	4	122,5	123,5	126,9	4,4	1,0
451	4	150,3	150,2	155,1	4,8	-0,1
452	4	222,3	220,6	224,7	2,3	-1,7
453	4	93,6	86,0	102,7	9,1	-7,6
454	4	209,5	207,6	211,9	2,4	-2,0
455	4	194,3	194,1	195,4	1,2	-0,2
456	4	195,7	197,9	193,7	-2,0	2,2
457	4	184,4	187,0	181,6	-2,9	2,6
458	4	176,6	175,8	178,4	1,8	-0,8
459	4	115,3	120,9	115,9	0,6	5,6
460	4	144,7	149,1	142,9	-1,8	4,4
461	4	119,0	121,7	120,5	1,6	2,8
462	4	189,3	185,6	193,6	4,3	-3,8
463	4	197,5	194,1	202,2	4,8	-3,4
464	4	156,7	160,9	154,7	-2,0	4,3
465	4	249,3	249,2	249,1	-0,2	-0,1
466	8	234,3	232,3	239,4	5,1	-2,1
467	8	201,9	204,1	204,4	2,5	2,2
468	8	232,1	244,8	227,1	-5,0	12,7
469	8	232,3	230,2	237,8	5,5	-2,1
470	8	235,1	230,4	240,6	5,6	-4,7
471	8	236,1	239,0	233,6	-2,6	2,9
472	8	255,3	255,3	254,8	-0,4	0,0
473	8	246,8	242,6	251,1	4,2	-4,2
474	8	263,2	263,2	264,6	1,4	0,0
475	8	185,8	184,0	191,7	5,9	-1,8
476	8	312,3	313,3	312,6	0,3	1,0
477	8	285,4	283,2	287,7	2,3	-2,3
478	8	254,9	256,6	252,8	-2,1	1,7
479	8	260,5	262,7	258,3	-2,3	2,1
480	8	240,4	239,5	241,4	1,0	-0,9
481	8	205,4	203,3	208,9	3,5	-2,1
482	8	164,1	162,1	169,0	4,8	-2,1
483	8	178,8	177,9	183,4	4,6	-0,9
484	8	267,9	266,7	269,6	1,7	-1,2
485	8	245,5	246,7	243,6	-1,9	1,2
486	8	170,1	167,2	176,7	6,6	-2,9
487	8	225,4	225,5	230,4	5,0	0,0
488	8	278,9	279,6	279,4	0,5	0,8
489	8	283,9	282,2	285,9	1,9	-1,8
490	8	226,7	224,3	222,0	-4,7	-2,4
491	8	291,8	292,2	291,2	-0,6	0,4
492	8	314,9	315,8	313,6	-1,2	0,9
493	8	274,8	273,4	276,3	1,4	-1,5
494	8	315,0	316,0	313,4	-1,6	1,0
495	8	280,2	279,6	280,3	0,1	-0,6
496	8	294,5	295,7	292,6	-1,9	1,2
497	8	228,5	229,2	227,9	-0,6	0,7
498	8	195,5	195,3	202,7	7,2	-0,2
499	8	274,2	269,7	278,1	3,9	-4,5
500	8	169,5	169,7	175,0	5,5	0,2
501	8	233,3	229,8	237,5	4,2	-3,4
502	8	233,7	233,6	232,9	-0,8	-0,1
503	8	168,2	173,8	169,0	0,7	5,6
504	8	210,3	208,1	218,3	8,0	-2,2
505	11	278,5	275,4	282,6	4,0	-3,2

MATEMÁTICA					2/2	
ITEM	SÉRIE	B_JUNTO	B_PAR	B_IMPAR	DIF_B_IMPAR_JUNTO	DIF_B_PAR_JUNTO
506	11	329,2	328,7	329,1	-0,1	-0,5
507	11	338,9	336,7	340,9	2,0	-2,2
508	11	305,9	308,7	301,2	-4,7	2,7
509	11	385,9	382,9	389,0	3,1	-3,0
510	11	313,8	313,6	314,0	0,1	-0,3
511	11	324,7	323,4	326,1	1,3	-1,3
512	11	326,8	324,5	329,1	2,3	-2,3
513	11	306,1	305,1	307,5	1,3	-1,0
514	11	353,4	351,9	354,4	1,1	-1,5
515	11	341,7	342,4	341,0	-0,7	0,7
516	11	274,4	271,5	286,3	11,9	-2,9
517	11	348,4	347,7	348,9	0,5	-0,8
518	11	311,8	310,5	312,3	0,5	-1,4
519	11	332,9	330,8	334,7	1,8	-2,1
520	11	309,9	306,1	314,1	4,2	-3,8
521	11	307,7	306,2	309,7	2,0	-1,5
522	11	327,1	325,3	329,3	2,1	-1,9
523	11	319,5	320,2	318,8	-0,8	0,7
524	11	333,3	334,0	331,6	-1,7	0,7
525	11	338,8	336,7	341,3	2,4	-2,1
526	11	342,1	341,5	342,5	0,3	-0,6
527	11	292,2	297,5	301,4	9,1	5,3
528	11	257,8	256,9	262,8	5,0	-1,0
529	11	365,1	366,4	363,1	-2,1	1,3
530	11	298,8	295,4	304,3	5,4	-3,4
531	11	330,0	327,4	332,6	2,6	-2,6
532	11	309,0	304,5	315,1	6,1	-4,5
533	11	314,9	310,3	322,6	7,7	-4,6
534	11	312,5	304,7	322,5	10,0	-7,7
535	11	269,9	265,5	276,6	6,7	-4,4
536	11	307,4	305,5	309,4	2,0	-1,8
537	11	333,1	333,4	332,2	-0,9	0,3
538	11	292,7	289,3	296,8	4,0	-3,4
539	11	258,9	254,7	268,7	9,8	-4,1
541	11	289,5	287,6	292,6	3,2	-1,9
542	11	318,6	319,6	317,4	-1,2	1,0
543	11	245,7	241,0	257,3	11,7	-4,6
544	11	301,9	301,1	303,1	1,2	-0,8

ANEXO 3

Comportamento do parâmetro b quando populações diferentes são submetidas ao mesmo design de teste e mesmos itens

Quadros com os valores dos parâmetros *b* dos itens nas UFs do Rio Grande do Norte, Ceará, Espírito Santo, Mato Grosso do Sul, Minas Gerais e Distrito Federal, para Língua Portuguesa e para Matemática, na Prova Brasil 2007, obtidos por dois procedimentos de *linkagens*: Via amostra nacional e via calibração dentro da própria UF.

