Instituto de Psicologia

Universidade de Brasília

A (In)Dependência da Habilidade Estimada pela Teoria de Resposta ao Item em Relação à Dificuldade da Prova: Um Estudo com os Dados do Saeb

FREDERICO NEVES CONDÉ

Orientador: PROF. JACOB ARIE LAROS

Brasília – DF

2002

A (In)Dependência da Habilidade Estimada pela Teoria de Resposta ao Item em Relação à Dificuldade da Prova: Um Estudo com os Dados do Saeb

Título Abreviado: (In)Dependência da Habilidade pela TRI em Relação à Dificuldade da Prova.

Dissertação apresentada ao curso de mestrado do Programa de Pós-graduação do Instituto de Psicologia da Universidade de Brasília como requisito à obtenção do grau de Mestre em Psicologia.

Brasília, 18 de outubro de 2002

Instituto de Psicologia

Universidade de Brasília

Esta dissertação será avaliada pela seguinte comissão examinadora:

_____

Prof. Jacob Arie Laros – Presidente

Universidade de Brasília

_____

Prof. Luiz Pasquali

Universidade de Brasília

_____

Prof. Hartmut Günther

Universidade de Brasília

Brasília, 18 de outubro de 2002

Dedico esse trabalho especialmente ao meu filho Arthur, essa criança linda que, com certeza, será meu grande companheiro nos melhores momentos da vida ("vida esta que ainda há de nos aproximar...");

Aos meus pais Paulo e Maria Léa, que novamente uniram e dedicaram todos os seus esforços para me incentivar em direção à consecução de mais esta etapa de minha formação;

Aos meus irmãos e companheiros Fabrício e Fabiano, pela verdadeira relação que nos une.

À minha namorada Maria Fernanda, essa pessoa especial que conheci há pouco tempo e aprendi a gostar muito. Agradeço toda sua atenção e apoio para a realização deste trabalho.

Agradecimentos

Ao Professor e Orientador Jaap Laros, excelente pessoa e profissional, com quem venho adquirindo um grande aprendizado na área de psicometria; à sua disponibilidade para discussão sobre o tema do presente estudo, zelo e precisão no acompanhamento dos aspectos técnicos da obra.

Ao professor Luiz Pasquali, maior responsável por meu interesse pelo estudo da medida em Psicologia.

Ao Instituto Nacional de Estudos e Pesquisas Educacionais, pela cessão da utilização das bases de dados do Saeb ao Laboratório de Pesquisa em Avaliação e Medida – LabPAM. Ao próprio LabPam e ao Centro de Pesquisa em Avaliação Educacional – CPAE da Universidade de Brasília - UnB, pela oportunidade da realização do estudo.

Aos professores Hartmut Günther e Batholomeu Tróccoli, pela aceitação do convite de participação como membros da comissão examinadora de minha dissertação e por todos os incentivos e ensinamentos em relação à pesquisa empírica em Psicologia.

À Profa Iza Locatelli, por seu imenso apoio à pesquisa enquanto Diretora da Avaliação da Educação Básica do INEP e por seu incentivo para o crescimento profissional e fortalecimento técnico.

Ao amigo Guilherme Coelho Rabello, por ter lançado a idéia principal da pesquisa e percebido que o delineamento do Saeb permite excelentes estudos sobre a invariância dos parâmetros. Agradeço-o por todos os seus ensinamentos na área de estatísticas e medidas sociais.

Ao amigo Eduardo de São Paulo, por toda a sua atenção e orientação no desenvolvimento do estudo e por sua dedicada cooperação em minha formação profissional.

Ao Prof. Ruben Klein e ao pesquisador Marcus Riether pela troca de idéias sobre algumas análises de dados realizadas.

Ao meu pai Paulo Condé por sua colaboração na revisão da dissertação.

À Amélia Regina Alves, minha amiga e professora, que me ajudou muito na definição de minhas diretrizes, tanto éticas, quanto técnicas de minha formação.

Aos amigos Paulo, Robson, Karina e Margarida pela oportunidade de discussões sempre bastante produtivas.

Índice

Lista de Tabelas

Lista de Figuras

Resumo

A Teoria de Resposta ao Item (TRI) assume a existência da propriedade de invariância dos parâmetros, que permite estimar a habilidade dos sujeitos (*theta*) independentemente da forma do teste utilizado. Esta propriedade se baseia em pelo menos duas condições relacionadas aos itens do teste: estar na mesma escala e atender ao pressuposto de unidimensionalidade. O objetivo do presente estudo é o de investigar se a estimativa de *theta* independe da dificuldade dos itens utilizados para estimá-la bem como verificar em que medida a unidimensionalidade da prova influencia nesta propriedade. Foram utilizados os dados secundários de 26 formas de prova de Matemática de 8ª Série do Ensino Fundamental (E.F.) do Sistema Nacional de Avaliação da Educação Básica (Saeb), aplicada em 1997, em uma amostra de 18.806 estudantes brasileiros de escolas públicas e particulares de cada uma das 27 Unidades da Federação brasileiras. Essas formas de prova foram respondidas por 26 grupos diferentes de estudantes, equivalentes em termos de habilidades. Foram correlacionados os resultados médios de índices de dificuldade das provas e habilidade dos estudantes calculados pela Teoria Clássica dos Testes (TCT) e pela TRI. Os resultados apontam para a existência de uma dependência do *theta* em relação à dificuldade dos cadernos (r = 0,68, com o *valor p*; r = - 0,69 com o *parâmetro b*), menor que a do índice de habilidade calculado pela TCT, o *escore total*, em relação à dificuldade (r = - 0,95, com o *parâmetro b*). A dependência entre o *theta* e a dificuldade diminui quando são excluídos da prova os itens com cargas fatoriais inferiores a 0,20 no fator único, que praticamente não contribuem para a unidimensionalidade. Observou-se, neste caso, um coeficiente de correlação com o *valor p* de 0,60 e, com o *parâmetro b*, de – 0,57. Conclui-se que o *theta* estimado depende da dificuldade dos itens que são utilizados para estimá-lo, não confirmando a propriedade de invariância dos parâmetros. O *theta* apresenta uma dependência menor com a dificuldade, que a observada pelo *escore total.* Por sua vez esta estimativa da TRI apresenta uma diminuição da dependência com relação à dificuldade quando a prova se aproxima da unidimensionalidade. Os resultados indicam ser vantajosa a utilização da TRI para estimar a habilidade dos estudantes, quando são utilizadas formas de teste com dificuldades diferentes, pois o *theta* é menos dependente da dificuldade que o *escore total*. Faz-se necessário um maior rigor no controle da condição de unidimensionalidade da prova para a obtenção de estimativas de *theta* mais invariantes.

Palavras-chave: invariância dos parâmetros, Teoria de Resposta ao Item, independência dos parâmetros, unidimensionalidade, BIB.

Abstract

The Item Response Theory (IRT) assumes the existence of a property of invariance of the parameters, which implies that the parameter that characterizes the ability of an examinee (*theta*) does not depend on the set of items of the test forms. This property is based on at least two conditions related to the test items: being in a common scale and being in accordance with the assumption of unidimensionality. The objective of the present study is to investigate whether the estimation of *theta* does not depend on the difficulty of the items used to estimate it, as well as how much the unidimensionality of the test items influences this property. In this study, secundary data were used of 26 test forms of Mathematics, eight grade, from the National System of Evaluation of the Basic Education (Saeb), applied in 1997, in a sample of 18.806 Brazilian students of public and particular schools of the 27 Brazilian states. The test forms were answered by 26 different groups of students, equivalent in terms of abilities. The results of indexes of the tests' difficulty and the students' ability were calculated and correlated using both the Classic Theory of Tests (CTT) and the IRT. The results point to the existence of a dependence of the *theta* in relation to the difficulty of the booklets (r = 0,68, for the *p value*; r = - 0,69 for the *parameter b*), smaller than the one of the ability index calculated by CTT, the *total score*, in relation to the difficulty (r = - 0,95, with the *parameter b*). The dependence between *theta* and the difficulty decreases when items with factor loadings less than 0,20 in the only factor of the test, that practically don't contribute to the unidimensionality, are excluded. A correlation with the *p value* of 0,60 and, with the *parameter b*, of - 0,57, was observed, in this case. It was concluded that the estimated of *theta* is related to the difficulty of the items, not confirming the property of parameter invariance. This relation is less strong than the one between the *total score* and the difficulty of the items. The estimated of *theta* presents a dependence that decreases with the difficulty when the test approaches the unidimensionality. The use of IRT is considered advantageous to esteem the students' ability, when test forms present different difficulties, *theta* is less dependent on the difficulty than the total score. It is considered necessary a greater rigidity on the control of the conditions for the unidimensionality of the tests in order to obtain more invariant estimates of *theta*.

Key-Words: parameters invariance, Item Response Theory, parameters independence, unidimensionality, BIB.

1. Introdução

1.1 A propriedade da invariância dos parâmetros pela Teoria de Resposta ao Item

A Teoria de Resposta ao Item (TRI) é composta de um conjunto de modelos estatísticos que se estrutura por meio de uma série de pressupostos e propriedades e envolve procedimentos de estimação de parâmetros. Sua aplicação na teoria psicométrica se mostrou bastante conveniente e útil acompanhando o paradigma baseado na Teoria do Traço Latente, também denominada de Modelos de Traço Latente ou de Modelos Estruturais Latentes (Requena, 1990). Esse paradigma especifica uma relação teórica entre as pontuações empíricas dos examinandos em uma prova ou teste e o traço latente não observável, teorizado como o responsável por tais pontuações.

A TRI fornece modelos que atribuem parâmetros para itens e para indivíduos separadamente de forma a predizer probabilisticamente a resposta de qualquer indivíduo a qualquer item. "As funções de resposta ao item estabelecem as relações, matematicamente formalizadas, de como cada resposta depende de um certo nível ou grau de habilidade (...) no traço considerado" (Requena, 1990). Quando a Psicometria se apropria desses modelos, percebe-se que seus parâmetros matemáticos podem ser utilizados como meio de caracterização de itens de testes. Geralmente, os itens podem ser avaliados por meio de modelos de um, dois ou três parâmetros. O modelo de um parâmetro envolve apenas a *dificuldade* (parâmetro *b*); o de dois envolve o parâmetro *b* e a *discriminação* (parâmetro *a*); e o de três parâmetros envolve os parâmetros *a*, *b* e o de *acerto ao acaso* (parâmetro *c*). Percebe-se também que um outro parâmetro, o *theta*, apresenta características que podem ser atribuídas à habilidade de cada um dos indivíduos testados.

Não se sabe ao certo os limites da TRI, mas ela é apresentada por estudiosos, como por exemplo, Hambleton, Swaminathan e Rogers (1991), como capaz de fornecer

contribuições na construção de testes, na identificação de viés de itens, na equalização de resultados de desempenho de examinandos em resposta a diferentes testes ou de diferentes formas de um mesmo teste e na apresentação ou relato desses resultados. Para esses autores, a TRI supera certas limitações teóricas graves que a Psicometria tradicional, baseada na Teoria Clássicas dos Testes (TCT), contém.

Se não a principal, uma das principais limitações da TCT é que as características dos examinandos e as características dos testes não podem ser separadas, sendo que umas só podem ser interpretadas no contexto das outras (Baker, 2001; Fernandez, 1990; Fan, 1998; Hambleton, Swaminathan e Rogers, 1991; Pasquali, 1996).

Baker (2001) define o escore total como a soma dos escores recebidos pelos examinandos nos itens do teste. Sob o enfoque da TCT, os escores totais que os examinandos obtiveram em resposta a uma prova dependem do teste utilizado. Sabe-se, assim, que o desempenho do examinando em um determinado teste pode variar em função da exigência das provas, ou seja, da dificuldade de seus itens. Desta forma, geralmente, quando um teste é difícil, o examinando tenderá a apresentar uma habilidade mais baixa e, quando é mais fácil, tenderá a apresentar uma habilidade mais alta.

Por outro lado, se um item ou uma prova (amostra de itens) é fácil ou difícil, depende da habilidade dos examinandos. Como o cálculo do índice de dificuldade dos itens se dá pelo percentual de examinandos que os acertou, um item é considerado difícil se esse percentual for baixo, e fácil se esse percentual for alto. Se a habilidade da amostra de examinandos, representada pelo escore total, for em média maior que a de uma outra amostra de examinandos, gerarão conjunto de índices de dificuldades diferentes.

Se, pela TCT, os índices de dificuldade dependem da habilidade dos examinados e a habilidade calculada depende da dificuldade dos itens da prova e, verifica-se uma espécie de dependência circular entre eles. O fato das habilidades serem dependentes do conjunto

particular de itens aplicados acarreta que os escores totais advindos da aplicação de duas provas diferentes podem não ser diretamente comparados. Por exemplo, os escores totais do grupo que respondeu ao caderno de prova *x* e os escores totais do grupo que respondeu ao caderno *y* não podem ser diretamente comparáveis se apresentarem dificuldades diferentes.

Uma das implicações práticas dos índices de dificuldade dos itens serem dependentes do grupo é que um mesmo conjunto de itens pode apresentar dois conjuntos diferentes de índices, se estes são calculados para duas amostras diferentes. Na administração de um banco de itens, por exemplo, isso é um problema de difícil solução. Poder-se-ia registrar cada um dos índices vinculado à informação da amostra para a qual foram calculados. Para uma, para duas ou mais amostras. Mas isso não é nada prático. E se quiséssemos montar uma nova prova utilizando itens já aplicados, qual conjunto de índices deveríamos utilizar como informações para tomada de decisão dos itens que a comporiam?

Verifica-se, também, uma outra implicação prática dos índices de dificuldade dos itens serem dependentes do grupo. Ainda no contexto da administração de um banco de itens, se quiséssemos ampliar o banco a partir da inclusão de novos itens, casos seus índices tivessem sido calculados com base nas respostas dos examinandos de outras amostras, a comparabilidade entre eles poderia ser questionada. Essa idéia do presente parágrafo foi apresentada por Hambleton, Swaminathan e Rogers (1991), antes de apresentar as vantagens da TRI.

A TRI, por sua vez, assume a propriedade de invariância dos parâmetros, considerada como a sua maior distinção da TCT. Esse princípio afirma que se podem estimar as habilidades dos sujeitos, independentemente do teste utilizado; bem como os parâmetros dos itens independentemente da amostra de examinandos que os responderam.

Uma das características anunciadas pela TRI é que os parâmetros dos itens não são dependentes do nível de habilidade dos examinandos que os responderam. Baker (2001) mostra um exercício de análise que apresenta dois grupos com habilidades diferentes.

Primeiramente, o autor calcula a proporção de respostas corretas a um item para todos os níveis de habilidade para cada um dos dois grupos. Num segundo momento, ele elabora um gráfico que apresenta as proporções para cada um dos níveis de habilidade do primeiro grupo e outro gráfico do segundo grupo. Posteriormente, utiliza o procedimento denominado *máxima verossimilhança* para ajustar uma curva característica do item (CCI) aos dados e aos valores numéricos dos parâmetros estimados. Como o primeiro grupo é aquele com habilidades baixas, a CCI se concentrou nos valores mais baixos da escala do eixo de *theta*. Como o segundo grupo foi aquele com habilidades altas, a CCI se concentrou nos valores mais altos da escala do eixo de *theta*.

O autor encontrou que, efetivamente como pressupõe a TRI, os parâmetros estimados, *a* e *b*, a partir do grupo com menores habilidades foram idênticos aos parâmetros estimados a partir do grupo com maiores habilidades. Além disso, integrando em um único gráfico as duas curvas, percebeu que elas se complementam e formam uma única curva logística.

Esse exemplo, associado aos resultados de outros quatro exercícios de análise apresentados aos leitores de sua obra, permite a conclusão que a CCI pode ser estimada a partir de qualquer segmento desta curva. Cabe ressaltar que esse exemplo carece de precisão de estimação devido ao pequeno número de examinandos da amostra. Este fornece, inclusive, a impressão de serem dados simulados. Considera-se, no entanto, que a apresentação desses resultados é uma importante forma de ilustração, que permite uma visão clara da propriedade da invariância dos parâmetros dos itens. Os exercícios apresentam indícios para a confirmação da propriedade de invariância dos parâmetros do item, estimados a partir de dois grupos com habilidades diferentes. Segundo Baker, esse estudo empírico “mostra que os valores dos

parâmetros do item são propriedades do item, e não dos grupos que responderam ao item", diferentemente dos resultados encontrados com estudos que envolvam a TCT.

Outro princípio anunciado pela TRI, e, como veremos posteriormente, será um elemento fundamental para o presente trabalho, é que a estimação da habilidade dos examinandos é invariante no que diz respeito aos itens utilizados para determiná-la.

Para ilustrar esse princípio, pode-se considerar a aplicação de duas provas com diferentes dificuldades a um mesmo examinando. Sua habilidade pode ser estimada primeiramente com base em suas respostas à prova *x*, mais fácil, que gera um parâmetro de *theta x*. Posteriormente, com base em suas respostas à prova *y*, mais difícil, que gera um parâmetro *theta y*. A propriedade de invariância dos parâmetros deve propiciar uma mesma estimativa de habilidade para esse examinando em resposta aos dois conjunto de itens (*theta x* = *theta y*).

Isto é possível porque a CCI abarca toda a extensão da escala de habilidades. Assim, independentemente se o item é fácil ou difícil, sempre existirá um ponto da CCI que corresponde à habilidade de interesse, variando apenas a probabilidade de acerto dos examinandos ao item. Em função da invariância dos parâmetros, esses pontos da CCI na escala de habilidades serão sempre os mesmos para itens fáceis ou difíceis. Baker (2001) considera como implicações desse princípio que, um teste composto de itens de qualquer ponto da escala de habilidades pode ser usado para estimar as habilidades dos examinandos, ou seja, a habilidade estimada é invariante em relação ao conjunto de itens utilizados para estimá-la.