LÍNGUA PORTUGUESA - 4ª SÉRIE **VARIAÇÕES DOS PARÂMETROS b DOS ITENS EM CADA UF**

ITEM	PARÂMETRO b							DIF_b_RN	DIF_b_CE	DIF_b_ES	DIF_b_MS	DIF_b_MG	DIF_b_DF
	SAEB	RN	CE	ES	MS	MG	DF						
429	219,1	218,1	222,7	216,3	226,1	216,7	220,6	-0,9	3,7	-2,7	7,1	-2,3	1,6
430	137,0	142,0	147,0	133,6	130,8	149,7	139,0	5,0	10,0	-3,4	-6,2	12,7	1,9
431	238,8	223,2	229,4	229,3	233,5	234,9	236,4	-15,6	-9,5	-9,5	-5,4	-3,9	-2,5
432	148,6	147,1	150,4	150,4	150,2	160,2	156,8	-1,4	1,9	1,9	1,7	11,7	8,3
433	198,4	205,7	194,4	203,2	209,9	183,0	196,8	7,3	-4,0	4,8	11,5	-15,4	-1,6
434	125,3	128,4	126,7	124,6	117,6	133,7	203,2	3,1	1,4	-0,7	-7,7	8,4	77,9
435	178,5	181,0	177,3	171,6	174,9	181,3	175,2	2,6	-1,2	-6,9	-3,6	2,8	-3,3
436	228,5	219,1	225,1	222,1	223,9	226,7	228,1	-9,4	-3,4	-6,4	-4,6	-1,8	-0,4
437	176,7	186,9	185,6	176,2	170,7	180,2	175,5	10,2	9,0	-0,5	-6,0	3,5	-1,2
438	208,5	209,3	209,7	208,2	214,8	212,8	210,1	0,8	1,2	-0,3	6,3	4,3	1,6
439	134,4	138,0	134,6	131,0	137,4	145,4	137,8	3,6	0,2	-3,4	3,0	11,0	3,4
440	90,9	93,0	91,4	93,9	84,8	110,4	214,7	2,0	0,4	2,9	-6,2	19,5	123,8
441	168,3	165,0	166,0	163,3	177,3	166,7	172,4	-3,3	-2,3	-5,1	8,9	-1,6	4,1
442	107,2	109,7	107,1	107,8	99,0	105,9	180,5	2,5	-0,1	0,6	-8,2	-1,3	73,3
443	237,4	239,6	233,5	234,8	228,9	234,8	245,7	2,2	-3,9	-2,6	-8,5	-2,5	8,4
444	153,5	150,1	149,3	140,2	145,3	168,7	140,7	-3,4	-4,3	-13,3	-8,2	15,2	-12,9
445	204,6	203,5	205,3	204,5	204,5	206,4	209,7	-1,1	0,7	-0,1	-0,1	1,8	5,1
446	318,1	x	312,0	309,1	313,0	313,6	316,1	x	-6,1	-9,0	-5,1	-4,5	-2,1
447	169,9	172,5	173,2	167,5	186,3	174,3	169,6	2,6	3,2	-2,4	16,4	4,3	-0,3
448	191,0	164,4	177,7	185,4	190,9	193,0	179,6	-26,5	-13,3	-5,6	-0,1	2,0	-11,3
449	229,3	x	x	x	x	x	x	x	x	x	x	x	x
450	290,1	x	x	292,2	288,5	283,4	284,5	x	x	2,2	-1,5	-6,7	-5,6
451	242,9	236,3	238,9	244,1	242,0	250,3	249,3	-6,6	-4,0	1,2	-0,9	7,4	6,5
452	248,1	255,2	251,4	253,8	245,3	250,2	252,7	7,1	3,4	5,7	-2,7	2,2	4,6
453	107,6	112,8	114,4	188,1	117,1	129,7	208,4	5,2	6,8	80,6	9,5	22,1	100,8
454	250,7	230,0	215,9	241,4	260,7	223,9	242,5	-20,7	-34,9	-9,3	10,0	-26,9	-8,2
455	259,3	246,4	256,2	265,5	261,5	251,2	268,7	-12,9	-3,1	6,2	2,1	-8,1	9,4
456	139,7	138,8	140,9	130,1	138,7	143,7	132,7	-0,9	1,3	-9,5	-1,0	4,1	-7,0
457	136,6	142,5	143,7	137,3	131,3	147,4	144,8	6,0	7,1	0,8	-5,2	10,8	8,2
458	216,1	223,1	223,4	224,5	219,7	225,1	227,1	7,0	7,3	8,3	3,5	9,0	10,9

LÍNGUA PORTUGUESA - 4ª SÉRIE **VARIAÇÕES DOS PARÂMETROS b DOS ITENS EM CADA UF** *continuação*

ITEM	PARÂMETRO b							DIF_b_RN	DIF_b_CE	DIF_b_ES	DIF_b_MS	DIF_b_MG	DIF_b_DF
	SAEB	RN	CE	ES	MS	MG	DF						
459	223,7	215,9	219,9	230,0	224,0	225,5	230,5	-7,8	-3,8	6,3	0,3	1,8	6,8
460	187,5	196,8	190,7	185,3	188,3	208,1	213,4	9,2	3,2	-2,2	0,8	20,6	25,9
461	164,1	165,3	169,5	158,3	166,8	165,4	171,5	1,2	5,4	-5,7	2,8	1,3	7,4
462	114,0	113,7	113,6	109,0	106,6	133,0	112,9	-0,3	-0,5	-5,0	-7,4	18,9	-1,1
463	251,7	250,1	250,8	248,4	248,9	247,1	251,7	-1,6	-1,0	-3,3	-2,8	-4,6	0,0
464	210,1	211,8	218,8	204,7	194,4	188,9	213,8	1,7	8,7	-5,4	-15,7	-21,2	3,7
465	211,8	208,0	209,4	212,5	211,3	217,0	214,9	-3,8	-2,4	0,7	-0,5	5,2	3,1
466	112,9	115,9	111,4	110,6	104,9	117,8	185,3	2,9	-1,5	-2,3	-8,1	4,9	72,4
467	171,6	171,3	173,0	173,7	182,3	187,6	192,1	-0,2	1,5	2,1	10,8	16,1	20,5
468	227,0	222,6	211,0	217,3	230,6	233,0	218,0	-4,4	-16,0	-9,7	3,6	6,0	-9,0
469	276,1	259,8	275,5	265,1	262,6	277,8	278,9	-16,2	-0,5	-10,9	-13,5	1,8	2,8
470	175,4	184,2	188,2	173,1	169,2	170,7	172,6	8,8	12,8	-2,3	-6,2	-4,7	-2,9
471	176,7	165,9	170,0	168,7	169,9	172,4	173,6	-10,8	-6,7	-8,0	-6,8	-4,3	-3,1
472	227,4	223,0	223,2	225,9	229,8	227,8	234,8	-4,4	-4,2	-1,5	2,4	0,4	7,4
473	226,9	255,4	232,5	229,1	223,3	219,0	224,0	28,5	5,6	2,3	-3,5	-7,9	-2,9
474	184,4	184,6	188,2	178,5	181,2	188,5	189,8	0,2	3,7	-5,9	-3,2	4,1	5,3
475	187,5	193,4	184,0	182,5	202,1	199,3	196,9	5,9	-3,5	-5,1	14,6	11,8	9,4
476	260,1	253,0	250,9	257,9	254,8	257,2	259,1	-7,2	-9,3	-2,3	-5,3	-2,9	-1,1
477	214,7	209,1	211,8	213,5	215,2	220,5	220,7	-5,6	-2,9	-1,2	0,4	5,8	6,0
478	181,2	186,8	184,8	167,0	170,2	178,0	164,1	5,5	3,6	-14,2	-11,0	-3,2	-17,1
479	172,3	190,9	193,8	168,4	170,9	176,6	163,1	18,5	21,5	-3,9	-1,4	4,3	-9,3
480	260,7	251,3	250,7	241,2	254,1	244,1	249,9	-9,4	-10,0	-19,6	-6,6	-16,6	-10,9
481	227,2	230,4	225,0	225,4	224,0	228,2	228,4	3,3	-2,1	-1,8	-3,1	1,0	1,3
482	163,4	157,4	152,3	151,5	156,7	160,5	166,8	-6,0	-11,1	-11,9	-6,7	-2,9	3,4
483	128,5	136,8	129,6	129,3	126,6	131,1	136,4	8,3	1,1	0,8	-1,9	2,6	7,9
484	270,3	267,5	265,8	269,9	265,3	267,7	269,5	-2,8	-4,4	-0,4	-5,0	-2,5	-0,8
485	186,6	182,1	189,5	190,0	189,9	193,7	193,1	-4,5	2,9	3,4	3,4	7,1	6,5
486	202,6	198,5	198,5	201,0	201,7	209,5	207,1	-4,1	-4,1	-1,6	-0,9	6,8	4,5