Embora esses exemplos ilustrem bem a propriedade de invariância dos parâmetros, Fan (1998) alerta para a escassez de estudos empíricos que busquem verificar essa propriedade, anunciada pela TRI com uma de suas grandes vantagens sobre a TCT. Ele ressalta que "(...) na medida psicológica, como em qualquer área da ciência, modelos teóricos

são importantes para guiar nossas pesquisas e práticas. No entanto, o mérito do modelo teórico deveria, em última instância, ser validado por meio de rigorosas investigações empíricas" (Fan, 1998).

Ele mesmo realizou uma investigação empírica que buscasse respostas (i) do quanto são comparáveis às estatísticas dos itens e dos examinandos geradas a partir da TCT e da TRI, e (ii) do quanto essas estatísticas da TCT e da TRI são invariantes, quando calculadas por meio de amostras diferentes. Utilizando uma base de dados de um programa de avaliação em larga escala, realizou a investigação empírica destas questões. Para tanto, foram utilizadas as respostas de examinandos a dois testes, um de Matemática, composto de 60 itens, e um de leitura, composto de 40 itens, todos com estrutura dicotômica. Um total de 193.000 examinandos respondeu a ambos os testes. Para o estudo do grau de invariância das estatísticas dos itens, foram utilizados três planos amostrais: (a) amostras selecionadas aleatoriamente; (b) amostras de homens e mulheres; e (c) amostras com baixas e altas habilidades.

Esse estudo apresentou, como resultados principais, que (i) as estatísticas dos examinandos pela TCT foram altamente comparáveis com as estimadas pela TRI, (ii) os índices de dificuldades calculados pela TCT foram muito comparáveis com aqueles estimados pela TRI, e (iii) o grau de invariância dos itens, pela TCT, foi altamente comparável com o grau de invariância em relação aos índices estimados pela TRI. Este último achado não confirma a superioridade teórica da TRI, com relação à invariância dos parâmetros dos itens.

Cabe ressaltar que o delineamento utilizado para a realização do estudo citado só permitia a verificação da invariância dos índices dos itens. Verifica-se um delineamento em que todos os examinandos da amostra respondem a uma prova de cada uma das duas áreas, Matemática e Leitura, o que torna inviável a verificação da invariância do parâmetro de habilidade, pelo menos diretamente, sem algum artifício de delineamento.

Para a realização da presente dissertação, foi realizada uma pesquisa bibliográfica que pudesse relatar resultados empíricos de verificação do princípio de invariância do parâmetro de habilidade. Não foram obtidos, no entanto, grandes êxitos neste levantamento. Observou-se que Fan (1998) parecia ter razão no sentido de alertar para a escassez de estudos empíricos na área de invariâncias dos parâmetros. Observou-se também que menor ainda é o número de estudos que buscasse verificar especificamente a invariância do parâmetro *theta*.

Uma das referências encontradas foi um estudo com os dados do Sistema Nacional de Avaliação da Educação Básica (Saeb), realizado por Condé e Rabello (2001). Com os dados de aplicação de 26 formas de provas de Língua Portuguesa do Saeb aplicado em 1997, os autores puderam verificar e comparar o comportamento dos índices de habilidade calculados pela TCT e pela TRI, quando correlacionados com índices de dificuldades. Embora carecesse de um pouco mais de precisão e aprofundamento teórico e metodológico, esse estudo empírico forneceu indícios para a conclusão que os índices de habilidades calculados pela TCT são mais dependentes da dificuldade das provas, que os parâmetros de habilidades estimados pela TRI.

É interessante ressaltar que a realização da presente dissertação foi motivada pelo estudo apresentado no parágrafo anterior (Condé e Rabello, 2001). Este, apesar de poder ter sido um pouco mais aprofundado, mostrou o quanto o delineamento do Saeb é útil para a investigação da propriedade de invariância do parâmetro de habilidade, estimado pela TRI.

Vemos que o delineamento das avaliações determina as possibilidades da investigação do princípio de invariância dos parâmetros. Ora pôde-se avaliar a invariância dos parâmetros dos itens pela TRI, ora pôde-se avaliar a invariância das habilidades pela TRI (Condé e Rabello, 2001).

Fan e Ping (1999), por sua vez, fizeram a verificação de ambas as invariâncias, dos parâmetros dos itens e dos parâmetros dos examinandos, utilizando as mesmas bases de

dados, utilizadas no estudo de Fan (1998). Acontece que esses autores utilizaram o artifício de construir provas fictícias a partir dos dados, considerando os 25% de itens mais fáceis e os 25% de itens mais difíceis de uma prova de 60 itens de Matemática, primeiramente, e de uma prova de leitura, posteriormente. Esse artifício pode ser útil para a investigação da propriedade de invariância dos parâmetros de habilidade da TRI desde que ressaltadas as suas limitações.

Observou-se que um quarto dos itens de Matemática com menores e maiores dificuldades equivalem aos 15 itens mais fáceis e 15 itens mais difíceis. Um quarto dos itens de Leitura com menores e maiores dificuldades equivalem aos 12 itens mais fáceis e 12 itens mais difíceis. Considerou-se que esse número de itens que compôs ambas as provas fictícias foi pequeno e seus resultados podem carecer de precisão.

Além disso, as provas fictícias apresentaram dificuldades médias, calculadas pela TCT (índice denominado por Fan e Ping, em seu estudo de 1999, como *valor p* [1] médio), com uma amplitude muito pequena. Essa diferença entre o *valor p* médio da prova mais difícil pelo *valor p* da prova mais fácil para a prova de Matemática foi de 0,09 e para a prova de Leitura foi de 0,13. Na prática, pode-se considerar essas provas como apresentando dificuldades iguais, o que poderia não discriminar os grupos por essa variável. Esses grupos que responderam a essas provas fictícias tenderiam a apresentar pouca variância de *theta*, mesmo se a propriedade de invariância dos parâmetros de *theta* fosse procedente.

O estudo de Fan e Ping (1999) procurava investigar o efeito do ajuste dos dados aos modelos de um e de três parâmetros da TRI, na invariância dos parâmetros. À luz de todas essas considerações abordadas nos parágrafos anteriores, os resultados da investigação empírica realizada pelos autores não foram conclusivos sobre o potencial efeito negativo do

[1] Para efeitos da presente dissertação, utilizou-se o termo *valor p* para representar a proporção de acertos ao item, ou seja, o índice de dificuldade calculado pela TCT (Nunnally e Bernstein, 1994; Fan e Ping, 1999).

desajuste dos dados aos modelos na propriedade de invariância dos parâmetros de habilidades estimadas pela TRI. Eles sugerem, no final de suas conclusões, que estudos com dados simulados (Monte Carlo) podem ser uma excelente alternativa para investigação do tema.

Tendo sido apresentados os estudos que buscaram verificar empiricamente a invariância dos parâmetros, para efeitos da presente dissertação, são focados aqueles cuja propriedade está relacionada à invariância do parâmetro de habilidade (*theta*).

1.2 A propriedade de invariância do parâmetro de habilidade da TRI

Baker (2001) considera que as habilidades dos examinandos são fixas, além de invariantes, com respeito aos itens usados para medi-las. Argumenta que a habilidade de um examinando é fixa apenas no caso que apresenta um valor particular em um dado contexto. Aborda também que ela deixará de ser fixa se, por exemplo, um processo de intervenção educacional gerasse um aprimoramento de sua habilidade. Assim, se um indivíduo responde a duas provas com dificuldades diferentes, os parâmetros só serão verificados fixos, se supusermos que ele não sofreu uma aprendizagem na resposta a duas provas. Se um processo de aprendizagem estiver envolvido, certamente os parâmetros de habilidade do sujeito, estimados a partir do resultado de duas provas aplicadas uma após a outra, não serão nem fixos e nem invariantes.

Baker (2001) considera que a invariância dos parâmetros depende de duas condições. Uma delas é a necessidade dos valores de todos os parâmetros dos itens estarem em uma métrica comum. Outra condição é a necessidade dos itens da prova estarem medindo um mesmo traço latente. Esta condição está relacionada ao pressuposto de unidimensionalidade dos itens. Isso quer dizer que, se os itens que compõem provas diferentes avaliam um mesmo traço latente, ou seja, são unidimensionais, elas tendem a propiciar estimativas de habilidades

pela TRI sem dependência com a amostra de examinandos que foi utilizada para estimá-la. Uma atenção especial será dada para esta segunda condição, ainda na introdução do presente trabalho, pois a relação entre a unidimensionalidade das provas e a invariância dos parâmetros comporá um dos problemas levantados na presente dissertação.

Tendo em vista as vantagens da independência do parâmetro de habilidade com referência aos parâmetros dos itens, anunciadas pela TRI, o presente estudo tem o objetivo de fornecer contribuições nesta área. Desta forma, a seguinte questão é apresentada: como a teoria pressupõe, o princípio de invariância dos parâmetros da TRI funciona empiricamente? Tendo em vista que essa questão envolve dois aspectos, a saber, (i) se os parâmetros dos itens independem da amostra de sujeitos utilizada para estimá-los e (ii) se o parâmetro de habilidade independe da prova utilizada para estimá-lo, considerou-se como recorte para viabilização do presente estudo a verificação da invariância do parâmetro de habilidade (aspecto ii).

Assim, especifica-se a questão anterior: a propriedade de invariância, no que diz respeito, ao parâmetro de habilidade (*theta*) é procedente quando verificamo-la empiricamente? Ou seja, a habilidade estimada pela TRI é invariante e independe do conjunto de itens utilizados para estimá-la?

Esta questão é de profunda relevância para a obtenção de informações sobre as vantagens da TRI sobre a TCT. Poder-se-ia pensar numa outra questão: é compensatória a utilização da "parafernália" da TRI, com toda a sua complexidade, no que diz respeito ao cálculo da habilidade de examinandos, quando são aplicadas provas distintas entre eles? As pessoas e instituições responsáveis por um determinado sistema de avaliação estarão, efetivamente, obtendo dados mais precisos e tendo em mãos uma ferramenta capaz de calcular essas habilidades, sob um delineamento que exija a aplicação de provas diferentes

para os examinandos? Essas são algumas questões secundárias que podem advir da pergunta principal.

As respostas a essas questões poderão ser analisadas pelo grau de atendimento de uma das condições básicas para que ocorra a invariância, ou seja, a existência de unidimensionalidade entre os itens. Como prevê a teoria, a unidimensionalidade entre os itens é determinante para a invariância dos parâmetros? Essa é uma outra questão, para a qual o presente estudo objetiva encontrar respostas.

Cabe ressaltar que o estudo que está sendo implementado tem por base uma investigação empírica. Pretende-se que os resultados advindos deste forneçam subsídios para um aprofundamento da discussão da teoria da área.

Para estudar a existência da invariância de *theta*, buscou-se encontrar um sistema de avaliação que apresentasse algumas características básicas e necessárias para sua viabilização. Selecionou-se uma avaliação que utilizasse um desenho que envolvesse a aplicação de mais de um modelo de prova. Assim, para a verificação do princípio da TRI, realizou-se um estudo com os dados advindos da aplicação dos modelos de prova de Matemática da 8ª Série do Sistema Nacional de Avaliação da Educação Básica (Saeb), aplicados em 1997, em uma amostra de estudantes brasileiros. Os aspectos envolvidos neste sistema de avaliação serão pormenorizados para melhor entendimento das análises e resultados do presente estudo.

1.3 O Sistema Nacional de Avaliação da Educação Básica (SAEB)[2]

1.3.1 Aspectos gerais do Saeb

"O Instituto Nacional de Estudos e Pesquisas Educacionais (INEP), do Ministério da Educação, vem obtendo informações sobre o desempenho dos alunos brasileiros desde 1991, por meio do Sistema Nacional de Avaliação da Educação Básica (Saeb)" (Pestana, 1999b).

Esse sistema de avaliação em larga escala avalia periodicamente estudantes da 4ª e 8ª séries do ensino fundamental (E.F.) e da 3ª série do ensino médio (E.M.) e tem como principais objetivos: "(a) monitorar a qualidade, a eqüidade e a efetividade do sistema de educação básica; (b) oferecer às administrações públicas de educação, informações técnicas e gerenciais que lhes permitam formular e avaliar programas de melhoria da qualidade do ensino; e (c) proporcionar aos agentes educacionais e à sociedade uma visão clara e concreta dos resultados dos processos de ensino e das condições em que são desenvolvidos e obtidos" (Rabello, 2001).

O Saeb, que foi realizado nos anos 1990, 1993, 1995, 1997, 1999 e 2001, busca avaliar o desempenho de estudantes em diversas disciplinas, a partir da aplicação de provas, e alguns fatores associados a esse desempenho, por meio de questionários contextuais. As disciplinas avaliadas pelo Saeb variaram de aplicação para aplicação. As disciplinas de Ciências da **N**atureza (Química, Física e Biologia) foram avaliadas pelo Saeb 97 e pelo Saeb 99; as disciplinas História e Geografia, pelo Saeb 99; e as disciplinas Língua Portuguesa e Matemática, por meio de todas as aplicações do Saeb.

[2] Procurou-se tratar dos aspectos de interesses principais para nossa investigação. Um maior detalhamento pode ser encontrado em Pestana (1999a), Pestana (1999b), Rabello (2001), Rodrigues (2002), Riether, M.M. e Rauter, R (2000) e Instituto Nacional de Estudos e Pesquisas Educacionais (2002).

A partir de 1995, o Saeb assumiu uma metodologia de elaboração dos testes e de análise de dados baseada na TRI, com o modelo de três parâmetros (Lord, 1980). O desempenho dos estudantes, sob esse enfoque teórico, é estimado por disciplina, conjuntamente, entre as séries e anos de avaliação. Assim, uma série de procedimentos de análise é utilizada de forma que os resultados dos estudantes possam ser colocados em uma mesma métrica e representados em uma mesma escala. A escala única do Saeb varia teoricamente de 0 a 500, sendo que geralmente os resultados de desempenho dos estudantes variam, na prática, de 100 a 400. Uma das vantagens da utilização de uma escala comum entre anos é a possibilidade da criação de uma série histórica que permite o monitoramento da variação desses resultados no decorrer do tempo.

1.3.2 A amostra do Saeb

Os instrumentos de levantamento de dados do Saeb são aplicados em uma amostra de estudantes de todas as Unidades da Federação brasileiras. A pesquisa por amostragem permite que medidas individuais dos estudantes sejam agregadas, de forma que se obtenham estatísticas, a partir das quais são feitas extrapolações para a população à qual essa amostra se refere. A amostra do Saeb é desenhada tendo em vista a avaliação do ensino em três diferentes séries, 4ª e 8ª E.F. e 3ª E.M. É estratificada, levando-se em conta as variáveis de escolas: zona (rural ou urbana), localização (capital ou interior) e rede administrativa (estadual, municipal ou particular). Não fazem parte da população pesquisada a zona rural da Região Norte, as escolas Federais, ou alunos de cursos profissionalizantes do ensino médio e os alunos de turmas multisseriadas no Ensino Fundamental.

A amostra do Saeb é aleatória para cada um dos estratos definidos. Assim a probabilidade de uma determinada escola participar da avaliação é a mesma que a de qualquer

outra. Além da amostra de escolas e de alunos avaliados pelo Saeb, a tabela 1.1 apresenta o número de diretores, funções-docentes[3], séries e disciplinas avaliadas. Optou-se pela apresentação dos resultados a partir do Saeb 95 e, pelo fato dos resultados do Saeb 2001 não terem sido divulgados ainda, não foi apresentada a composição da amostra desse ano.

Tabela 1.1 – Composição da amostra do Saeb 95, 97 e 99.

Participantes	**Ano de Realização do Saeb**		
	1995	**1997**	**1999**
Escolas	2.839	1.933	6.890
Diretores	2.839	1.933	6.890
Funções Docentes	4.967	18.077	53.815
Alunos	90.499	167.196	279.764
Séries avaliadas	4ª, 8ª, 2ª e 3ª	4ª, 8ª e 3ª	4ª, 8ª e 3ª
Disciplinas avaliadas	Matemática Língua Portuguesa	Matemática Língua Portuguesa Ciências (4ª e 8ª) Química, Física e Biologia (3ª)	Matemática Língua Portuguesa Geografia História Ciências (4ª e 8ª) Química, Física e Biologia (3ª)

1.3.3 O instrumento de avaliação do desempenho dos estudantes

Até 1993, o Saeb utilizou provas clássicas de 30 itens para avaliar o desempenho dos estudantes. Sabe-se, no entanto, das limitações desse modelo de instrumento. Dentre essas limitações, pode-se citar que o pequeno número de itens utilizados não permite uma grande abrangência dos conteúdos e competências desenvolvidos e que se espera que os estudantes desenvolvam quando cursam uma determinada série escolar. Uma disciplina como Matemática, por exemplo, envolve temas como *espaço e forma*, *grandezas e medidas*, *número e operações*, *álgebra e funções* e *tratamento da informação* (grandes áreas de conteúdo

[3] A variável *funções-docentes* está relacionada aos professores. É assim tratada pois acontece de um mesmo professor ser computado duas vezes na amostra se for professor de duas disciplinas, salas de aula ou escolas diferentes que foram contemplados pelo Saeb.

considerados por INEP, 2002, nesta disciplina). Como poderíamos avaliar uma grande gama de competências de se trabalhar com esses conteúdos com uma prova de 30 itens? É praticamente impossível. Mesmo se limitássemos o nosso interesse de avaliação apenas parcelas de conteúdos e competências, teríamos um pequeno número de itens para abordar cada um deles, o que acarretaria uma precisão certamente baixa. Ainda teríamos a desvantagem da avaliação não ser representativa dessa abrangência de conteúdos que são tratados (ou pelo menos deveriam ser tratados) em sala de aula em cada uma das séries.