LÍNGUA PORTUGUESA - 8ª SÉRIE **VARIAÇÕES DOS PARÂMETROS b DOS ITENS EM CADA UF**

ITEM	PARÂMETRO b							DIF_b_RN	DIF_b_CE	DIF_b_ES	DIF_b_MS	DIF_b_MG	DIF_b_DF
	SAEB	RN	CE	ES	MS	MG	DF						
430	137,0	142,0	147,0	133,6	130,8	149,7	139,0	-5,0	-10,0	3,4	6,2	-12,7	-1,9
434	125,3	128,3	126,7	124,6	117,6	133,7	203,2	-3,1	-1,4	0,7	7,7	-8,4	-77,9
437	176,7	186,9	185,6	176,2	170,7	180,2	175,5	-10,2	-9,0	0,5	6,0	-3,5	1,2
443	237,4	239,6	233,5	234,8	228,9	234,8	245,7	-2,2	3,9	2,6	8,5	2,5	-8,4
446	318,1	307,6	312,0	309,1	313,0	313,6	316,1	10,5	6,1	9,0	5,1	4,5	2,1
450	290,1	291,9	297,3	292,2	288,5	283,4	284,5	-1,9	-7,2	-2,2	1,5	6,7	5,6
452	248,1	255,2	251,4	253,8	245,3	250,2	252,7	-7,1	-3,4	-5,7	2,7	-2,2	-4,6
458	216,1	223,1	223,4	224,4	219,7	225,1	227,1	-7,0	-7,3	-8,3	-3,5	-9,0	-10,9
460	187,5	196,8	190,7	185,3	188,3	208,1	213,4	-9,2	-3,2	2,2	-0,8	-20,6	-25,9
463	251,7	250,1	250,8	248,4	248,9	247,1	251,7	1,6	1,0	3,3	2,8	4,6	0,0
469	276,1	x	x	x	x	x	x	x	x	x	x	x	x
470	175,4	184,2	188,2	173,1	169,2	170,7	172,5	-8,8	-12,8	2,3	6,2	4,7	2,9
472	227,4	223,0	223,2	225,9	229,8	227,8	234,8	4,4	4,2	1,5	-2,4	-0,4	-7,4
476	260,1	252,9	250,9	257,9	254,8	257,2	259,1	7,2	9,3	2,3	5,3	2,9	1,1
477	214,7	209,1	211,8	213,5	215,2	220,5	220,7	5,6	2,9	1,2	-0,4	-5,8	-6,0
478	181,2	186,8	184,8	167,0	170,2	178,0	164,1	-5,5	-3,6	14,2	11,0	3,2	17,1
479	172,3	190,8	193,8	168,4	170,9	176,6	163,1	-18,5	-21,5	3,9	1,4	-4,3	9,3
480	260,7	251,3	250,7	241,2	254,1	244,1	249,9	9,4	10,0	19,6	6,6	16,6	10,9
484	270,3	267,5	265,8	269,9	265,3	267,7	269,5	2,8	4,4	0,4	5,0	2,5	0,7
486	202,6	198,5	198,5	201,0	201,7	209,5	207,1	4,1	4,1	1,6	0,9	-6,8	-4,5
487	249,9	266,0	274,0	262,9	283,1	269,3	269,6	-16,1	-24,1	-13,0	-33,3	-19,4	-19,7
488	330,0	337,5	332,9	327,5	324,6	327,8	334,5	-7,5	-2,9	2,5	5,4	2,2	-4,5
489	183,1	178,5	183,2	195,4	189,5	205,2	217,8	4,6	-0,1	-12,3	-6,3	-22,1	-34,7
490	192,0	197,6	193,3	189,8	188,3	195,8	199,4	-5,6	-1,3	2,2	3,7	-3,7	-7,4
491	334,6	342,0	345,5	334,4	345,0	339,3	347,7	-7,4	-10,9	0,2	-10,5	-4,7	-13,1
492	315,8	302,7	300,5	301,9	298,9	300,3	310,7	13,1	15,3	13,9	16,9	15,5	5,1
493	219,4	229,5	228,9	228,8	223,9	225,1	228,4	-10,2	-9,5	-9,4	-4,5	-5,7	-9,0
494	339,5	333,1	326,3	336,4	340,4	332,0	357,4	6,4	13,2	3,1	-0,9	7,5	-17,9
495	357,0	332,2	333,5	345,8	331,9	347,1	350,2	24,8	23,5	11,2	25,1	9,9	6,8
496	286,2	285,6	284,3	283,3	284,4	294,6	296,5	0,6	1,9	2,8	1,7	-8,4	-10,3

LÍNGUA PORTUGUESA - 8ª SÉRIE **VARIAÇÕES DOS PARÂMETROS b DOS ITENS EM CADA UF** *continuação*

ITEM	PARÂMETRO b							DIF_b_RN	DIF_b_CE	DIF_b_ES	DIF_b_MS	DIF_b_MG	DIF_b_DF
	SAEB	RN	CE	ES	MS	MG	DF						
497	326,4	311,5	310,7	319,7	315,5	329,9	332,2	14,9	15,7	6,7	11,0	-3,5	-5,8
498	177,7	182,5	183,5	188,2	179,1	185,7	189,8	-4,8	-5,8	-10,5	-1,4	-8,0	-12,1
499	342,0	311,0	321,3	331,5	328,7	334,7	334,0	31,1	20,8	10,6	13,3	7,3	8,0
500	199,8	200,5	200,0	200,4	200,1	202,7	208,8	-0,6	-0,2	-0,6	-0,3	-2,9	-8,9
501	310,1	296,9	294,3	306,7	311,8	311,2	310,9	13,2	15,8	3,4	-1,7	-1,1	-0,8
502	207,3	192,4	197,5	189,4	189,3	191,2	194,3	14,9	9,9	18,0	18,1	16,1	13,1
503	294,5	337,6	312,0	304,3	292,2	292,8	311,2	-43,1	-17,5	-9,8	2,3	1,7	-16,6
504	266,5	251,6	249,8	266,8	266,3	268,3	270,2	15,0	16,7	-0,3	0,2	-1,7	-3,7
505	278,0	283,4	292,1	280,8	287,5	285,3	297,5	-5,4	-14,1	-2,8	-9,4	-7,3	-19,5
506	246,5	248,3	244,3	244,5	241,4	248,7	251,3	-1,8	2,2	2,1	5,2	-2,1	-4,7
507	270,2	272,3	270,3	255,3	260,3	272,5	270,2	-2,1	-0,1	14,9	10,0	-2,2	0,1
508	215,1	197,3	200,9	197,9	252,8	221,4	203,8	17,8	14,2	17,2	-37,7	-6,2	11,3
509	433,3	368,5	375,7	395,7	382,8	371,0	390,7	64,8	57,6	37,6	50,4	62,3	42,6
510	258,2	259,4	256,2	253,4	258,2	249,1	239,0	-1,2	2,0	4,8	0,0	9,1	19,2
511	313,6	306,8	307,4	320,0	312,0	312,6	323,6	6,9	6,3	-6,4	1,6	1,1	-10,0
512	340,1	340,0	343,7	339,8	324,8	334,6	332,4	0,1	-3,6	0,3	15,3	5,5	7,7
513	353,1	338,0	364,3	355,1	360,3	364,8	368,6	15,1	-11,2	-2,0	-7,2	-11,7	-15,5
514	252,3	245,3	249,9	265,9	246,3	267,8	270,2	7,0	2,4	-13,6	6,0	-15,5	-17,9
515	283,1	270,5	285,5	280,9	287,8	280,6	301,2	12,6	-2,4	2,2	-4,7	2,5	-18,1
516	302,6	291,9	294,4	302,3	307,7	300,2	318,4	10,7	8,2	0,3	-5,2	2,4	-15,8
517	264,8	260,6	260,6	265,8	263,0	265,8	272,8	4,2	4,2	-1,0	1,8	-1,0	-8,0
518	309,3	307,3	311,5	317,3	312,5	322,8	322,3	2,0	-2,2	-8,0	-3,2	-13,5	-13,0
519	255,0	259,5	261,2	262,3	271,5	265,0	251,7	-4,6	-6,2	-7,3	-16,6	-10,1	3,2
520	232,0	224,4	232,5	240,6	223,2	245,4	238,7	7,6	-0,6	-8,6	8,7	-13,4	-6,8
521	222,5	237,3	237,7	242,8	231,6	233,8	239,2	-14,8	-15,2	-20,4	-9,1	-11,4	-16,8
522	351,6	x	x	355,0	361,2	363,5	355,0	x	x	-3,4	-9,6	-11,9	-3,4
523	221,2	234,6	227,5	225,3	218,6	225,6	211,4	-13,3	-6,3	-4,0	2,6	-4,4	9,8
524	315,4	310,3	313,8	320,1	314,7	318,9	330,4	5,1	1,6	-4,8	0,7	-3,6	-15,0
525	220,4	224,5	227,6	228,0	229,7	226,9	232,3	-4,1	-7,2	-7,6	-9,4	-6,6	-11,9
526	298,8	255,4	257,2	299,4	297,2	307,7	294,3	43,4	41,6	-0,6	1,6	-8,9	4,5

LÍNGUA PORTUGUESA - 8ª SÉRIE **VARIAÇÕES DOS PARÂMETROS b DOS ITENS EM CADA UF** *continuação*

ITEM	PARÂMETRO b							DIF_b_RN	DIF_b_CE	DIF_b_ES	DIF_b_MS	DIF_b_MG	DIF_b_DF
	SAEB	RN	CE	ES	MS	MG	DF						
527	270,2	274,9	270,5	274,8	272,5	273,3	284,3	-4,7	-0,3	-4,5	-2,2	-3,1	-14,1
528	274,0	281,7	285,6	287,0	283,0	286,0	298,8	-7,7	-11,6	-13,0	-9,0	-12,0	-24,8
529	306,3	307,1	309,4	303,4	296,7	300,5	303,7	-0,8	-3,0	2,9	9,6	5,8	2,6
530	294,1	284,1	279,3	284,9	285,1	279,7	294,8	10,0	14,8	9,1	8,9	14,3	-0,8
531	174,1	174,4	186,9	173,4	171,5	186,2	183,6	-0,3	-12,8	0,7	2,6	-12,1	-9,5
532	278,8	284,1	285,3	277,8	275,8	272,7	280,4	-5,3	-6,4	1,1	3,0	6,1	-1,5
533	228,3	240,5	239,4	227,5	221,2	232,3	232,5	-12,2	-11,1	0,8	7,0	-4,0	-4,2
534	309,6	311,0	310,2	307,1	308,8	303,8	313,9	-1,3	-0,6	2,6	0,8	5,9	-4,2
535	354,7	340,2	325,3	332,6	336,0	330,3	352,7	14,5	29,4	22,1	18,7	24,4	2,0

MATEMÁTICA - 4ª SÉRIE **VARIAÇÕES DOS PARÂMETROS b DOS ITENS EM CADA UF**

ITEM	PARÂMETRO b							DIF_b_RN	DIF_b_CE	DIF_b_ES	DIF_b_MS	DIF_b_MG	DIF_b_DF
	SAEB	RN	CE	ES	MS	MG	DF						
430	159,1	162,1	153,5	155,6	160,0	173,3	161,1	-3,0	5,7	3,5	-0,9	-14,1	-2,0
431	291,2	296,3	304,6	300,8	284,2	285,5	283,0	-5,2	-13,5	-9,6	6,9	5,7	8,2
432	228,8	221,1	225,7	234,0	237,6	228,4	214,1	7,7	3,1	-5,2	-8,8	0,4	14,7
433	160,9	154,9	152,6	162,6	162,3	161,2	160,0	6,1	8,4	-1,7	-1,4	-0,2	0,9
434	147,0	142,7	150,5	142,7	151,4	153,3	151,2	4,3	-3,5	4,3	-4,4	-6,4	-4,2
435	287,9	279,8	287,1	279,2	280,3	286,6	284,5	8,1	0,8	8,7	7,6	1,3	3,4
436	266,2	259,1	249,2	262,0	272,5	265,0	254,1	7,1	17,0	4,2	-6,2	1,2	12,1
437	184,0	169,3	189,9	190,1	194,5	206,3	188,5	14,7	-6,0	-6,1	-10,5	-22,3	-4,5
438	271,4	248,4	260,4	265,4	270,3	276,8	269,6	23,0	11,0	5,9	1,1	-5,4	1,7
439	192,3	191,2	186,7	184,9	201,5	208,0	190,5	1,1	5,6	7,4	-9,2	-15,6	1,8
440	327,2	319,6	319,2	325,9	319,8	324,9	336,5	7,6	8,0	1,3	7,4	2,3	-9,2
441	328,2	x	x	x	x	319,2	303,5	x	x	x	x	9,0	24,8
442	131,5	137,5	136,0	127,3	127,8	139,5	196,5	-5,9	-4,5	4,3	3,7	-7,9	-65,0
443	219,0	223,7	226,1	210,0	215,6	220,5	211,4	-4,8	-7,1	8,9	3,4	-1,6	7,6
444	218,3	202,0	207,8	216,8	220,1	226,3	227,6	16,3	10,5	1,5	-1,8	-8,0	-9,3
445	270,6	259,4	262,0	267,0	268,5	273,2	269,0	11,3	8,7	3,6	2,1	-2,6	1,7
446	225,4	212,9	224,9	222,6	221,5	238,6	230,9	12,5	0,5	2,8	3,8	-13,3	-5,5
447	261,2	263,7	258,5	263,4	259,1	253,7	268,2	-2,5	2,7	-2,1	2,1	7,6	-6,9
448	311,0	312,4	310,3	305,8	306,7	308,0	311,0	-1,4	0,6	5,2	4,3	3,0	-0,1
449	307,9	297,0	287,9	304,8	315,2	316,7	303,1	11,0	20,0	3,1	-7,2	-8,8	4,8
450	193,6	198,7	200,9	191,9	197,2	181,1	186,0	-5,2	-7,3	1,7	-3,6	12,5	7,6
451	248,6	242,2	250,3	238,5	259,5	265,3	244,2	6,4	-1,7	10,1	-10,9	-16,6	4,4
452	302,2	289,7	289,5	289,1	289,3	298,8	302,4	12,6	12,7	13,1	13,0	3,4	-0,2
453	166,5	170,3	168,0	163,3	169,8	162,7	157,4	-3,8	-1,5	3,2	-3,3	3,8	9,1
454	213,4	196,3	182,1	211,7	196,6	192,7	192,8	17,0	31,2	1,7	16,7	20,6	20,5
455	286,0	282,4	285,9	285,4	284,2	297,8	289,1	3,6	0,1	0,7	1,9	-11,7	-3,0
456	296,3	248,8	293,7	296,9	291,8	275,0	292,8	47,5	2,6	-0,6	4,4	21,2	3,5
457	283,8	266,2	281,4	281,7	291,1	255,6	281,2	17,6	2,4	2,1	-7,3	28,2	2,7
458	241,0	249,1	244,2	237,7	246,5	238,0	237,6	-8,1	-3,1	3,3	-5,5	3,0	3,4
459	349,0	330,3	336,5	341,0	347,7	350,9	332,4	18,7	12,6	8,0	1,3	-1,8	16,7