Buscando corrigir limitações como essas geradas pela instrumentação clássica, a partir de 1995 passou a apresentar características bastante peculiares, resumidamente explicitadas a seguir:

(i) **matrizes de referência**. São as tabelas de especificação da avaliação. Em 1995 essas tabelas eram compostas de objetivos de ensino. A partir de 1997 assumiram a estrutura composta por descritores que contemplam os conteúdos e as competências que embasam a avaliação (ver Pestana, 1999a, e Instituto Nacional de Estudos e Pesquisas Educacionais, 2002).

(ii) **provas.** Apresentam um número de itens maior que o apresentado pelas provas tradicionais. As provas do Saeb são compostas por aproximadamente 150 itens por série e disciplina. Essa característica possibilita uma ampla cobertura das competências e conteúdos que se espera que os estudantes tenham desenvolvido e que podem ser representados comportamentalmente pelas respostas aos itens da prova. É claro que não seria viável para um estudante responder a 150 itens em função do tempo e do cansaço. Por isso, para viabilizar a utilização desse grande número de itens, o Saeb incorpora uma metodologia baseada na amostragem matricial de itens, que utiliza o esquema de montagem e aplicação de provas por Blocos Incompletos Balanceados (BIB).

Sob esse delineamento, são montados primeiramente 13 blocos de itens que podem variar de tamanho. Desde 1999, o Saeb vem utilizando um número de 13 itens por bloco, mas já utilizou tamanhos que variavam de 10 a 13 itens por bloco. São montados 26 cadernos a partir da combinação, três a três, desses blocos de itens por meio da orientação fornecida pela matriz do BIB, apresentada na tabela 1.2.

Tabela 1.2 – Delineamento de Blocos Incompletos Balanceados (BIB).

Caderno	Primeiro Bloco	Segundo Bloco	Terceiro Bloco	Caderno	Primeiro Bloco	Segundo Bloco	Terceiro Bloco
1	1	2	5	**14**	1	3	8
2	2	3	6	**15**	2	4	9
3	3	4	7	**16**	3	5	10
4	4	5	8	**17**	4	6	11
5	5	6	9	**18**	5	7	12
6	6	7	10	**19**	6	8	13
7	7	8	11	**20**	7	9	1
8	8	9	12	**21**	8	10	2
9	9	10	13	**22**	9	11	3
10	10	11	1	**23**	10	12	4
11	11	12	2	**24**	11	13	5
12	12	13	3	**25**	12	1	6
13	13	1	4	**26**	13	2	7

Essa distribuição de itens por blocos e de blocos por cadernos permite, dentre outros aspectos, que um mesmo conjunto de itens esteja localizado na primeira posição (primeiro bloco) em dois cadernos de prova, na segunda posição, em outros dois cadernos e na terceira posição, em outros dois. Por exemplo, o bloco 1 está localizado na primeira posição nos cadernos 1 e 14; na segunda posição nos cadernos 13 e 25; e na terceira posição nos cadernos 10 e 20.

(iii) **aplicação da provas pela amostra**. Cada aluno recebe uma forma de prova, ou caderno de uma das disciplinas. Como podemos observar na tabela 1.1, no Saeb 97, por exemplo, em que foram avaliadas as disciplinas Matemática, Língua Portuguesa, Ciências (Química, Física e Biologia, para a 3ª série do ensino médio). O aluno que, no dia da prova,

está sentado na primeira carteira recebe o caderno 1 da disciplina Matemática. O aluno que está sentado na segunda carteira recebe o caderno 1 de Língua Portuguesa. O terceiro recebe o caderno 1 de Ciências e, assim, sucessivamente para as outras disciplinas, se for o caso. Os próximos recebem o caderno 2 de cada uma das disciplinas, o caderno 3 e, assim, sucessivamente.

A seqüência de aplicação dos cadernos de provas foi mantida para uma próxima sala e, seqüencialmente, para uma próxima. Desta forma, caso o último estudante que respondeu à prova de Matemática tivesse recebido o caderno número 15, por exemplo, o primeiro estudante da outra sala, que responde ao caderno de Matemática, recebe o caderno 16. Esse esquema de aplicação foi previsto para dentro da escola, entre as escolas de um determinado município, entre os municípios e, num nível mais amplo, para a unidade federativa brasileira.

1.3.4 A equivalência dos os grupos de estudantes que respondem aos cadernos

Com o esquema amostral e o delineamento BIB adotado, consegue-se uma aplicação em que um número aproximado de estudantes, dos mais diversos estratos da amostra, responda a cada um dos cadernos. Também permite que os estudantes que respondem a um determinado caderno apresentem, proporcionalmente, características semelhantes aos grupos que responderam aos outros cadernos, visto que a alocação dos cadernos aos alunos é aleatória. Ou seja, garante que todos os grupos apresentem, por exemplo, as mesmas proporções de estudantes de baixa habilidade, de média ou de alta habilidade; de classes sociais menos ou mais favorecidas; com culturas e etnias diversas. Em outras palavras, os grupos de estudantes que respondem a cada um dos cadernos de prova do Saeb podem ser considerados equivalentes.

1.4 A verificação da invariância do parâmetro de habilidade da TRI a partir dos dados do Saeb

O trabalho com análises pela TRI e um delineamento no qual formas de provas diferentes são aplicadas a grupos de estudantes, teoricamente, com características semelhantes, faz dos resultados do Saeb um excelente material de pesquisa da invariância do *theta*, em relação à dificuldade das provas aplicadas. Assim, para atingirmos os objetivos e obtermos respostas para as questões levantadas neste estudo, trabalhou-se então com as respostas de estudantes aos vinte e seis cadernos de prova de Matemática da 8ª Série aplicados no Saeb 97.

Hipotetiza-se que o *theta* representativo dos grupos que responderam a cada um dos cadernos não apresentem diferenças, pois pela propriedade de invariância dos parâmetros, esses independem dos itens que foram usados para estimá-los. Praticamente, os *theta* estimados para os 26 grupos de estudantes diferentes que responderam aos cadernos, para que não se rejeite esta hipótese nula, deverão ser bastante próximos, mesmo que haja diferenças entre as dificuldades entre esses cadernos de provas.

Para a obtenção de respostas à questão da unidimensionalidade como condição para verificação da invariância do parâmetro de *theta*, serão usados os mesmos dados de aplicação da prova de Matemática da 8ª Série do Saeb 97. Caso exista alguma variação entre os *theta* estimados dos grupos de estudantes que responderam a cada um dos cadernos, o pressuposto da unidimensionalidade entre os índices desta prova será analisado. Apresenta-se a hipótese que, quanto maior é o grau de unidimensionalidade entre os itens que compõem a prova, menor é a dependência do *theta* com relação à dificuldade dos cadernos.

Para o estudo da condição de unidimensionalidade entre os itens para a invariância dos parâmetros, será apresentado um aprofundamento teórico e uma investigação empírica, apresentada a seguir.

1.5 O pressuposto de unidimensionalidade dos itens

Unidimensionalidade é um pressuposto da TRI em que apenas uma habilidade é medida por um conjunto de itens em um teste. Ela está relacionada á idéia da existência de um único traço latente subjacente ao conjunto de itens. De maneira mais prática, considera-se uma prova unidimensional se esta apresenta um componente ou fator dominante que influencia o desempenho dos examinandos em um teste.

Para a estimação dos parâmetros dos itens e das habilidades pela TRI, a verificação da unidimensionalidade da prova utilizada se torna fundamental. Laros, Pasquali & Rodrigues (2000) apresentaram quatro efeitos negativos que podem surgir quando é violado o pressuposto da unidimensionalidade dos itens na utilização da TRI.

O primeiro efeito negativo é que a ausência de unidimensionalidade de um conjunto de itens conduz à diminuição da validade de construto do teste, dificultando a interpretação dos escores. Esse aspecto coloca a validade da prova em questão.

O segundo aspecto é a função diferencial do item que surge para grupos de diferentes culturas, por exemplo. Esse viés está associado à validade de construto. “Se num teste falta validade de construto, o teste conterá itens que estarão medindo outras habilidades que não aquelas que se propôs medir e, portanto, o potencial para viés do item também existe” (Laros, Pasquali & Rodrigues, 2000).

O terceiro efeito trata do efeito negativo da violação do pressuposto para a equalização dos resultados de várias formas de uma prova, o que a torna impossível de ser realizada mesmo para modelos multidimensionais da TRI.

O quarto efeito está relacionado à estimação da proficiência do aluno. Quando se quer avaliar a habilidade de examinandos em tópicos amostrados de um domínio conceitual e

unidimensional, por exemplo, a "probabilidade de *theta*, dado o padrão de resposta, não é válida e as estimativas e os desvios-padrão de *theta* podem ser errôneos" (Laros, Pasquali & Rodrigues, 2000).

Quando estes quatro efeitos da violação da unidimensionalidade são analisados, verifica-se que eles são todos inter-relacionados. Como a estimação do *theta* é feita (ou pode ser feita) conjuntamente à equalização, a falta de unidimensionalidade pode enviesar esse parâmetro. E ele estará representando que traço latente? Não se sabe, caso os itens da prova não estejam avaliando um único fator. Além disso, como vimos, a falta de validade de construto pode tornar os itens enviesados para examinandos com mesma habilidade (questão da função diferencial do item).

Laros, Pasquali & Rodrigues (2000) realizaram uma revisão da literatura psicométrica e relataram cinco índices para determinar a unidimensionalidade de um conjunto de itens. "São eles (1) índices baseados em padrões de resposta; (2) índices baseados na fidedignidade; (3) índices baseados na análise de componentes principais; (4) índices baseados na análise fatorial e (5) índices baseados na TRI". Os autores chegaram à conclusão, em acordo com Hattie (1985), que os índices baseados na TRI são os mais adequados para a avaliação da unidimensionalidade.

1.6 A unidimensionalidade da prova de 8ª série de Matemática, do Saeb 97: um estudo empírico

Esses autores realizaram também um estudo empírico para verificação da dimensionalidade das provas do Sistema Nacional de Avaliação da Educação Básica (Saeb) aplicadas em 1997. Seu trabalho tinha como objetivo avaliar se as provas do Saeb 97 eram unidimensionais.

Utilizaram, para tal feito, o método de análise fatorial *full information,* baseado na TRI, um dos índices considerados por Hattie (1985) como um dos mais adequados para verificação da unidimensionalidade de um conjunto de itens dicotômicos como o Saeb. Utilizou também os índices complementares porcentagem de variância explicada pelo primeiro fator, a correlação bisserial item-total e a correlação tetracórica entre os itens.

O referido estudo envolveu todas as disciplinas avaliadas pelo Saeb. Para efeitos da presente dissertação, serão apresentados os resultados da verificação da unidimensionalidade da prova de Matemática da 8ª Série do Saeb 97.

Esta prova foi composta de 161 itens. Um item foi anulado. Assim foram considerados, para a avaliação da unidimensionalidade da prova de Matemática, 160 itens. Primeiramente, foram calculados os valores de qui-quadrado para os modelos de um e dois fatores com seus respectivos graus de liberdade, as diferenças para os modelos de um e dois fatores, as diferenças dos valores de qui-quadrado corrigidas e os índices de unidimensionalidade. Esses resultados estão apresentados a seguir, na tabela 1.3. O qui-quadrado é um índice de falta de ajuste dos dados ao modelo. Se o modelo de dois fatores exibe um qui-quadrado maior do que o modelo com um fator, este se ajusta melhor aos dados do que o modelo de 2 fatores.

Tabela 1.3. - Informações para verificação do número de fatores das provas de Matemática da 8ª série do Saeb 97.

Modelo	**1 Fator**	**2 Fatores**
Qui-quadrado	366.403	366.440
Graus de Liberdade	18.473	18.314
Diferença no qui-quadrado		-37,0
Diferença corrigida no qui-quadrado		-12,3
Graus de Liberdade da diferença		159
Índice de unidimensionalidade		-0,07

Os resultados do estudo de Laros, Pasquali & Rodrigues (2000) para a prova de Matemática indicaram que o modelo de dois fatores exibe um qui-quadrado maior do que o

modelo com um fator. Dessa forma, o modelo de um fator se ajustou melhor que o de dois fatores, ou seja, a prova, como um todo, apresenta unidimensionalidade.

Além disso, o índice de unidimensionalidade, que avalia as vantagens da utilização do modelo de dois fatores, foi menor que 2. O valor encontrado neste teste estatístico foi de -0,07, ou seja, não haveria melhoria significativa se considerassem dois fatores para a prova.

No entanto, nem todos os itens contribuem na mesma quantidade para a unidimensionalidade da prova. Foram encontrados, do conjunto total de itens da prova, 26 itens (16% dos itens avaliados) com cargas fatoriais inferiores a 0,20 no primeiro e único fator. Os autores sugeriram a exclusão destes itens, que praticamente não contribuem para a unidimensionalidade. Depois da exclusão desses itens, a prova de Matemática pode ser considerada unidimensional e pode ser analisada pela Teoria de Resposta ao Item, sem a violação do seu pressuposto principal.

Após a explicitação dos aspectos teóricos e de alguns achados empíricos sobre a Teoria de Resposta ao Item, invariância dos parâmetros e unidimensionalidade e após a abordagem dos aspectos principais do Saeb, pode-se partir para o cumprimento dos objetivos do presente estudo.

Assim, partiu-se para a investigação da invariância do parâmetro *theta* estimado a partir de provas, que se espera que apresentem dificuldades diferentes, de grupos, que se espera que sejam equivalentes. É claro que estas verificações são fundamentais para o trabalho e não é nenhuma surpresa adiantarmos que, na comparação das dificuldades médias e na verificação da equivalência entre os grupos, verificou-se que as dificuldades dos cadernos eram diferentes e os grupos eram equivalentes.

No capítulo 2, será apresentado o método utilizado para fornecer respostas às questões: pode-se verificar independência ou invariância do *theta* em função dos itens que foram utilizados para estimá-lo? É procedente a afirmação de vantagens da TRI sobre a TCT,

no que diz respeito à dependência da habilidade com a prova? Quando a prova se aproxima mais da unidimensionalidade, a dependência da estimação do *theta* com os itens que foram utilizados para estimá-lo é menor?

2. Metodologia

2.1 Participantes

A prova de 8ª Série de Matemática do SAEB 1997 foi respondida por uma amostra de 18.806 estudantes da rede pública e particular. Essa amostra foi delineada para produzir resultados de desempenho representativo para as 27 unidades da federação e, dentro delas, para subpopulações de interesse.

Para tanto, foi definida uma população de referência, que foi estratificada em diversas subpopulações ou estratos. "A amostra do Saeb 97 foi estratificada levando-se em conta as variáveis de escolas: zona (rural ou urbana), localização (capital ou interior) e rede administrativa (estadual, municipal e particular)" (Instituto Nacional de Estudos e Pesquisas Educacionais, 1998). Dentro dos estratos, houve um sorteio dos elementos que participaram do Saeb. Esse modelo de amostragem nos permite dizer que o grupo de alunos estudado foi representativo da população de alunos de 8ª Série E.F. do Brasil.

Um maior detalhamento da amostra já foi apresentado no tópico 1.3.2 desta dissertação.

2.2 Instrumento

Os 161 itens da prova foram construídos com base em uma matriz de referência (Pestana, 1997) de conteúdos e competências, validada em nível nacional em termos do currículo efetivo, com base no que estava sendo ensinado aos estudantes. Foram construídos itens de 4ª e 8ª Séries de Matemática[1], exclusivamente de múltipla escolha (com quatro e cinco alternativas, sendo apenas uma a correta).

[1] O motivo de estar-se levando em conta a elaboração de itens de 4ª Série será abordado posteriormente.

O item foi composto de estímulo, enunciado e alternativas. O estímulo serve como elemento que auxilia o examinando na resposta ao item e, normalmente, apareceu em formato de gráficos, tabelas e ilustrações. O enunciado apresentava o problema a ser resolvido e as alternativas eram as opções de escolha.

Um conjunto de itens foi construído e validado pedagogicamente em oficinas, respectivamente, de elaboração e revisão de itens, compostas por especialistas na disciplina de Matemática, normalmente professores das séries em questão, com experiência em elaboração e revisão de itens. Os itens foram construídos com base em uma série de normas técnicas (Instituto Nacional de Estudos e Pesquisas Educacionais, 2001). Dentre elas, cada item foi elaborado de forma a avaliar um único descritor da Matriz de Referência, a apresentar apenas um problema, a não exigir do aluno um tempo de leitura excessivo e cuja linguagem fosse acessível aos alunos.

As mesmas normas foram utilizadas como critério de revisão ou validação teórica dos itens posteriormente. Nesta, eles eram submetidos a uma revisão técnica, lingüística e pedagógica. Na primeira, eram avaliados com referência ao atendimento das normas de construção, estrutura e editoração. Na revisão lingüística, procurou-se verificar se a linguagem que estava sendo utilizada estava de acordo com a norma culta da Língua Portuguesa e se era de leitura acessível aos alunos da série. Já na revisão pedagógica, avaliou-se se o item efetivamente avaliava o domínio curricular a que ele se referia. Como produto da validação teórica, os itens foram classificados em aceitos integralmente, aceitos com reformulação (e então reformulados) e rejeitados.

Ainda no processo de validação, os itens aceitos (com ou sem reformulações) foram submetidos a uma pré-testagem em nível nacional. A análise dos dados, pela TCT, indicou as características estatísticas dos itens. Para a composição da prova, priorizaram-se aqueles itens com correlação com o escore total (coeficiente bisserial) maior ou igual a 0,20.

Foram escolhidos 161 itens para a composição da prova de 8ª Série de Matemática do Saeb 97. Deste total, 44 itens foram provenientes da prova de Matemática da 4ª Série do Saeb 97, 11 itens, da 4ª Série do Saeb 95 e 22 itens da 8ª Série do Saeb 95, sendo que, deste quantitativo, alguns deles foram utilizados em mais de uma destas avaliações (ver tabela 2.1). Eles foram inseridos na prova de 8ª com a função de serem utilizados como base para a equalização dos resultados das provas entre estas séries e anos.

Do conjunto total de itens selecionados, foram montados 13 blocos compostos de 11 ou 13 itens. Houve uma preocupação, da mesma forma, em montar blocos que tivessem itens com referência a uma variedade de descritores da Matriz de Referência e itens de complexidades variadas. Quando se tratava de blocos compostos de itens de 4ª Série, buscou-se selecionar, de preferência, blocos inteiros mais difíceis da série original, com base nas estatísticas do pré-teste e do Saeb 95, na tentativa de viabilizar uma aproximação da dificuldade dos blocos da prova de Matemática de 8ª Série do Saeb 97.

Tabela 2.1 – Número de itens dos blocos da prova de 8ª Série de Matemática do Saeb.

Bloco	**Número de itens**			
	Total do bloco	**Comuns à 4ª Série do Saeb 97**	**Comuns à 4ª Série do Saeb 95**	**Comuns à 8ª Série do Saeb 95**
1	11	11	0	0
2	11	11	0	0
3	11	11	0	0
4	11	11	11	11
5	13	0	0	0
6	13	0	0	0
7	13	0	0	0
8	13	0	0	0
9	13	0	0	0
10	13	0	4	11
11	13	0	0	0
12	13	0	0	0
13	13	0	0	0
Total	161	44	15	22

Optou-se pela utilização do delineamento por BIB para a montagem dos cadernos de prova e posterior esquematização da aplicação. Esse delineamento permite uma ampla cobertura da Matriz de Referência, pois é possível a utilização de um número grande de itens (no caso 161), sem exigir do aluno que responda a um número excessivo de itens. Desta forma, foram compostos 26 cadernos, a partir da combinação de blocos, três a três. A tabela 2.2 apresenta o número de itens que compuseram cada um dos cadernos. Pode-se notar que os cadernos apresentaram um número mínimo de 35 e máximo de 39 itens.

Tabela 2.2 - Número de itens por caderno e por bloco de 8a Série de Matemática.

Caderno	**Número de itens**			
	Primeiro Bloco	**Segundo Bloco**	**Terceiro Bloco**	**Total do caderno**
1	11	11	13	35
2	11	11	13	35
3	11	11	13	35
4	11	13	13	37
5	13	13	13	39
6	13	13	13	39
7	13	13	13	39
8	13	13	13	39
9	13	13	13	39
10	13	13	11	37
11	13	13	11	37
12	13	13	11	37
13	13	11	11	35
14	11	11	13	35
15	11	11	13	35
16	11	13	13	37
17	11	13	13	37
18	13	13	13	39
19	13	13	13	39
20	13	13	11	37
21	13	13	11	37
22	13	13	11	37
23	13	13	11	37
24	13	13	13	39
25	13	11	13	37
26	13	11	13	37

* A tabela 1.2, que apresenta o BIB, é referência para a composição desta tabela.

2.3 Procedimentos

Garantida a devida padronização nos procedimentos de aplicação, cada aluno respondeu a um único caderno de provas de Matemática. Estes cadernos foram distribuídos seqüencialmente em termos dos números de cadernos. O primeiro aluno, que respondeu a prova de Matemática, recebeu o caderno 1, o segundo aluno que respondeu à prova de Matemática, recebeu o caderno 2, e assim sucessivamente. Os cadernos desta disciplina foram distribuídos alternadamente com os cadernos das outras disciplinas aplicadas no Saeb, como já foi mencionado na introdução do presente trabalho. Desta forma, apenas parte da turma respondeu à prova de Matemática.

A prova foi aplicada por pessoal contratado que utilizou cerca de uma semana para cobrir toda a amostra. Anteriormente à aplicação da prova, foi aplicado um questionário sócio-demográfico.

O tempo de aplicação da prova foi de 75 minutos, divididos em três períodos de 25 minutos, um para resposta a cada bloco. Os alunos de uma determinada sala iniciavam ao mesmo tempo o preenchimento de cada bloco de itens. Esse procedimento, juntamente com o delineamento de distribuição de blocos de itens pelos cadernos, permitiu que os dados ausentes fossem distribuídos, até certo ponto, igualmente pelos blocos respondidos, não acarretando em uma perda significativa apenas das respostas aos últimos itens dos cadernos.

Cada estudante recebeu um caderno de prova e uma folha de leitura óptica, em que marcavam as respostas. Foi fornecido, ao final da aplicação, um tempo extra para os estudantes terminarem de preencher a folha de leitura óptica. Os dados coletados foram então lidos oticamente e um banco de dados com as respostas dos estudantes foi estruturado. O banco com as respostas dos estudantes às provas continha um campo para registro do código identificador, das respostas (A, B, C, D ou E) para cada um dos itens dos cadernos e do peso

amostral. Os dados ausentes foram diferenciados por caracteres distintos quando os alunos deixavam questões em branco no meio dos blocos, quando deixavam em branco no final dos blocos ou quando marcavam mais de uma alternativa. Essa diferenciação teve impacto, posteriormente, nas análises.

Foi realizada uma análise da base de dados e excluídos os estudantes que não responderam à prova ou que responderam incorretamente às folhas de respostas. O número de integrantes da amostra efetiva foi extraído após essa depuração.

Os dados foram analisados primeiramente pela TCT para os quais foram calculadas as estatísticas dos itens. Para cada um dos itens, calculou-se (i) o percentual de acerto, que expressa o índice de dificuldade (*valor p*), (ii) o percentual de estudantes que optou por cada uma das alternativas, (iii) a diferença do percentual de acerto dos 27% dos alunos com melhor desempenho pelo percentual de acerto dos 27% dos alunos com pior desempenho na prova (índice de discriminação), e (iv) o coeficiente de correlação item-total, especificamente o coeficiente de correlação bisserial para todas as alternativas (alternativa correta e distratores). Para cálculo dessas estatísticas, foi utilizado um software elaborado pela empresa contratada para análise de dados (Klein & Klein, 1998).

Os coeficientes de correlação bisserial foram utilizados para a definição permanência ou não de cada um dos itens nas próximas fases de análise. Esperavam-se bisseriais positivos e altos na alternativa correta e negativos nos distratores (alternativas incorretas). Foram mantidos, para a estimação dos parâmetros dos itens e habilidades pela TRI, apenas os itens que apresentaram bisserial maior que 0,20. Esse procedimento é justificado, pois aqueles itens que não apresentaram boa qualidade pela TCT poderiam prejudicar estimação dos parâmetros pela TRI. Utilizou-se um critério até certo ponto leniente para que não se perdesse uma grande quantidade de itens.

Posteriormente, os dados foram analisados pela TRI, para os quais foram estimados os parâmetros para cada um dos 161 itens da prova (discriminação – *parâmetro a*, dificuldade – *parâmetro b*, e acerto ao acaso – *parâmetro c*) e as habilidades dos estudantes (*theta*). Para tanto, foi utilizado o *software* BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). Tanto os parâmetros dos itens, quanto as habilidades, foram estimados conjuntamente a partir da utilização deste *software* que implementa a Teoria dos Grupos Múltiplos (Bock e Zimowski, 1995). Essa teoria permite a estimação conjunta dos parâmetros várias amostras não equivalentes.

Os parâmetros dos itens da prova e das habilidades foram estimados conjuntamente aos dos parâmetros das provas de Matemática de 4ª e 8ª Séries E.F. e 3ª Série do E.M. dos Saebs 95 e 97. Essas estimativas, portanto, se encontram na mesma escala entre séries e anos do Saeb. Os parâmetros de dificuldade estimados pela TRI e o parâmetro de habilidade *theta* são registrados em uma escala que varia, geralmente, de –3 a 3 e que apresenta média 0, com desvio-padrão de 1. Para efeito de divulgação de seus resultados do Saeb, eles são transformados para uma escala que varia, geralmente, de 0 a 500. Ressalta-se que, para efeitos do presente trabalho, os resultados serão sempre utilizados na escala original.

Para investigar uma possível influência da dificuldade dos cadernos de provas na estimação das habilidades dos examinandos pela TRI, então, tinham-se disponíveis resultados de 26 grupos de estudantes diferentes, fruto da aplicação de 26 cadernos diferentes. Se os grupos que responderam aos cadernos apresentassem características semelhantes, esperar-se-ia que as estimativas de habilidade desses grupos, pela TRI, fossem também semelhantes, mesmo que estes tenham sido submetidos a provas com dificuldades diferentes, como prevê a propriedade de invariância dos parâmetros.

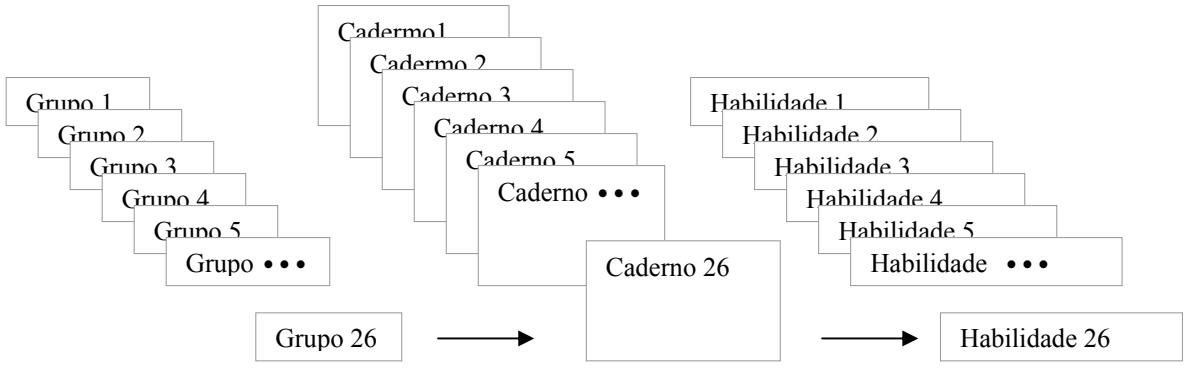
Duas condições foram fundamentais para a viabilização do estudo. Em primeiro lugar, foi fundamental, para nossa investigação, que os grupos que responderam aos cadernos de

provas apresentassem características iguais em termos de habilidades, ou seja, que fossem equivalentes. Em segundo lugar, foi fundamental que os cadernos de provas apresentassem variabilidade em suas dificuldades, ou seja, que fossem de diferentes dificuldades médias.

2.3.1 Estudo da equivalência entre os grupos de estudantes

Para se conseguir estudar o impacto das dificuldades dos cadernos na estimativa de habilidade estimada pela TRI, foi necessária uma etapa de estudos sobre algumas condições de comparabilidade desses grupos de estudantes que responderam a cada uma das formas de prova (caderno 1 a 26). Não eram os mesmos examinandos que estavam respondendo a esses 26 cadernos. No entanto, para os interesses do delineamento do presente estudo, é fundamental que todos os grupo apresentem características iguais em termos de habilidades. O diagrama a seguir ilustra essa nuance do delineamento.

Figura 2.1 – Delineamento em que grupos de examinandos respondem a diferentes cadernos de prova e apresentam resultados específicos em termos de habilidades.



Onde as habilidades dos grupos de examinandos que responderam a cada um dos cadernos sejam iguais. Ou seja,

Habilidade 1 = Habilidade 2 = Habilidade 3 = • • • = Habilidade 26

Vimos no tópico 1.3 da introdução que a utilização de uma estrutura amostral aleatória (dentro dos estratos) e do delineamento de montagem e aplicação de provas por BIB fazem com que os grupos de estudantes que responderam a cada um dos cadernos sejam equivalentes. Apenas a análise dessas características amostrais, de montagem e distribuição das provas pela amostra já seriam suficientes para considerarmos a equivalência entre os grupos.

No entanto, procurou-se realizar alguns outros estudos para confirmação dessa equivalência. Procurou-se verificá-la por meio de análises estatísticas, investigando se os grupos que responderam a cada caderno apresentavam características semelhantes em termos de habilidades. Para a verificação estatística dessas semelhanças de características de habilidades dos estudantes de cada grupo, utilizou-se um procedimento que compara o desempenho de grupos diferentes de estudantes que responderam a um mesmo bloco de itens, em termos de escores totais.

Sabendo-se que cada um dos blocos aparece em seis cadernos diferentes (duas na primeira posição, duas na segunda e duas na terceira, como pode ser visto na tabela 1.2) e, portanto, são respondidos por seis grupos diferentes, pôde-se comparar o desempenho desses estudantes em termos da diferença de escores totais médios nos blocos e distribuição de freqüências de escores totais nos blocos. A partir desses procedimentos de análise, considerou-se cada bloco, como uma sub-prova, com os mesmos itens respondidos por grupos distintos de estudantes.

Foram obtidos, desta forma, seis escores totais médios e seus respectivos desvios-padrão, referentes aos seis grupos de examinandos que responderam a cada um dos 13 blocos. Caso os grupos de estudantes que responderam a cada um dos blocos sejam equivalentes, ou seja, caso a média e a distribuição de freqüências dos escores totais desses grupos sejam bem

semelhantes, poder-se-ia considerar que os grupos de estudantes que responderam a cada um dos cadernos também são equivalentes.

Para verificação do que significaria essas diferenças entre médias em termos de desvio padrão, na tentativa de verificar a variabilidade dessas estimativas para cada bloco, foram realizados procedimentos de normalização semelhantes ao índice *d* de Cohen ou *Cohen's d* (Shaughnessy, J.J., Zechmeister E.B., & Zechmeister, J.S., 2000). Por meio desta, divide-se a diferença entre as médias pelo desvio padrão dos escores totais. Assim,

$$d = (X_1 - X_2) / \sigma$$

O índice *d* de Cohen considera σ (o desvio padrão) como sendo a raiz quadrada da expressão que representa a variância

$$[(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2] / N$$

em que

n_1 = tamanho da amostra 1
n_2 = tamanho da amostra 2
s_1^2 = variância do grupo 1
s_2^2 = variância do grupo 2
$N = n_1 + n_2$

Para o presente trabalho, foram utilizados procedimentos de normalização que divide a diferença entre as médias pela média dos desvios-padrão dos escores totais. O σ do índice *d* de Cohen foi aqui considerado como

$$(\sigma_1 - \sigma_2) / 2$$

Com essa normalização, pôde-se verificar o quanto essas médias dos escores totais dos estudantes em resposta a um determinado bloco se diferenciam em termos de desvios-padrão, de forma a se avaliar essas diferenças. Pode-se utilizar a seguinte classificação como critérios de interpretação dos resultados normalizados pelo índice *d* de Cohen

Diferença pequena: $d < 0,50$
Diferença média: *d* varia de 0,50 a 0,80
Diferença grande: $d > 0,80$

Num primeiro momento, verificou-se o desempenho dos seis grupos (dois a dois) que respondiam aos blocos, quando estes eram apresentados em uma determinada posição do caderno. Das seis médias obtidas para cada bloco, analisou-se, num segundo momento, o desempenho dos grupos independentemente da posição em que esses blocos estavam situados nos cadernos. Essas análises forneceram estatísticas descritivas que ajudaram a conhecer aspectos básicos sobre os escores dos grupos.

Para aprofundar ainda mais a investigação da equivalência entre os grupos, além das análises descritivas com os resultados de escores por bloco, foram realizados testes de estatísticas inferenciais para verificação da existência de diferenças significativas entre as características dos grupos. Definiu-se a hipótese nula (H_o) de dois determinados grupos que responderam ao mesmo bloco na mesma posição do caderno, mas em cadernos diferentes, apresentarem a mesma distribuição de freqüências de escores totais.

Para testar a hipótese, utilizou-se um teste não-paramétrico denominado prova *U* de *Mann-Witney* (Siegel, 1975; Kvanli, Guynes & Pavur, 1991; e SPSS, 1999). Esse teste estatístico, que trabalha a partir de combinações das observações ou escores, relacionado-os por ordem ascendente, foi uma alternativa à opção de não utilizar o teste t pelo SPSS (*Statistical Package for the Social Sciences*). Este pacote estatístico não considera

adequadamente os pesos de amostras complexas como a do Saeb. O teste t exige um estimação precisa do erro-padrão e este software acaba por subestimá-lo. As médias de escores totais são consideradas significativamente diferentes entre blocos, mesmo que essas na realidade não fossem. Já o teste U, como trabalha por postos, não exige a consideração do erro-padrão, motivo da adoção de sua utilização no presente estudo. Este teste pode ser utilizado quando o número de observações dos grupos é diferente (amostras de tamanhos diferentes) e quando se quer realizar uma prova bilateral. É considerado eficiente para análises de grandes amostras. Como pressupostos, esse teste exige que (i) as amostras sejam aleatórias, (ii) as observações não sejam pareadas, e (iii) os dados sejam no mínimo ordinais. Todos os pressupostos foram atendidos, o que permitiu a utilização do referido teste estatístico.

Brogan (1997) fez alguns estudos com o SAS (SAS Institute Inc., 1993), um software de características similares às do SPSS. Ele constatou que este software subestima os erros-padrão, apontando para uma falsa rejeição da hipótese nula. Ele encontrou que o software SUDAAN seria uma solução para a estimação "justa" dos erros-padrão.

Brogan (1997) observou também que, utilizando os pesos amostrais, o SAS considera o tamanho da amostra como sendo a soma dos valores desses pesos, o que resulta em um valor muito alto, impossibilitando a realização apropriada do teste. Os resultados da utilização dos testes tendem sempre a ser significativos. A alternativa a esse problema foi utilizar pesos amostrais normalizados.

Em seu artigo, Brogan (1997) utiliza como uma de suas variáveis os pesos amostrais normalizados para o tamanho da amostra. Segundo ele esse procedimento vem sendo utilizado por pesquisadores, que consiste que a soma dos pesos seja igual ao número de ocorrências da amostra, utilizando valores de pesos individuais substancialmente menores. "Essa utilização é recomendada por alguns analistas de dados como capaz de fornecer resultados

aproximadamente iguais aos fornecidos por *softwares* especializados em amostragem tipo '*survey*'" (Brogan, 1997), como o SUDAAN (Shah, 1996).