MATEMÁTICA - 4ª SÉRIE **VARIAÇÕES DOS PARÂMETROS b DOS ITENS EM CADA UF** *continuação*

ITEM	PARÂMETRO b							DIF_b_RN	DIF_b_CE	DIF_b_ES	DIF_b_MS	DIF_b_MG	DIF_b_DF
	SAEB	RN	CE	ES	MS	MG	DF						
460	244,1	224,4	229,9	233,5	247,3	245,2	231,6	19,7	14,2	10,6	-3,2	-1,1	12,5
461	208,0	210,2	207,9	208,6	210,7	214,7	208,1	-2,2	0,1	-0,5	-2,7	-6,6	-0,1
462	172,7	176,7	174,2	173,9	170,1	174,8	176,8	-4,0	-1,6	-1,2	2,6	-2,2	-4,2
463	211,9	186,0	196,0	205,1	213,7	209,6	212,7	25,9	15,8	6,8	-1,8	2,3	-0,8
464	274,3	266,5	272,4	278,5	286,5	270,9	277,1	7,9	2,0	-4,2	-12,2	3,4	-2,8
465	136,4	133,6	137,1	133,8	130,1	152,9	152,0	2,8	-0,7	2,7	6,3	-16,4	-15,6
466	188,3	174,5	185,7	179,8	190,1	182,8	191,5	13,7	2,6	8,5	-1,8	5,5	-3,2
467	103,9	106,5	107,1	94,8	92,8	115,7	219,7	-2,6	-3,2	9,1	11,0	-11,8	-115,8
468	291,4	289,7	295,9	291,6	283,0	290,4	291,6	1,7	-4,5	-0,2	8,4	1,0	-0,2
469	227,6	216,9	230,5	222,1	230,8	231,2	229,0	10,7	-2,9	5,6	-3,2	-3,6	-1,4
470	142,9	161,6	147,1	135,1	136,2	162,5	134,6	-18,7	-4,2	7,8	6,7	-19,6	8,3
471	291,4	260,5	271,7	289,3	272,1	278,2	302,6	30,8	19,7	2,1	19,2	13,2	-11,2
472	244,1	223,0	233,0	238,4	239,3	244,0	246,8	21,2	11,1	5,8	4,8	0,2	-2,7
473	231,6	253,6	264,3	238,4	263,4	231,8	244,8	-22,1	-32,8	-6,8	-31,8	-0,3	-13,2
474	230,8	224,0	232,5	224,8	240,4	223,0	225,4	6,8	-1,7	6,0	-9,6	7,8	5,4
475	156,6	150,4	155,1	156,3	147,5	167,4	159,6	6,2	1,5	0,2	9,1	-10,8	-3,0
476	207,6	202,7	215,8	202,2	218,8	213,3	208,6	5,0	-8,1	5,4	-11,2	-5,6	-0,9
477	250,8	238,7	243,3	249,2	251,7	254,9	251,9	12,1	7,5	1,6	-0,9	-4,1	-1,1
478	237,5	x	x	x	x	x	x	x	x	x	x	x	x
479	218,0	209,9	216,1	218,2	225,9	225,7	224,9	8,2	1,9	-0,2	-7,9	-7,7	-6,9
480	214,8	192,8	192,7	213,4	206,3	213,3	217,9	21,9	22,1	1,4	8,5	1,5	-3,1
481	293,5	x	291,5	287,6	285,0	290,0	292,5	x	2,0	5,9	8,5	3,5	1,0
482	230,2	226,6	225,2	228,3	234,3	224,9	225,1	3,5	5,0	1,9	-4,1	5,3	5,0
483	93,9	95,5	99,4	91,7	172,3	126,3	120,6	-1,6	-5,5	2,2	-78,4	-32,4	-26,7
484	160,5	141,8	157,6	153,4	168,8	164,2	160,2	18,7	2,9	7,1	-8,3	-3,7	0,3
485	135,3	136,3	136,7	129,4	143,8	145,7	145,9	-0,9	-1,4	6,0	-8,4	-10,4	-10,5
486	246,5	237,3	239,0	242,3	252,7	242,7	234,8	9,1	7,5	4,2	-6,2	3,8	11,7
487	341,5	352,0	329,9	363,0	363,8	343,2	341,4	-10,5	11,6	-21,5	-22,2	-1,7	0,2
488	146,0	164,2	155,3	144,6	147,4	170,9	140,2	-18,2	-9,2	1,5	-1,4	-24,8	5,9