Seguindo esse procedimento como uma opção de análise, foi realizada a normalização dos pesos amostrais originais (que chamaremos de *peso_o*). Para um examinando j, o valor de cada do peso amostral normalizado (que chamaremos de *peso_n*) é

$$peso\_n(j) = (18.806) \bullet [peso\_o(j) / 2.512.018]$$

Onde, peso_n(j) é o peso normalizado do examinando j e o peso_o(j), o peso original do Examinando j. Em que o valor 2.512.018 é a soma dos pesos originais (peso_o) e a estimativa total dos estudantes; e o valor 18.806 é o número de estudantes da amostra.

A partir desta transformação, os pesos amostrais utilizados para a realização da prova *U* de *Mann-Whitney* foram os pesos normalizados (peso_n). Nesta dissertação, para todas as análises que exigiam a utilização dos pesos amostrais, como as que utilizam as bases de examinandos, foram também utilizados os peso_n.

2.3.2 Estudo da diferença entre as dificuldades dos cadernos de provas e levantamento dos índices de habilidades dos grupos que os responderam

Também como condição para a continuação do presente estudo, foi necessário verificar se os cadernos apresentam dificuldades diferentes. Para tanto, utilizou-se como índices de verificação das dificuldades dos cadernos de prova (i) a estatística de dificuldade pela TCT (que estamos chamando de *valor p* no presente trabalho) que é o percentual de acertos do grupo que respondeu a cada um dos cadernos, e (ii) o parâmetro de dificuldade da TRI (representado pelo *parâmetro b*). Foram calculados os índices *valor p* e *parâmetro b* médios dos cadernos, considerando os pesos amostrais, o que gerou 26 índices de dificuldade

para cada um dos modelos teóricos (TCT e TRI). Como essa condição era fundamental para a continuação do presente estudo, já se pode adiantar que os cadernos apresentavam, geralmente, dificuldades diferentes.

Também foram calculadas as estimativas de habilidade dos grupos de examinandos que responderam a cada um dos cadernos. Utilizou-se, para tanto, (i) o *escore total* da TCT, e (ii) o parâmetro de habilidade da TRI (representado por *theta*). A exemplo dos índices de dificuldades, utilizou-se a média das informações de *escore total* e *theta* dos 26 grupos que responderam a cada um dos 26 cadernos como índices.

Já tendo sido calculado índices ou estimativas de dificuldade (*valor p* e *parâmetro b*), bem como dois índices de habilidades (*escore total* e *theta*), tinham-se disponíveis informações de ambas as teorias, o que permitiu correlações e comparações entre elas.

Para a realização de comparações entre os resultados do índice de *escore total* médios e de correlações entre os resultados deste índice com os de dificuldade (*valor p* e *parâmetro b*) médios, no entanto, deve-se notar que o número diferenciado de itens para cada um dos cadernos torna injusto o cômputo dos *escores totais*. Os grupos de estudantes que responderam a um número menor de itens teriam menores chances de conseguir um *escore total* mais alto. Por outro lado, aqueles grupos que responderam a um caderno com um número maior de itens poderiam ter conseguido *escore total* mais alto, por terem maiores oportunidades de acerto. Na tentativa de minimizar a influência do número de itens dos cadernos nos resultados médios de *escore total* dos examinandos, procedeu-se um ajuste desses escores, simulando uma situação em que todos os estudantes tivessem respondido a um caderno de 39 itens.

Assim o escore total de cada caderno foi dividido pelo seu número de itens. Obteve-se taxa de acerto por item.

Assim,

taxa de acerto = (Média original / n itens caderno)

Multiplicou-se essa taxa por 39, que é o número de itens do caderno para o qual a média será ajustada, obtendo-se assim o escore total médio ajustado.

Assim,

escore total ajustado = taxa de acerto · 39

Quando estivermos nos referindo ao *escore total* médio dos estudantes que responderam a cada um dos cadernos, consideraremos um índice transformado para 39 itens. Não foi útil, neste momento da análise, realizar qualquer tipo de equalização, visto que a variabilidade nas médias do *escore total* seria eliminada.

Utilizando a base de dados constituída pelas 26 informações de dificuldade e de habilidade, pelas duas teorias, foram calculadas a média, o desvio padrão e a amplitude entre elas.

Para o cálculo da amplitude das médias de dificuldade e habilidade, em unidades de desvio padrão, dividiu-se a amplitude original pela média dos desvios-padrão dos índices dos cadernos, sem ajuste. Foi utilizada a média dos desvios-padrão sem ajuste, pois não se tinha disponível estes resultados com o ajuste. Certamente a variabilidade, após um ajuste para 39 itens seria maior.

2.3.3 A associação entre as dificuldades dos cadernos e as habilidades dos estudantes

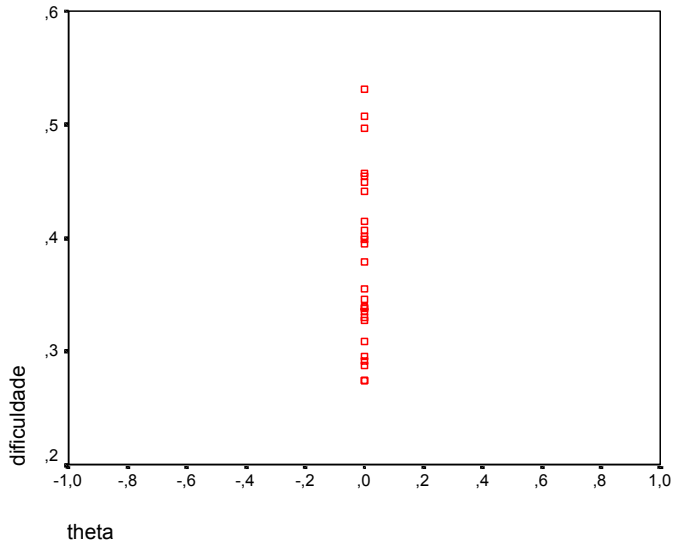
Foi estruturada uma base de dados que continha os resultados referentes (i) aos 26 índices de dificuldade (*valor p*) médios, pela TCT, (ii) aos 26 parâmetros de dificuldade (*parâmetro b*) médios, pela TRI, para cada um dos cadernos, (iii) aos 26 índices de *escore*

total médios dos estudantes, pela TCT, e (iv) às 26 estimativas de habilidade da TRI (*theta*) médias para os estudantes que responderam a cada um dos cadernos.

Foram calculadas as correlações (r de Pearson) e criados gráficos de dispersão entre (i) os valores médios do *parâmetro b* dos cadernos e os índices de *escore total* médios dos estudantes, (ii) o *valor p* médio dos cadernos com o *theta* médio dos estudantes, e (iii) o *parâmetro b* médio dos cadernos com o *theta* médio dos estudantes.

A análise desses resultados foi realizada à luz da TRI. As vantagens que essa teoria supostamente apresenta sobre a TCT, em termos da invariância dos parâmetros de habilidade foi investigada. Esperavam-se correlações mais baixas quando o *theta* estava envolvido e, mais altas, quando o *escore total* estava envolvido. No caso do *theta*, esperava-se encontrar gráficos de dispersão com nuvens de pontos que se aproximavam de uma reta vertical, considerando o eixo X como o *theta* e o eixo Y como a dificuldade. Assim verificaríamos o quão invariante é o *theta* dos estudantes, tendo em vista a variabilidade da dificuldade dos cadernos.

Figura 2.2 – Gráfico de dispersão entre a dificuldade dos cadernos e o theta esperado pela propriedade de invariância dos parâmetros.



No caso da associação entre o *escore total* e a dificuldade, por sua vez, esperavam-se gráficos com nuvens inclinadas que sugeririam uma correlação perfeita. Aí estaria representada a dependência circular entre dificuldade e *escore total*.

2.3.4 A unidimensionalidade como condição da invariância do parâmetro de habilidade pela TRI

Um outro aspecto que se procurou investigar no âmbito dessas correlações foi a influência de itens que praticamente não contribuíam para a unidimensionalidade na invariância do parâmetro de habilidade pela TRI. Retoma-se neste momento do estudo, o que Baker (2001) considerou como uma das condições básicas para a existência de invariância do *theta*: a de "todos os itens medirem o mesmo traço latente". Levantou-se a seguinte questão: será que retirando os itens da prova com baixas cargas fatoriais no fator único, as correlações entre os índices *valor p* e *parâmetro b* com o *theta* não tenderiam a ser menores? Ou seja, será que a dependência entre o *theta* e os índices de dificuldade não seria menor?

Utilizando uma análise prévia da unidimensionalidade das provas do Saeb 97 (Laros, Pasquali & Rodrigues, 2000), pôde-se identificar aqueles itens da prova de Matemática de 8ª Série que apresentavam baixas cargas fatoriais no fator único (abaixo de 0,20). Foram excluídos da análise clássica e da estimação dos parâmetros da TRI os 26 itens que apresentaram cargas fatoriais no fator único menores que 0,20. Foram excluídos também outros quatro itens que já deixaram de serem usados para o cálculo dos parâmetros clássicos nas análises sem a exclusão dos itens. Cabe ressaltar que as cargas fatoriais destes quatro itens não ultrapassaram 0,30. Ressalta-se também que a prova apresentava, após a exclusão dos itens, um total de 131 itens, divididos em 26 cadernos que variaram de 19 a 39 itens.

Após a retirada 26 itens, foram novamente calculados ou estimados os índices de dificuldade e de habilidade, pela TCT e pela TRI, médios para os 26 cadernos. Na estimação dos parâmetros de dificuldade e habilidade pela TRI, a partir do BILOG-MG, todos os comandos do programa foram mantidos os mesmos da primeira estimação, à exceção do comando *groups*, que indica os itens que entrariam nesta estimação. Esse procedimento é justificado pois, esperava-se comparar as estimativas de *theta* sem a exclusão e, posteriormente, com a exclusão dos itens e quanto menor a influência de outros fatores que não essa eliminação de itens poderia influenciar os resultados.

Com a estimativa do novo *theta*, já se tinha disponível as estimativas do novo *parâmetro b*, que foram calculados conjuntamente. Retirando-se os mesmos itens, foram também calculados os novos índices *valor p* e *escore total*. Essas estatísticas compuseram uma segunda base de dados.

De posse dessas informações de médias, foram calculadas a média, o desvio padrão e a amplitude entre elas. Também foram novamente calculadas as correlações (r de Pearson) e criados gráficos entre (i) o *parâmetro b* dos cadernos com o *escore total* dos estudantes, (ii) o *valor p* dos cadernos com o *theta* dos estudantes, e (iii) o *parâmetro b* dos cadernos com os *theta* dos estudantes, agora com a exclusão dos itens que não contribuíam significativamente para a unidimensionalidade. Os resultados foram comparados com os gráficos e coeficientes de correlação sem a exclusão dos itens. Esperava-se que, se efetivamente a unidimensionalidade fosse uma condição para que ocorra a invariância do parâmetro de *theta*, a correlação entre o *theta* e a dificuldade fosse menor que no caso sem a exclusão de itens. Esperavam-se gráficos com nuvens de dispersão mais próximas ainda à idéia de uma reta vertical.

Para diferenciar os índices sem a exclusão de itens daqueles com a exclusão de itens, utilizou-se, para o segundo, uma notação cujo índices estavam acompanhados pela letra d (de

dimensionalidade). Assim, quando estivermos nos referindo aos índices e parâmetros de dificuldade e aos índices de habilidade, após a exclusão dos itens, utilizaremos: *valor p_d*, *parâmetro b_d*, *escore total_d* e *theta_d*.

3. Resultados

A fase inicial do estudo foi realizar o levantamento de informações sobre as dificuldades dos cadernos de provas, já que esta variável é uma "peça-chave" para o alcance dos objetivos propostos. Só se poderia investigar a influência da diferença das dificuldades de provas sobre a variável de habilidade estimada pela TRI se houvesse variabilidade entre essas dificuldades. No entanto, percebeu-se que, para a realização de certas comparações entre as dificuldades médias de 26 cadernos, seria necessário que os 26 grupos que o responderam fossem equivalentes em termos de habilidades. De tal forma, seria como se fosse o mesmo grupo respondendo a provas diferentes. Partiu-se, assim, para a investigação dos resultados da equivalência entre os grupos.

3.1 Verificação da equivalência entre os grupos

Como resultado da análise do delineamento amostral e dos procedimentos de aplicação das provas, considerou-se que, a amostragem probabilística, aliada à forma de distribuição dos itens pelos cadernos e dos cadernos pela amostra, garante a equivalência dos grupos. No sentido apenas de confirmação dessa equivalência, a seguir são apresentados os resultados de investigações com bases estatísticas.

Todos os procedimentos estatísticos utilizados para confirmar a equivalência entre os grupos sempre foram baseados na investigação das características dos examinandos de diferentes grupos, em resposta a um mesmo bloco de itens. A análise realizada por bloco permite comparar diretamente o desempenho na prova de dois grupos de alunos.

3.1.1 Estatísticas descritivas do escore total do estudante em resposta aos blocos de itens

As médias de escores totais dos estudantes por bloco, bem como a diferença entre elas considerando os blocos de mesma posição e, também, independentemente da posição que ocupa nos cadernos estão apresentadas no anexo I. Neste são apresentadas também as diferenças entre médias normalizadas nessas duas situações (os procedimentos de normalização já foram abordados no tópico 2.3.1).

As informações deste anexo são muito importantes, pois se consegue ter uma visão geral das médias e conseqüentemente de suas diferenças e verificar que as médias dos escores totais dos blocos são muito próximas. Por exemplo, na primeira posição o escore total médio do grupo que respondeu ao caderno 1 foi de 7,39 e o do grupo que respondeu ao caderno 14 foi de 7,31. Essa diferença de 0,04 d.p. está indicando que os grupos apresentam praticamente a mesma habilidade média. Se atentarmos para os desvios-padrão de cada uma das médias de escores totais, verificamos que a variabilidade de cada um dos grupos também é praticamente a mesma. Veja que os dois grupos que responderam ao bloco 1, quando este estava situado na mesma posição (cadernos 1 e 14), apresentam desvios-padrão de 2,22 e 2,19, resultados bastante próximos. Para os outros cadernos deste bloco não eram muito diferentes, se considerarmos a posição em que esse bloco se encontra.

Buscando sintetizar as informações do anexo I, são apresentadas as principais estatísticas descritivas relativas às diferenças entre médias e às diferenças entre médias normalizadas.

Tabela 3.1 – Diferenças entre os escores totais e diferenças entre os escores totais normalizados dos estudantes nos blocos que compunham os cadernos de Matemática do Saeb.

	Diferença entre *escores totais* brutos		**Diferença entre *escores totais* normalizados**	
	Blocos na mesma posição	**Blocos independentemente da posição**	**Blocos na mesma posição**	**Blocos independentemente da posição**
Número de Comparações	39	13	39	13
Média	0,27	0,69	0,12	0,30
DP	0,22	0,27	0,10	0,10
Mínimo	0,01	0,31	0,01	0,17
Máximo	0,93	1,29	0,38	0,53
Amplitude	0,92	0,98	0,38	0,37

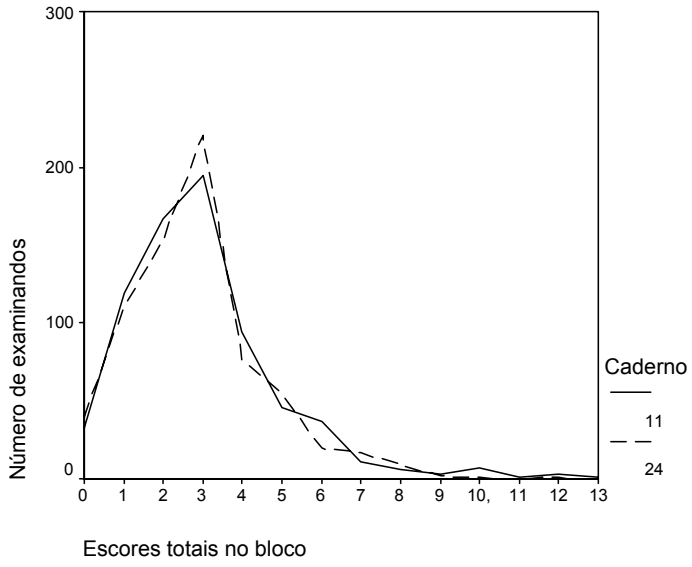
Observa-se, para os blocos de mesma posição, uma diferença média de 0,12 d.p. nos escores totais nos blocos, com desvio-padrão das diferenças entre as médias de 0,10. Essa diferença média é considerada pequena pela interpretação dos resultados do índice *d* de Cohen (Ver a classificação apresentada no tópico 2.3.1). A menor diferença encontrada foi de 0,01 d.p. (ou seja, praticamente nenhuma diferença), para um determinado bloco, e a maior de 0,38 d.p. Já na situação em que não se controlou a posição em que o bloco estava localizado no caderno, as diferenças normalizadas foram em média de 0,30 d.p., com desvio-padrão de 0,10 d.p. A menor diferença normalizada foi de 0,17 d.p. e a maior, de 0,53 d.p.

Os resultados indicam que, em média, as diferenças entre os escores totais foram menores para blocos de mesma posição que para blocos independentemente da posição. Analisando-se apenas as diferenças médias entre os escores totais, quando os blocos estão situados em uma mesma posição do caderno, observa-se que uma amplitude de 0,38 d.p. Essa variabilidade é pequena em termos de unidades de desvio-padrão (*d* de Cohen < 0,50).