MATEMÁTICA - 8ª SÉRIE **VARIAÇÕES DOS PARÂMETROS b DOS ITENS EM CADA UF**

ITEM	PARÂMETRO b							DIF_b_RN	DIF_b_CE	DIF_b_ES	DIF_b_MS	DIF_b_MG	DIF_b_DF
	SAEB	RN	CE	ES	MS	MG	DF						
431	x	x	x	x	x	x	291,2	-13,5	-5,2	-9,6	5,7	8,2	6,9
435	287,1	279,8	279,2	286,6	284,5	280,3	287,9	0,8	8,1	8,7	1,3	3,4	7,6
437	189,9	169,3	190,1	206,3	188,5	194,5	184,0	-6,0	14,7	-6,1	-22,3	-4,5	-10,5
439	x	x	x	x	x	x	192,3	5,6	1,1	7,4	-15,6	1,8	-9,2
440	319,2	319,6	325,9	324,9	336,5	319,8	327,2	8,0	7,6	1,3	2,3	-9,2	7,4
443	226,1	223,7	210,0	220,5	211,4	215,6	219,0	-7,1	-4,8	8,9	-1,6	7,6	3,4
444	207,8	202,0	216,8	226,3	227,6	220,1	218,3	10,5	16,3	1,5	-8,0	-9,3	-1,8
445	262,0	259,4	267,0	273,2	269,0	268,5	270,6	8,7	11,3	3,6	-2,6	1,7	2,1
447	258,5	263,7	263,4	253,7	268,2	259,1	261,2	2,7	-2,5	-2,1	7,6	-6,9	2,1
448	310,3	312,4	305,8	308,0	311,0	306,7	311,0	0,6	-1,4	5,2	3,0	-0,1	4,3
452	289,5	289,7	289,1	298,8	302,4	289,3	302,2	12,7	12,6	13,1	3,4	-0,2	13,0
453	168,0	170,3	163,3	162,7	157,4	169,8	166,5	-1,5	-3,8	3,2	3,8	9,1	-3,3
459	336,5	330,3	341,0	350,9	332,4	347,7	349,0	12,6	18,7	8,0	-1,8	16,7	1,3
461	207,9	210,2	208,6	214,7	208,1	210,7	208,0	0,1	-2,2	-0,5	-6,6	-0,1	-2,7
462	174,2	176,7	173,9	174,8	176,8	170,1	172,7	-1,6	-4,0	-1,2	-2,2	-4,2	2,6
468	295,9	289,7	291,6	290,4	291,6	283,0	291,4	-4,5	1,7	-0,2	1,0	-0,2	8,4
470	147,1	161,6	135,1	162,5	134,6	136,2	142,9	-4,2	-18,7	7,8	-19,6	8,3	6,7
477	243,3	238,7	249,2	254,9	251,9	251,7	250,8	7,5	12,1	1,6	-4,1	-1,1	-0,9
479	216,1	209,9	218,2	225,7	224,9	225,9	218,0	1,9	8,2	-0,2	-7,7	-6,9	-7,9
481	291,5	288,0	287,6	290,0	292,5	285,0	293,5	2,0	5,5	5,9	3,5	1,0	8,5
487	329,9	352,0	363,0	343,2	341,4	363,8	341,5	11,6	-10,5	-21,5	-1,7	0,2	-22,2
489	318,5	314,5	324,1	328,6	327,0	326,0	325,7	7,3	11,3	1,7	-2,8	-1,3	-0,3
490	209,5	198,7	221,8	221,3	194,1	205,8	243,1	33,6	44,4	21,3	21,8	49,0	37,3
491	334,9	326,8	328,7	339,4	336,5	328,2	333,6	-1,3	6,8	4,9	-5,7	-2,9	5,4
492	286,3	294,4	290,5	292,7	293,2	296,4	292,6	6,4	-1,7	2,1	-0,1	-0,6	-3,8
493	287,0	287,8	287,7	290,8	271,4	289,8	288,0	1,0	0,3	0,3	-2,8	16,7	-1,8
494	347,0	359,8	360,8	364,5	356,0	357,0	355,5	8,5	-4,3	-5,3	-9,0	-0,5	-1,5
495	358,0	342,8	370,3	367,4	359,7	369,0	362,3	4,3	19,6	-8,0	-5,1	2,7	-6,7
496	336,6	335,5	338,5	338,8	342,4	336,5	339,0	2,4	3,5	0,5	0,2	-3,4	2,5
497	279,4	277,0	278,1	284,3	273,1	271,2	283,8	4,3	6,8	5,7	-0,6	10,6	12,6