3.1.2 Análise gráfica da distribuição de freqüências dos escores totais nos blocos

Os resultados da análise gráfica da distribuição dos alunos dos grupos em função dos escores totais no bloco indicaram para linhas muito próximas, na maioria das vezes. Veja o exemplo do dos grupos de estudantes que responderam ao bloco 11.

Figura 3.1 – Gráfico da freqüência de estudantes por escore no bloco 11, localizado na primeira posição do caderno.



Cabe relatar que as linhas de freqüência dos escores totais dos estudantes em alguns blocos não eram tão próximas. Esses desajustes, no entanto, foram poucos e não podem ser considerados tão expressivos.

3.1.3 Teste U de Mann Whitney dos escores totais nos blocos

Os resultados da prova *U* de Mann Whitney realizada para os grupos de estudantes que responderam a cada um dos blocos, quando esses estavam localizados na primeira posição do caderno, estão apresentados na tabela 3.2. Como pode ser observado, para aqueles grupos que responderam ao blocos 3, 6, 7 e 13, não se pode dizer que apresentam características

semelhantes, pois as diferenças entre as distribuições de freqüências dos escores totais nos blocos foram significativas para um nível de 0,05.

Tabela 3.2 – Resultados do teste U de Mann-Whitney das distribuições de freqüências dos escores totais dos estudantes no bloco, situados na primeira posição do caderno.

Bloco	**Caderno**	**N**	***escore total***				
			Média	**Diferença entre médias**	**U**	**z**	**Sig.**
1	1	756	7,39		259407,5	-0,61	0,543
	14	699	7,31	0,08			
2	2	726	3,73		253064,5	-0,42	0,678
	15	706	3,81	0,08			
3	3	700	6,26		234664,0	-2,00	0,050*
	16	714	6,59	0,33			
4	4	760	7,83		259550,0	-1,59	0,112
	17	717	7,57	0,26			
5	5	753	4,47		264485,0	-0,21	0,831
	18	707	4,49	0,02			
6	6	752	4,13		218544,5	-4,70	0,000*
	19	678	4,66	0,52			
7	7	734	4,61		197417,0	-6,90	0,000*
	20	681	5,53	0,93			
8	8	746	3,76		244678,5	-1,86	0,062
	21	695	4,00	0,24			
9	9	741	3,36		257204,0	-0,65	0,517
	22	708	3,43	0,07			
10	10	761	4,25		262081,0	-0,95	0,340
	23	709	4,37	0,11			
11	11	719	2,57		246808,0	-0,64	0,521
	24	700	2,51	0,06			
12	12	750	3,40		266292,5	-0,32	0,747
	25	717	3,33	0,07			
13	13	716	3,08		260713,5	-2,38	0,017*
	26	783	2,77	0,32			

* Diferenças significativas ao nível p de 0,05.

A partir da análise dos grupos que responderam ao bloco 6 na primeira posição (aqueles que responderam ao caderno 6 e ao caderno 19), observou-se que a hipótese nula que eles apresentam idênticas distribuições foi rejeitada. No entanto, sabe-se que o grupo que respondeu ao caderno 6 respondeu também ao bloco 10 (Ver tabela 1.2), quando esse estava na terceira posição. Testando-se as respostas dos estudantes ao bloco 10 na terceira posição

(cadernos 6 e 16), verifica-se que a hipótese nula de idênticas distribuições de freqüências de escores totais não foi rejeitada, o que aponta para a equivalência dos estudantes que responderam ao caderno 6 com aqueles que responderam ao caderno 16. O grupo que respondeu ao caderno 16 apresentou características semelhantes ao que respondeu ao caderno 3 (bloco 3 na primeira posição) e, assim, sucessivamente.

Os resultados da testagem estatística indicam que, embora diretamente não se tenha encontrado equivalência entre os grupos que responderam aos cadernos 6 e 19, indiretamente pode-se concluir que esses grupos são equivalentes a grupos que responderam a outros cadernos, que por sua vez, são equivalentes a grupos que responderam a outros cadernos. Observou-se que todos os estudantes que responderam aos cadernos 3 e 16 (presença do bloco 3 na primeira posição), 6 e 19 (bloco 6), cadernos 7 e 20 (bloco 7) e cadernos 13 e 26 (bloco 13) apresentaram características comuns com outros grupos que responderam a outros cadernos, a partir da análise das respostas aos blocos localizados na segunda e terceira posições.

3.1.4 Síntese dos resultados da verificação da equivalência entre os grupos

Se levarmos em consideração os resultados (i) das estatísticas descritivas, (ii) da análise gráfica das distribuições de freqüências, e (iii) das estatísticas não-paramétricas das distribuições de freqüências, todos com referência aos escores totais nos blocos, a hipótese que os grupos que responderam aos cadernos do Saeb são equivalentes não pode ser rejeitada.

3.2 Dificuldades dos cadernos de prova do Saeb

Uma vez que foi verificado que os grupos de estudantes que responderam aos 26 cadernos são equivalentes, foram inicialmente calculadas as estatísticas de dificuldade dos

itens que compunham esses cadernos. Os resultados das estatísticas de dificuldade da TCT (*valor p*) podem ser encontrados na tabela 3.3, a seguir.

Tabela 3.3 – Índice de dificuldade clássica *(valor p)* dos itens dos cadernos de Matemática do Saeb.

Caderno	**N**	**n itens**	***valor p***			
			Média	**DP**	**Mínimo**	**Máximo**
1	757	33	0,46	0,21	0,10	0,85
2	731	34	0,41	0,18	0,16	0,78
3	705	35	0,53	0,21	0,16	0,89
4	742	35	0,45	0,21	0,10	0,89
5	762	36	0,34	0,13	0,10	0,63
6	752	39	0,35	0,12	0,16	0,81
7	735	38	0,31	0,15	0,07	0,81
8	744	36	0,29	0,12	0,12	0,62
9	729	36	0,29	0,10	0,11	0,62
10	771	37	0,38	0,21	0,07	0,85
11	722	36	0,27	0,14	0,07	0,59
12	757	36	0,34	0,19	0,11	0,78
13	725	34	0,51	0,25	0,11	0,89
14	688	34	0,50	0,21	0,16	0,85
15	707	32	0,45	0,23	0,12	0,89
16	698	36	0,41	0,18	0,10	0,78
17	717	37	0,40	0,22	0,07	0,89
18	706	38	0,34	0,16	0,10	0,81
19	686	37	0,30	0,10	0,11	0,51
20	680	35	0,44	0,21	0,12	0,85
21	693	35	0,33	0,12	0,17	0,59
22	691	35	0,36	0,20	0,07	0,78
23	714	37	0,40	0,22	0,14	0,89
24	704	37	0,27	0,13	0,07	0,63
25	727	37	0,40	0,20	0,14	0,85
26	761	35	0,33	0,16	0,11	0,81
Média	723,29	-	0,38	0,18	0,11	0,78
Mediana	723	36	0,37	0,19	0,11	0,81
DP	26,78	-	0,07	0,04	0,03	0,12
Mínimo	680	32	0,27	0,10	0,07	0,51
Máximo	771	39	0,53	0,25	0,17	0,89
Amplitude	91	7	0,26	0,15	0,10	0,38

A média dos índices de dificuldade dos cadernos pela TCT (*valor p*) foi de 0,38, com desvio-padrão de 0,07. O percentual de acertos aos itens que compõem os cadernos variou de 0,27 (cadernos 11 e 24, que são, em média, os mais difíceis) a 0,53 (caderno 3, que, em média, é o mais fácil). Pode-se observar que existem grandes diferenças em relação à

dificuldade dos cadernos. A diferença entre o caderno mais fácil e mais difícil é de 0,26, que representa uma variabilidade de 1,44 d.p, que pode ser considerada grande.

Os resultados referentes ao índice de dificuldade estimado pela TRI (*parâmetro b*), estão apresentados na tabela 3.4.

Tabela 3.4 – Índice de dificuldade pela TRI (*parâmetro b*) dos itens dos cadernos de Matemática do Saeb.

Caderno	**N**	**n itens**	***parâmetro b***			
			Média	**DP**	**Mínimo**	**Máximo**
1	757	33	0,91	1,34	-1,76	5,15
2	731	34	0,90	1,00	-0,90	2,71
3	705	35	0,43	1,20	-1,65	3,60
4	742	35	0,83	1,12	-1,65	3,31
5	762	36	1,27	0,90	-0,62	2,71
6	752	39	1,18	0,89	-1,38	3,60
7	735	38	1,57	0,98	-1,38	3,60
8	744	36	1,59	0,83	-0,62	3,31
9	729	36	1,55	0,82	-0,62	3,40
10	771	37	1,24	1,28	-1,76	5,15
11	722	36	1,69	0,83	-0,14	3,38
12	757	36	1,50	1,10	-0,90	3,40
13	725	34	0,81	1,60	-1,76	5,15
14	688	34	0,79	1,39	-1,76	5,15
15	707	32	0,68	1,15	-1,65	2,69
16	698	36	0,91	0,87	-0,90	2,07
17	717	37	1,04	1,22	-1,65	3,38
18	706	38	1,42	1,00	-1,38	3,60
19	686	37	1,64	0,84	0,08	3,40
20	680	35	1,01	1,45	-1,76	5,15
21	693	35	1,24	0,71	-0,14	3,31
22	691	35	1,28	1,13	-0,90	3,38
23	714	37	0,94	1,08	-1,65	2,77
24	704	37	1,81	0,82	-0,15	3,40
25	727	37	1,23	1,33	-1,76	5,15
26	761	35	1,49	1,08	-1,38	3,60
Média	723,31	-	1,19	1,08	-1,16	3,67
Mediana	723	36	1,23	1,08	-1,38	3,40
DP	26,80	-	0,36	0,23	0,61	0,90
Mínimo	680	32	0,43	0,71	-1,76	2,07
Máximo	771	39	1,81	1,60	0,08	5,15
Amplitude	91	7	1,38	0,89	1,85	3,08

A média do *parâmetro b* para os 26 cadernos é de 1,19, com desvio-padrão de 0,36. O caderno 3 é aquele com menor dificuldade, com *parâmetro b* médio de 0,43, e o caderno 24 é

aquele com maior dificuldade, com *parâmetro b* médio de 1,81. Isso representa uma amplitude de 1,38 entre as médias do *parâmetro b*. Se essa amplitude for dividida pela média dos desvios-padrão, observa-se que é de 1,28 d.p., considerada grande.

É muito importante chamarmos a atenção para um determinado valor máximo do *parâmetro b* que é de 5,15. Ele se refere a um item com uma dificuldade muito alta que, por opção, não foi retirado da base e que certamente elevou a média desse índice. Esse item estava presente nos cadernos 1, 10, 13, 14, 20 e 25.

Conclui-se sobre a investigação dos índices de dificuldade da TCT e da TRI que a variabilidade dos cadernos é grande. Verifica-se que, a partir de ambos os modelos de análise, o caderno 3 se mostrou o mais fácil e o caderno 24, um dos mais difíceis.

3.3 Habilidades dos estudantes

Partindo-se para os resultados de investigação dos índices de habilidade dos examinandos, primeiramente são apresentadas, na tabela 3.5, as estatísticas descritivas, pela TCT, do *escore total* médios dos estudantes e os resultados deste índice ajustado em resposta a cada um dos cadernos de prova.

Tabela 3.5 – *Escore total* dos estudantes que responderam aos cadernos de Matemática do Saeb.

Caderno	N	n itens	*escore total*				Taxa de acerto por item	*escore total* ajustado
			Média	d.p.	Mínimo	Máximo		
1	757	33	15,80	5,52	2	32	0,48	18,67
2	731	34	14,10	6,30	0	33	0,41	16,18
3	705	35	19,03	6,33	1	35	0,54	21,20
4	742	35	16,05	5,44	0	35	0,46	17,89
5	762	36	11,55	6,21	0	34	0,32	12,51
6	752	39	13,31	7,39	1	37	0,34	13,31
7	735	38	10,89	5,19	0	33	0,29	11,18
8	744	36	10,48	4,99	0	31	0,29	11,35
9	729	36	9,71	5,43	0	33	0,27	10,52
10	771	37	14,09	5,77	1	37	0,38	14,85
11	722	36	9,86	4,64	0	36	0,27	10,68
12	757	36	12,11	5,00	0	30	0,34	13,12
13	725	34	16,85	5,45	0	32	0,50	19,33
14	688	34	17,95	5,71	3	32	0,53	20,59
15	707	32	15,34	5,78	1	32	0,48	18,70
16	698	36	15,44	7,08	0	36	0,43	16,73
17	717	37	15,14	6,14	1	35	0,41	15,96
18	706	38	12,25	5,92	0	36	0,32	12,57
19	686	37	11,00	5,11	0	33	0,30	11,60
20	680	35	16,21	5,92	1	33	0,46	18,07
21	693	35	11,50	6,03	0	35	0,33	12,82
22	691	35	11,72	5,43	0	32	0,33	13,05
23	714	37	14,58	6,16	0	35	0,39	15,37
24	704	37	9,54	4,28	0	33	0,26	10,06
25	727	37	14,49	5,32	1	34	0,39	15,27
26	761	35	11,16	4,63	0	31	0,32	12,44
Média	723,29	-	13,47	5,66	-	-	0,38	14,77
Mediana	724	36	13,70	5,62	0	33	0,36	14,08
DP	26,78	-	2,68	0,71	-	-	0,08	3,31
Mínimo	680	32	9,54	4,28	0	30	0,26	10,06
Máximo	771	39	19,03	7,39	3	37	0,54	21,20
Amplitude	91	7	9,48	3,11	3	7	0,29	11,14

A média nos escores totais ajustados dos caderno é de 14,77, com desvio-padrão das médias de 3,31. O caderno em que os estudantes obtiveram o menor *escore total* ajustado foi o caderno 24, com 10,06. O maior *escore total* foi o caderno 3, com 21,20. Comparando com os índices de dificuldades calculados, esses resultados já indicam que para o caderno mais fácil, o caderno 3, os estudantes obtiveram os maiores escores médios. Da mesma forma, para

um dos cadernos mais difíceis, o caderno 24, os estudantes obtiveram os menores escores médios. A amplitude de 11,14 é de quase 2 d.p, resultado alcançado da divisão da amplitude do escore total ajustado pela média dos desvios-padrão deste índice sem o ajuste.

As estatísticas relacionadas às habilidades estimadas pela TRI (*theta*) dos examinandos que responderam aos cadernos de Matemática estão apresentados na tabela 3.6, a seguir.

Tabela 3.6 – Estimativas da habilidade pela TRI (*theta*) dos estudantes que responderam aos cadernos de Matemática do Saeb.

Caderno	N	*theta*			
		Média	**d.p.**	**Mínimo**	**Máximo**
1	757	0,13	0,96	-2,10	2,77
2	731	0,00	0,94	-1,89	2,71
3	705	0,08	0,92	-2,21	2,85
4	742	0,10	0,83	-2,09	2,99
5	762	-0,09	0,91	-1,72	2,97
6	752	-0,06	0,93	-1,81	2,82
7	735	-0,11	0,81	-1,68	2,68
8	744	-0,04	0,86	-1,44	2,46
9	729	-0,09	0,84	-1,38	2,87
10	771	0,02	0,88	-1,93	3,03
11	722	-0,01	0,85	-1,34	3,27
12	757	-0,09	0,91	-1,72	2,66
13	725	-0,03	0,95	-2,22	2,73
14	688	0,18	0,93	-2,10	2,47
15	707	0,16	0,87	-2,10	2,98
16	698	0,11	0,97	-1,81	2,93
17	717	0,11	0,94	-2,08	2,71
18	706	-0,08	0,94	-1,73	2,87
19	686	0,06	0,80	-1,41	3,01
20	680	0,13	0,91	-1,96	2,79
21	693	-0,02	0,87	-1,53	3,05
22	691	-0,17	0,92	-1,85	2,63
23	714	-0,01	0,87	-2,04	3,01
24	704	-0,19	0,80	-1,32	2,82
25	727	0,01	0,86	-1,96	2,75
26	761	-0,06	0,82	-1,74	2,60
Média	723,31	0,00	0,89	-1,81	2,82
Mediana	723	-0,01	0,90	-1,83	2,82
DP	26,80	0,10	0,05	0,28	0,19
Mínimo	680	-0,19	0,80	-2,22	2,46
Máximo	771	0,18	0,97	-1,32	3,27
Amplitude	91	0,37	0,18	0,91	0,80

A média dos resultados de *theta* dos estudantes em reposta aos cadernos foi de 0 (zero), o que equivale ao centro da escala, que varia de –3 a 3, com desvio-padrão de 0,10. O *theta* dos estudantes variou de –0,19, referente ao grupo que respondeu ao caderno 24, a 0,18 , referente ao grupo que respondeu ao caderno 14, o que representa uma amplitude de 0,37. O

tamanho desta amplitude é de 0,42 d.p[1]., *d* de Cohen considerado pequeno. Cabe observar, também, que os desvios-padrão do *theta* dos estudantes nos cadernos foram bastante semelhantes. Esses variaram em 0,05 d.p.

É interessante notar que, da mesma forma que para o índice de *escore total* dos estudantes que responderam aos cadernos, o grupo com a menor habilidade respondeu ao caderno 24. Por outro lado, o grupo com maior *theta* não respondeu ao caderno 3, como foi para o *escore total*. Verifica-se, pelo menos preliminarmente, que não parece existir uma relação perfeita entre a dificuldade e o *theta*. Observa-se, por exemplo, que o *theta* do grupo que respondeu ao caderno 3 (0,08) é inferior ao *theta* do grupo que respondeu a sete cadernos (1, 4, 14, 15, 16, 17 e 20), mesmo sendo considerado o caderno com menor dificuldade.