MATEMÁTICA - 8ª SÉRIE **VARIAÇÕES DOS PARÂMETROS b DOS ITENS EM CADA UF** *continuação*

ITEM	PARÂMETRO b							DIF_b_RN	DIF_b_CE	DIF_b_ES	DIF_b_MS	DIF_b_MG	DIF_b_DF
	SAEB	RN	CE	ES	MS	MG	DF						
498	x	x	398,8	414,1	388,2	383,2	409,5	172,0	172,0	10,6	-4,7	21,3	26,3
499	317,1	310,5	323,2	323,4	322,6	320,0	320,7	3,6	10,2	-2,5	-2,7	-1,9	0,8
500	199,9	207,0	200,4	210,1	213,6	203,8	214,5	14,6	7,5	14,1	4,5	0,9	10,8
501	330,1	321,5	338,2	339,5	347,4	338,1	327,1	-3,0	5,6	-11,2	-12,4	-20,3	-11,0
502	328,6	324,8	334,0	329,9	333,8	334,8	342,3	13,7	17,6	8,3	12,4	8,5	7,5
503	274,2	268,1	281,3	285,5	281,4	285,2	283,9	9,7	15,8	2,6	-1,5	2,5	-1,3
504	327,1	342,3	340,9	341,8	339,9	348,0	332,3	5,2	-10,0	-8,5	-9,5	-7,6	-15,7
505	331,4	312,8	340,9	342,5	346,3	345,4	342,6	11,1	29,8	1,7	0,1	-3,7	-2,8
506	334,2	343,4	348,3	346,7	337,7	341,4	348,6	14,4	5,2	0,4	1,9	11,0	7,2
507	292,7	255,4	292,7	308,8	273,8	300,3	316,9	24,2	61,4	24,2	8,1	43,1	16,6
508	315,3	319,4	318,9	320,4	325,7	327,5	318,3	2,9	-1,1	-0,7	-2,1	-7,5	-9,2
509	197,3	191,4	196,1	196,4	195,9	195,1	195,0	-2,3	3,6	-1,1	-1,4	-0,9	-0,1
510	304,8	303,3	302,2	309,2	309,3	299,8	312,4	7,7	9,1	10,2	3,2	3,1	12,6
511	x	x	423,3	394,8	387,1	x	399,9	162,4	162,4	-23,4	5,1	12,8	162,4
512	350,6	345,3	357,5	359,7	360,6	349,3	359,8	9,2	14,5	2,3	0,2	-0,8	10,6
513	389,8	381,6	382,3	387,7	398,8	387,8	384,5	-5,3	2,9	2,2	-3,2	-14,3	-3,3
514	256,4	258,1	244,4	246,4	257,4	251,4	257,6	1,2	-0,5	13,3	11,2	0,3	6,3
515	373,6	346,6	358,7	359,4	358,8	369,2	356,6	-17,0	10,0	-2,1	-2,9	-2,2	-12,7
516	340,2	355,1	342,7	350,3	344,3	345,8	347,0	6,8	-8,1	4,3	-3,3	2,7	1,2
517	342,9	346,2	352,0	357,1	355,4	355,0	355,1	12,2	8,9	3,1	-2,0	-0,3	0,1
518	342,3	352,5	344,7	341,5	348,9	350,5	349,1	6,9	-3,4	4,4	7,6	0,2	-1,4
519	320,1	318,5	328,6	337,4	348,2	327,2	352,2	32,1	33,8	23,6	14,8	4,0	25,0
520	270,5	266,9	269,4	270,3	267,6	269,6	270,7	0,2	3,8	1,2	0,4	3,1	1,1
521	322,6	307,0	324,1	325,5	322,6	319,9	333,0	10,4	26,0	8,8	7,5	10,3	13,1
522	264,0	260,8	259,5	266,2	267,9	267,6	280,7	16,7	19,9	21,2	14,5	12,8	13,0
523	294,1	294,8	288,8	291,2	302,3	299,3	300,3	6,2	5,5	11,5	9,0	-2,1	1,0
524	330,2	317,8	332,0	327,0	324,5	326,0	330,5	0,3	12,7	-1,5	3,5	6,1	4,5
525	348,6	338,6	354,8	350,2	355,5	339,3	344,7	-3,9	6,1	-10,1	-5,6	-10,8	5,3
526	236,6	240,1	242,0	244,4	241,8	240,0	239,8	3,2	-0,4	-2,2	-4,7	-2,0	-0,3
527	336,0	336,2	345,4	346,9	355,1	342,3	358,7	22,6	22,4	13,3	11,8	3,6	16,3

MATEMÁTICA - 8ª SÉRIE **VARIAÇÕES DOS PARÂMETROS b DOS ITENS EM CADA UF** *continuação*

ITEM	PARÂMETRO b							DIF_b_RN	DIF_b_CE	DIF_b_ES	DIF_b_MS	DIF_b_MG	DIF_b_DF
	SAEB	RN	CE	ES	MS	MG	DF						
528	270,0	271,8	267,9	267,1	273,1	267,4	275,8	5,8	4,0	7,8	8,6	2,7	8,4
529	316,9	309,8	316,2	320,5	314,9	313,5	321,5	4,6	11,6	5,2	1,0	6,6	7,9
530	308,0	318,0	312,9	322,1	335,3	318,7	330,0	22,0	12,0	17,1	7,9	-5,3	11,3
531	229,6	221,7	226,2	226,3	220,9	220,2	226,7	-2,9	5,0	0,5	0,4	5,8	6,6
532	322,8	337,7	330,9	345,8	337,0	329,7	338,7	15,9	1,0	7,8	-7,1	1,7	9,0
533	286,8	281,9	287,0	293,2	305,7	287,5	304,6	17,8	22,6	17,5	11,4	-1,2	17,0
534	272,5	268,7	271,5	274,7	274,7	268,5	277,5	5,0	8,9	6,0	2,8	2,9	9,0
535	270,5	273,2	270,2	272,0	274,8	272,5	269,9	-0,6	-3,3	-0,3	-2,1	-4,8	-2,6
536	329,1	327,8	328,6	328,2	341,9	343,1	332,1	3,1	4,3	3,5	3,9	-9,8	-10,9
537	287,7	279,4	287,6	285,6	294,9	299,7	292,0	4,3	12,7	4,5	6,4	-2,9	-7,6