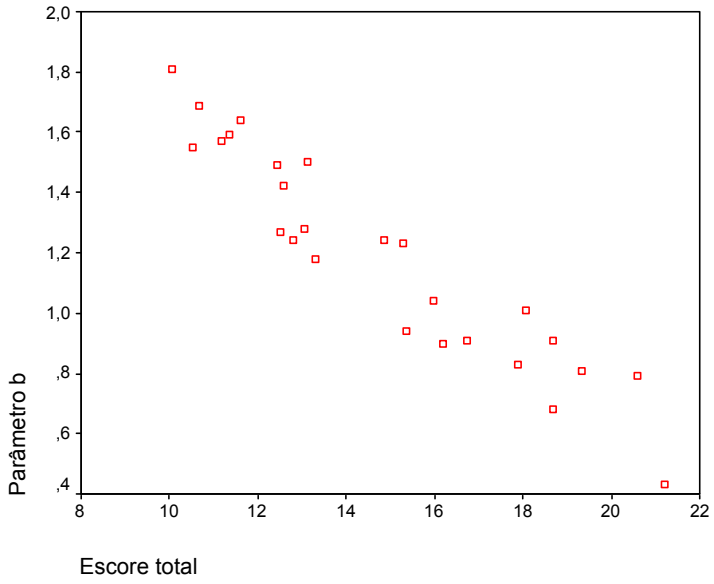
3.4 Associação entre as dificuldades dos cadernos e as habilidades dos estudantes

Em posse dos dois índices de dificuldade (*valor p* e *parâmetro b*) e dos dois índices de habilidade (*escore total* e *theta*) médios para cada um dos cadernos, foram realizadas as seguintes correlações: o *valor p* com o *theta* e o *parâmetro b* com o *escore total* e o *theta*.

A correlação entre o *valor p* e o *escore total* é, por definição, perfeita. Associando o *parâmetro b* com o *escore total*, é observada uma correlação alta e negativa (r = -0,95), em que quanto maior é o *parâmetro b* do caderno, menor é o *escore total* dos estudantes. Na figura 3.2, o gráfico de dispersão entre estas duas variáveis é apresentado.

[1] Resultado da divisão da amplitude média pela média dos desvios-padrão de *theta*.

Figura 3.2 – Gráfico de dispersão entre o índice de dificuldade dos cadernos pela TRI (*parâmetro b*) e o *escore total* dos respondentes aos cadernos de Matemática do Saeb.



Na figura 3.2, cada ponto do gráfico representa um caderno. O caderno 3, o mais fácil, está situado na extrema direita do eixo X, onde estão situados os escores mais altos. Já o caderno 24, o mais difícil, está situado na extrema esquerda do eixo X, onde estão situados os escores mais baixos. Verifica-se com essa associação que o *escore total é* altamente dependente do índice de dificuldade estimado pela TRI.

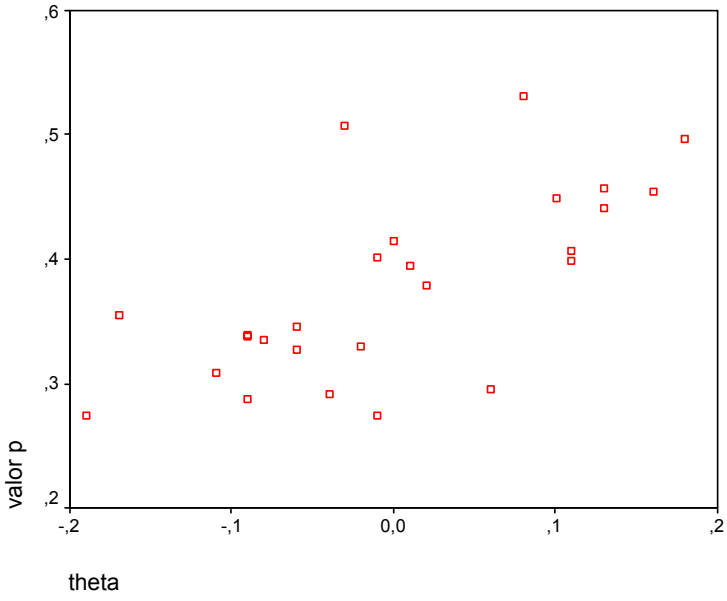
Da mesma forma, se tivéssemos correlacionado o *escore total* com o *valor p* dos cadernos obteríamos uma correlação perfeita de 1, a correlação entre o *parâmetro b* e o *escore total* é extremamente alta. A correlação de -0,95 implica que 90% da variância do *escore total* estão associadas à variância do *parâmetro b*.

Observa-se também que a correlação entre o *parâmetro b* e o *valor p* também foi extremamente alta (– 0,95), o que indica que o *parâmetro b* é uma estimativa de dificuldade que apresenta características semelhantes à proporção de acerto dos estudantes aos itens (*valor p*).

Utilizou-se, em outro momento, uma outra variável para medir a habilidade, o *theta*, em associação com os índices de dificuldade. Primeiramente, este foi correlacionado com o *valor p* e, posteriormente, com o *parâmetro b*.

A associação entre *valor p* e *theta* forneceu um coeficiente de correlação r de 0,68. A dispersão dos pontos dessa correlação pode ser observada na figura 3.3.

Figura 3.3 – Gráfico de dispersão entre o índice de dificuldade clássico dos cadernos pela TCT (*valor p*) e as habilidades estimadas pela TRI (*theta*) dos respondentes aos cadernos de Matemática do Saeb.

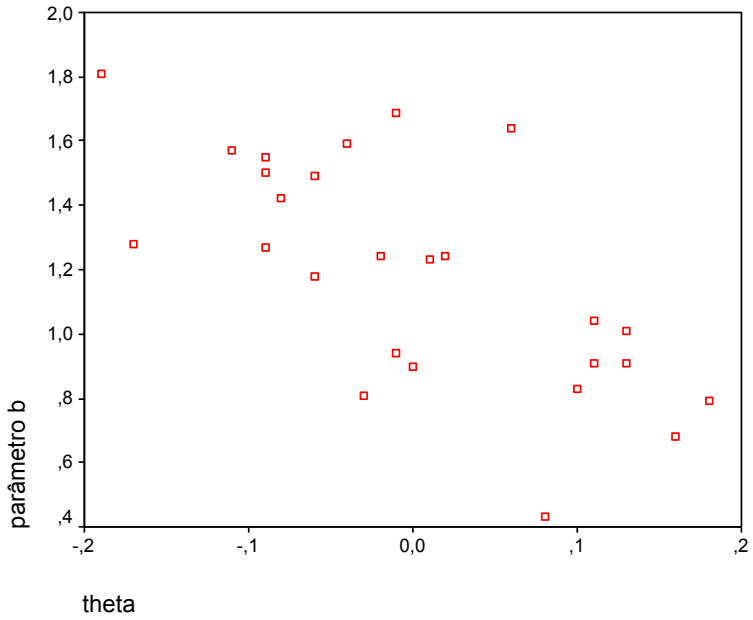


O gráfico aponta para uma associação mais fraca que a do *parâmetro b* com o *escore total*. A nuvem de pontos se encontra bem mais dispersa. Cabe ressaltar que o eixo X da figura 3.3 está variando de – 0,2 e 0,2, refletindo um foco da escala que varia de –3,0 a 3,0. Apresentaram-se os resultados sob essa configuração com o objetivo de se conseguir clarificar as associações. Essa observação faz-nos considerar que, embora a figura não transpareça, a variabilidade de *theta* é pequena, mesmo com a grande variabilidade da dificuldade.

Quando o outro índice de dificuldade, o *parâmetro b*, é associado com o *theta*, observa-se um coeficiente de correlação r de –0,69, que, em módulo é praticamente o mesmo

valor da associação deste índice de habilidade com a dificuldade clássica. A figura 3.4 pode fornecer uma visão mais clara dessa correlação.

Figura 3.4 – Gráfico de dispersão entre o índice de dificuldade dos cadernos pela TRI (*parâmetro b*) e as habilidades estimadas pela TRI (*theta*) dos respondentes aos cadernos de Matemática do Saeb.



Semelhantemente aos dados de correlação com o *valor p*, o *theta* se mostrou bem menos linearmente associado ao *parâmetro b* que o *escore total*. Não se pode concluir, entretanto, que a habilidade estimada pela TRI é isenta de qualquer dependência do conjunto de itens que está sendo aplicado. Note-se que ainda existe uma correlação de - 0,69. Isso quer dizer que 48% da variância do *theta* estão associadas à variância do *parâmetro b* da prova. Percebe-se que, embora a correlação entre as variáveis de dificuldade (*valor p* ou *parâmetro b*) com o *theta* seja menor que a correlação destas variáveis com o *escore total*, existe ainda uma associação forte entre elas.

É fundamental novamente aqui considerar que a figura apresentada está apresentando o eixo X variando de – 0,2 e 0,2, refletindo um foco da escala que varia de –3,0 a 3,0. Embora exista uma associação com a dificuldade, a variabilidade de *theta* é pequena, mesmo com a

grande variabilidade da dificuldade. Se aumentássemos a escala do eixo X para –3,0 a 3,0, por sua vez, observaríamos uma nuvem de pontos que se aproxima de uma reta vertical.

Por sua vez, a Teoria de Resposta ao Item considera que uma das duas condições necessárias para que haja a independência entre a habilidade, estimada pela TRI, e a dificuldade dos itens é que se esteja avaliando o mesmo traço latente (Baker, 2001). Essa suposição deu margem a outro tipo de investigação. Quando retiramos os itens que praticamente não contribuíam na mensuração do fator único, será que a dependência entre os índices de dificuldade e *theta* tende a diminuir? Será que esse seria um cuidado que realmente se justificaria ser tomado?

3.5 Dificuldades dos cadernos de prova do Saeb, após a exclusão dos itens

Além dos itens que já vinham sendo excluídos das análises realizadas até esse momento do trabalho, excluindo também aqueles que, pelo estudo de Laros, Pasquali e Rodrigues (2000), apresentaram cargas fatoriais iguais ou abaixo de 0,20 no fator único com relação ao item-total, qual seria o impacto em termos de dependência entre os parâmetros? Na tabela 3.7 são apresentados os índices de dificuldade pela TCT médios para cada um dos cadernos, após a exclusão dos itens (*valor p_d*[2]).

[2] Quando qualquer notação é acompanhada de *"_d"*, significa que houve a retirada dos itens que praticamente não contribuem para a unidimensionalidade (cargas fatoriais menores que 0,20). A letra *d* vem de dimensionalidade.

Tabela 3.7 – Índice de dificuldade clássica *(valor p_d)* dos itens dos cadernos de Matemática do Saeb, após a exclusão do itens que praticamente não contribuem para a unidimensionalidade.

Caderno	**N**	**N itens**	***valor p_d***			
			Média	**d.p.**	**Mínimo**	**Máximo**
1	757	28	0,45	0,22	0,10	0,85
2	731	34	0,41	0,18	0,16	0,78
3	705	32	0,51	0,21	0,16	0,81
4	742	19	0,45	0,22	0,10	0,81
5	762	31	0,31	0,11	0,10	0,62
6	752	39	0,35	0,12	0,16	0,81
7	735	28	0,33	0,15	0,16	0,81
8	744	24	0,31	0,13	0,12	0,62
9	729	34	0,30	0,10	0,12	0,62
10	771	35	0,40	0,20	0,16	0,85
11	722	30	0,30	0,13	0,14	0,59
12	757	30	0,37	0,20	0,14	0,78
13	725	29	0,50	0,24	0,16	0,85
14	688	26	0,56	0,20	0,16	0,85
15	707	29	0,42	0,21	0,12	0,81
16	698	31	0,39	0,18	0,10	0,78
17	717	32	0,38	0,19	0,16	0,81
18	706	29	0,33	0,15	0,10	0,81
19	686	27	0,30	0,09	0,16	0,51
20	680	35	0,44	0,21	0,12	0,85
21	693	27	0,34	0,12	0,17	0,59
22	691	33	0,37	0,20	0,12	0,78
23	714	30	0,39	0,19	0,14	0,81
24	704	28	0,26	0,09	0,10	0,41
25	727	33	0,42	0,20	0,14	0,85
26	761	33	0,34	0,15	0,16	0,81
Média	723,23	-	0,38	0,17	0,14	0,75
Mediana	723	30	0,38	0,19	0,14	0,81
DP	26,70	-	0,07	0,04	0,03	0,12
Mínimo	680	19	0,26	0,09	0,10	0,41
Máximo	771	39	0,56	0,24	0,17	0,85
Amplitude	91	20	0,30	0,15	0,07	0,44

Após a exclusão dos itens que praticamente não contribuem para a unidimensionalidade, a média do índice de dificuldade dos cadernos pela TCT (*valor p_d*) foi de 0,38, com desvio-padrão de 0,07, os mesmos observados sem a exclusão dos itens. Observa-se também que as médias de *valor p_d* para os 26 cadernos variam de 0,26 (caderno 24, que é o mais difícil) a 0,56 (caderno 14, que é o mais fácil). É interessante notar que os caderno 3 e 13 se mostraram como uns dos mais fáceis. A amplitude de *valor p_d* foi de 0,30,

que demonstra uma variabilidade de 1,76 d.p., maior que a variabilidade sem a exclusão dos itens que foi de 1,44 d.p.

Os resultados referentes aos índices médios de dificuldade da TRI, após a exclusão do itens (*parâmetro b_d*), são apresentados na tabela 3.8.

Tabela 3.8 – Índice de dificuldade pela TRI (parâmetro *b_d*) dos itens dos cadernos de Matemática do Saeb, após a exclusão dos itens que praticamente não contribuem para a unidimensionalidade.

Caderno	**N**	**n itens**	***parâmetro b_d***			
			Média	**d.p.**	**Mínimo**	**Máximo**
1	757	28	0,95	1,29	-1,76	4,20
2	731	34	0,88	0,99	-0,89	2,72
3	705	32	0,55	1,16	-1,36	3,60
4	742	19	0,84	1,02	-0,76	2,09
5	762	31	1,37	0,84	-0,65	2,72
6	752	39	1,17	0,88	-1,36	3,60
7	735	28	1,42	0,95	-1,36	3,60
8	744	24	1,42	0,83	-0,65	2,69
9	729	34	1,46	0,74	-0,65	2,72
10	771	35	1,10	1,13	-1,76	4,20
11	722	30	1,48	0,73	-0,15	2,69
12	757	30	1,27	1,04	-0,89	2,72
13	725	29	0,79	1,43	-1,76	4,20
14	688	26	0,50	1,33	-1,76	4,20
15	707	29	0,83	1,07	-0,76	2,67
16	698	31	0,98	0,85	-0,89	2,09
17	717	32	1,07	1,00	-0,76	2,72
18	706	29	1,42	0,93	-1,36	3,60
19	686	27	1,52	0,78	0,09	2,72
20	680	35	0,97	1,38	-1,76	4,20
21	693	27	1,13	0,65	-0,15	2,44
22	691	33	1,16	1,05	-0,89	2,67
23	714	30	0,92	0,85	-0,76	2,69
24	704	28	1,85	0,40	1,07	2,72
25	727	33	1,03	1,24	-1,76	4,20
26	761	33	1,39	1,03	-1,36	3,60
Média	723,23	-	1,13	0,98	-0,96	3,16
Mediana	723	30	1,12	1,00	-0,89	2,72
DP	26,70	-	0,32	0,24	0,68	0,72
Mínimo	680	19	0,50	0,40	-1,76	2,09
Máximo	771	39	1,85	1,43	1,07	4,20
Amplitude	91	20	1,35	1,02	2,84	2,12

A média do *parâmetro b*_d é de 1,13, com desvio-padrão de 0,32. O caderno com menor dificuldade é o caderno 14, com *parâmetro b_d* médio de 0,50, e o caderno com maior

dificuldade é o 24, com *parâmetro b_d* médio de 1,85. Isso representa uma amplitude de 1,35, que significa uma variabilidade de 1,38 d.p. Cabe ressaltar que essa variabilidade também é maior, comparada à informação sem a exclusão de itens, que foi de 1,28 d.p., que já tinha sido considerada grande. As medidas de variabilidade do *parâmetro b_d* indicam que os cadernos são de dificuldades diferentes.

Note-se o valor máximo do *parâmetro b* de 4,20, que se refere a um item muito difícil presente nos cadernos 1, 10, 13, 14, 20 e 25. Esse item aparece nos mesmos cadernos em que foi encontrado um item com *parâmetro b* de 5,15, sem a exclusão dos itens.

Conclui-se sobre a investigação dos índices de dificuldade da TCT e da TRI que os cadernos são de diferentes dificuldades, após a exclusão dos itens que praticamente não contribuem para a unidimensionalidade. Verifica-se que, a partir de ambos os modelos de análise, o caderno 14 se mostrou o mais fácil e o caderno 24, o mais difícil.

3.6 Habilidades dos estudantes, após a exclusão dos itens

Os resultados de *escore total_d* da investigação dos parâmetros de habilidade, após a exclusão dos itens que praticamente não contribuem para a unidimensionalidade, estão apresentados na tabela 3.9. Esta apresenta também as médias ajustadas para 39 itens desse índice.

Tabela 3.9 – *Escore total_d* dos estudantes que responderam aos cadernos de Matemática do Saeb, após a exclusão dos itens que praticamente contribuem para a unidimensionalidade.

Caderno	N	n itens	*escore total_d*				Taxa de acerto por item	*escore total* ajustado
			Média	d.p.	Mínimo	Máximo		
1	757	28	13,22	4,60	2	28	0,47	18,41
2	731	34	14,10	6,30	0	33	0,41	16,18
3	705	32	16,73	5,78	1	32	0,52	20,39
4	742	19	8,67	3,20	0	19	0,46	17,79
5	762	31	9,07	5,45	0	29	0,29	11,41
6	752	39	13,31	7,39	1	37	0,34	13,31
7	735	28	8,53	3,91	0	25	0,30	11,87
8	744	24	7,33	3,81	0	21	0,31	11,91
9	729	34	9,41	5,37	0	32	0,28	10,80
10	771	35	13,88	5,74	0	35	0,40	15,46
11	722	30	8,93	4,47	0	30	0,30	11,61
12	757	30	10,84	4,86	0	27	0,36	14,10
13	725	29	14,36	4,82	0	28	0,50	19,31
14	688	26	15,46	4,71	2	26	0,59	23,19
15	707	29	12,96	5,25	1	29	0,45	17,43
16	698	31	12,89	6,13	0	31	0,42	16,21
17	717	32	12,62	5,61	0	31	0,39	15,38
18	706	29	9,10	4,82	0	28	0,31	12,24
19	686	27	8,36	4,01	0	26	0,31	12,07
20	680	35	16,21	5,92	1	33	0,46	18,07
21	693	27	9,01	4,92	0	27	0,33	13,01
22	691	33	11,53	5,36	0	31	0,35	13,62
23	714	30	11,80	5,50	0	30	0,39	15,34
24	704	28	7,02	3,43	0	26	0,25	9,78
25	727	33	13,79	5,33	0	32	0,42	16,30
26	761	33	10,82	4,56	0	30	0,33	12,79
Média	723,23	-	11,54	5,05	-	-	0,38	14,92
Mediana	724	30	11,66	5,08	0	30	0,38	14,72
DP	26,70	-	2,80	0,94	-	-	0,08	3,31
Mínimo	680	19	7,02	3,20	0	19	0,25	9,78
Máximo	771	39	16,73	7,39	2	37	0,59	23,19
Amplitude	91	20	9,71	4,19	2	18	0,34	13,41

O *escore total_d* ajustado apresenta média de 14,92 e desvio-padrão de 3,31. O caderno em que os estudantes obtiveram o maior escore total médio foi o 14, com 23,19. Aquele em que os estudantes obtiveram o menor escore total médio foi o caderno 24 (9,78). A amplitude dos escores totais médios dos cadernos foi de 13,41, ou seja, de 2,66 d.p.

As estatísticas relacionadas às habilidades estimadas pela TRI dos examinandos que responderam aos cadernos de Matemática, após a exclusão dos itens (*theta_d*), estão apresentados na tabela 3.10.

Tabela 3.10 – Estimativas da habilidade pela TRI (*theta_d*) dos estudantes que responderam cadernos de Matemática do Saeb, após a exclusão dos itens que praticamente não contribuem para a unidimensionalidade.

Caderno	**N**	***Theta_d***			
		Média	**d.p.**	**Mínimo**	**Máximo**
1	757	0,13	0,95	-2,03	2,89
2	731	0,00	0,94	-1,89	2,71
3	705	0,07	0,91	-2,08	2,84
4	742	0,05	0,82	-1,71	2,69
5	762	-0,14	0,92	-1,58	2,93
6	752	-0,07	0,94	-1,81	2,80
7	735	-0,09	0,77	-1,59	2,81
8	744	-0,07	0,85	-1,35	2,37
9	729	-0,09	0,85	-1,40	2,82
10	771	0,03	0,89	-1,94	3,04
11	722	-0,02	0,84	-1,34	3,20
12	757	-0,08	0,91	-1,72	2,62
13	725	-0,04	0,94	-2,14	2,64
14	688	0,20	0,93	-2,10	2,61
15	707	0,13	0,86	-1,85	2,96
16	698	0,12	0,96	-1,77	2,88
17	717	0,11	0,91	-1,81	2,71
18	706	-0,10	0,89	-1,59	2,74
19	686	0,05	0,79	-1,33	2,94
20	680	0,13	0,91	-1,96	2,78
21	693	-0,03	0,86	-1,43	2,85
22	691	-0,17	0,92	-1,86	2,63
23	714	0,00	0,87	-1,76	2,89
24	704	0,01	0,64	-0,90	2,95
25	727	0,00	0,86	-1,98	2,71
26	761	-0,06	0,82	-1,74	2,51
Média	723,23	0,00	0,87	-1,72	2,79
Mediana	723	0,00	0,89	-1,77	2,80
DP	26,70	0,09	0,07	0,29	0,17
Mínimo	680	-0,17	0,64	-2,14	2,37
Máximo	771	0,20	0,96	-0,90	3,20
Amplitude	91	0,37	0,32	1,23	0,83

O *theta_d* médio foi de 0, com desvio-padrão de 0,09. Apresentou amplitude de 0,37, ou seja, de 0,42 d.p., sendo que os estudantes que responderam ao caderno 14 obtiveram os

maiores *theta_d* médios (0,20) e os que responderam ao caderno 22 obtiveram os menores (-0,17).

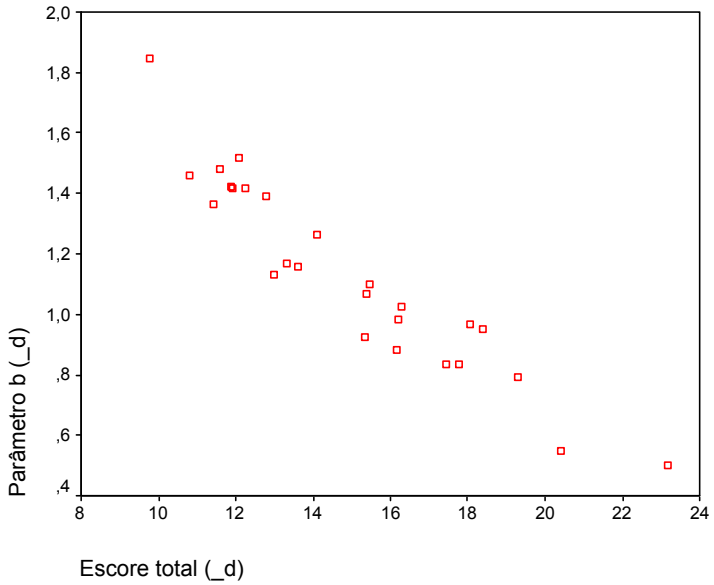
O caderno 24, que tinha se apresentado como o mais difícil para os estudantes e cujos *escores totais* tinham sido os mais baixos sistematicamente, não foi o que apresentou os menores resultados de *theta_d*. Para quinze outros cadernos, essas estimativas foram menores.

3.7 Associação entre as dificuldades dos cadernos e as habilidades dos estudantes, após a exclusão dos itens

De posse dos dois índices de dificuldade (*valor p_d* e *parâmetro b_d*) e dos dois índices de habilidade (*escore total_d* e *theta_d*) médios para cada um dos cadernos, após a exclusão dos itens, foram realizadas as seguintes correlações: o *valor p_d* com o *theta_d* e o *parâmetro b_d* com o *escore total_d* e o *theta_d*.

Relacionando o *escore_d* com o *parâmetro b_d*, é observada uma correlação alta (r = -0,95). A figura 3.5 apresenta o gráfico de dispersão entre essas duas variáveis.

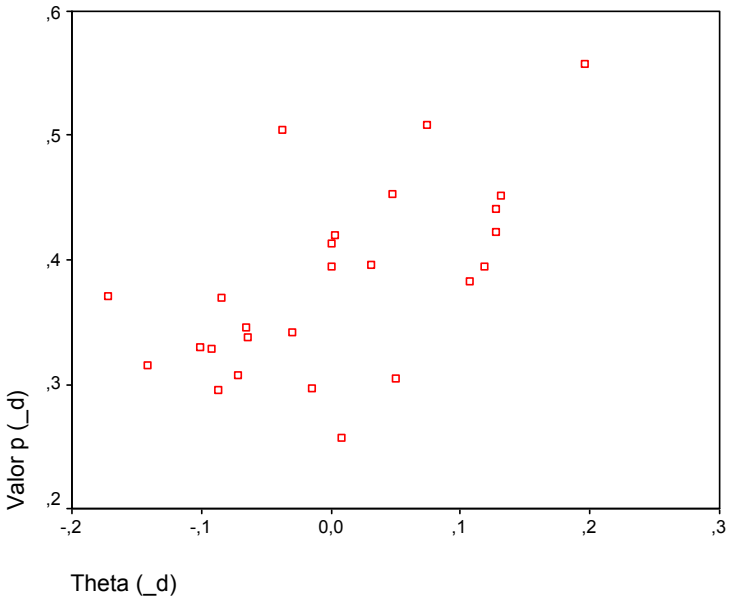
Figura 3.5 – Gráfico de dispersão entre o índice de dificuldade dos cadernos pela TRI (*parâmetro b_d*) e o *escore total(_d)* dos respondentes aos cadernos de Matemática do Saeb, após a exclusão dos itens que praticamente não contribuem com a unidimensionalidade.



Observa-se uma correlação alta e negativa na qual quanto maior o escore médio dos examinandos, menor a dificuldade do caderno. A exemplo da associação do *valor p_d* com o *escore total_d*, que é de 1, em função da própria natureza dos índices, este parece ser dependente do parâmetro de dificuldade do item estimado pela TRI (*parâmetro b_d*).

Além do *escore total_d*, investigou-se a habilidade estimada por meio da TRI, após a exclusão dos itens, quando associadas a ambos os índices de dificuldades. A Figura 3.6, a seguir, apresenta a correlação entre *valor p_d* e *theta_d*

Figura 3.6 – Gráfico de dispersão entre o índice de dificuldade pela TCT(*valor p_d*) dos cadernos e as habilidades estimadas pela TRI (*theta_d*) dos respondentes aos cadernos de Matemática do Saeb, após a exclusão dos itens.

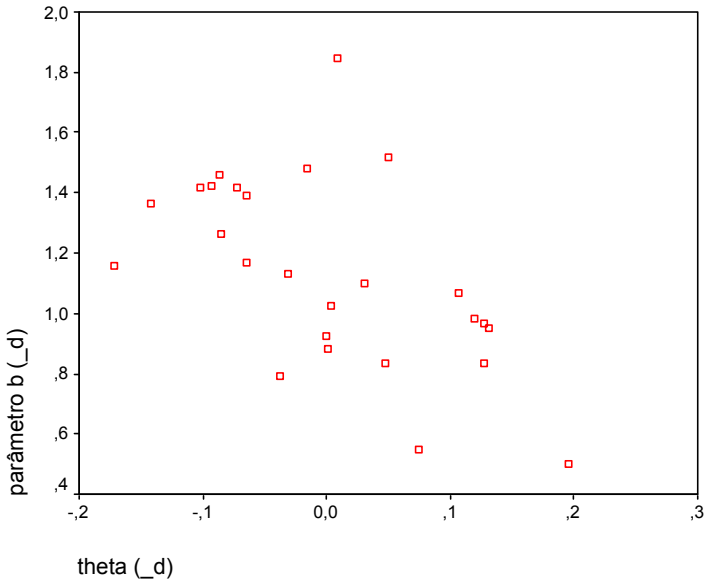


A associação entre *valor p_d* e *theta_d* forneceu um coeficiente de correlação de Pearson, r, de 0,60. É importante, a partir da análise dessas informações, lembrar que o coeficiente de correlação entre *valor p* e theta, sem a exclusão de itens, foi 0,68. Com a exclusão dos itens que praticamente não contribuem para a unidimensionalidade, a percentagem da variância de *theta* associada ao *valor p* diminuiu de 46% para 36%. Vale a pena ressaltar também que não se pode concluir que não existe dependência entre *valor p_d* e

theta_d. Embora menor do que quando se relaciona esse índice de dificuldade com o *escore total* existe ainda uma associação.

Quando o parâmetro *parâmetro b_d* é utilizado em associação com o *theta_d*, observa-se um coeficiente de correlação r de –0,57. O gráfico de dispersão a seguir (figura 3.7) ilustra essa associação. Esse resultado indica que 32% da variância de *theta_d* estão associados ao *parâmetro b_d*.

Figura 3.7 – Gráfico de dispersão entre o índice de dificuldade dos cadernos pela TRI (*parâmetro b_d*) e as habilidades estimadas pela TRI (*theta_d*) dos respondentes aos cadernos de Matemática do Saeb, após a exclusão dos itens.



Percebe-se que, embora a correlação entre as variáveis de dificuldade (*valor p_d* ou *parâmetro b_d*) com o *theta_d* seja menor que a correlação destas variáveis com o *escore total*, existe ainda uma associação forte entre elas.

Note-se, no entanto, que após a exclusão dos itens que praticamente não contribuíam para a mensuração do fator único, a correlação entre a habilidade estimada pela TRI e a dificuldade diminui de 0,68 para 0,60, quando o *valor p* é utilizado, e de –0,69 para – 0,57, quando o *parâmetro b* é utilizado. A exclusão dos itens com cargas fatoriais inferiores a 0,20

tornou a prova mais unidimensional, apontando para a importância deste pressuposto para o funcionamento do da propriedade da TRI de invariância dos parâmetros. No capítulo 4, a análise desses resultados será aprofundada.

4. Discussão e Conclusões

Pela propriedade da invariância dos parâmetros da TRI, a habilidade estimada deve ser independente do parâmetro dos itens que foram utilizados para estimá-la. Será que essa propriedade anunciada pelos teóricos da TRI é observada empiricamente? O presente estudo procurou verificar essa propriedade, com o foco no parâmetro da dificuldade e a sua influência no *theta*. Em que medida a dependência do parâmetro de habilidade com relação ao parâmetro de dificuldade dos itens que foram utilizados para estimá-lo se manifesta? Esse modelo teórico supera limitações da TCT?

Uma síntese dos resultados das correlações entre os índices de dificuldade e os de habilidade é apresentada na tabela 4.1.

Tabela 4.1 – Correlações entre os índices de dificuldade e habilidade (*).

		Índices de Dificuldade		**Índices de Habilidade**	
		Valor p	***parâmetro b***	***escore total***	***theta***
Índices de Dificuldade	***valor p***	-	-0,95	1,00	0,68
	parâmetro b	-	-	-0,95	-0,69
Índices de Habilidade	***escore total***	-	-	-	0,77
	theta	-	-	-	-

(*) Com base em 26 observações.

A correlação entre o *valor p* e o *parâmetro b* (-0,95) faz-nos concluir que estes índices de dificuldade estão fortemente associados. Praticamente e de forma sistemática, quanto maior é o *valor p* (menor é a percentagem de estudantes que, em média, acerta os itens), menor é o *parâmetro b*. Estes foram muito próximos e coerentes entre si, mesmo sendo calculados por meio de dois modelos teóricos e metodológicos diferente. Isto indica que qualquer um dos dois índices parece ser adequado como representativo da dificuldade dos cadernos.

A associação entre *valor p* e *escore total* apresenta coeficiente de correlação, por definição, de valor 1,0. A correlação entre o outro índice de dificuldade, o *parâmetro b*, e o *escore total* foi de – 0,95. Verifica-se que o índice de habilidade calculado pela TCT, o *escore total*, é associado fortemente à dificuldade dos itens que compõem os cadernos. Quase sistematicamente, quanto maior a dificuldade, menor o *escore total*, ou seja, menor a habilidade dos estudantes. Verifica-se uma extrema dependência entre a habilidade pela TCT e a dificuldade, tanto quando esta é calculada pela TCT, quando é estimada pela TRI.

Como a estimativa de habilidade da TRI, o *theta*, é anunciada como menos dependente da dificuldade da prova que o *escore total*, esperava-se uma fraca associação entre esses índices. Seria indiferente usarmos um ou outro índice, baseado pela TCT ou pela TRI, caso sua correlação fosse muito próxima de 1. O coeficiente de correlação entre *escore total* e *theta* foi de 0,77, que não pode ser considerada fraca, mas que aponta para um distanciamento dessas estimativas. Elas efetivamente estão, pelo menos parcialmente, se comportando de maneira não-perfeita.

Por sua vez, pela propriedade de invariância dos parâmetros, esperava-se que o *theta* estimado para cadernos muito fáceis ou difíceis não apresentasse uma grande variabilidade para grupos de iguais características de habilidade. Esperava-se uma nuvem de pontos que se aproximasse de uma reta vertical, representando o caso em que o *theta* dos estudantes fosse o mesmo, ainda que exista variabilidade entre as dificuldades dos cadernos. Foi observado, no entanto, uma associação entre o *theta* e os dois índices de dificuldade: o *valor p* (0,68) e o *parâmetro b* (-0,69), que indica dependência entre essa estimativa e a dificuldade.

Essa constatação empírica, embora lance dúvidas quanto a independência dos parâmetros, demonstra que a estimativa de habilidade pela TRI sofre uma influência menor com relação à dificuldade que o *escore total*, o que sugere vantagens quanto à sua utilização em situações que são utilizadas provas que podem apresentar dificuldades diferentes.

Uma análise mais detalhada dos pontos dos gráficos de dispersão (figuras 3.3 e 3.4) reforça a constatação de uma menor dependência de *theta* em relação à dificuldade. Na região central do eixo de *theta*, pode-se notar alguns pontos associados a índices de dificuldade bastante diferentes (uns muito altos outros muito baixos). Mesmo assim, nessa região, observam-se estimativas de *theta* bem semelhantes, o que eram esperadas para grupos equivalentes em termos de habilidades.

De qualquer forma, mesmo considerando essa vantagem da TRI, a completa independência dos parâmetros não foi observada. Que fatores podem estar influenciando para um distanciamento dos dados empíricos em relação à teoria?

Os resultados discutidos até agora, no entanto, foram obtidos sem o controle de certas variáveis (condições) que seriam fundamentais para a ocorrência da invariância dos parâmetros. É possível que a existência de dependência entre a dificuldade e o *theta* possa estar ocorrendo pela falta de controle de algumas dessas condições como, por exemplo, o ajuste dos dados ao modelo e a unidimensionalidade.

É importante saber, por exemplo, se o *theta*, além de menor dependência com os índices de dificuldade, está funcionando como uma boa estimativa de habilidade. Um aprofundamento do estudo da forma como o *theta* é estimado pode indicar se, além de não ser tão dependente da amostra de itens que são usados para estimá-lo (quanto o *escore total*), ele é um índice preciso para a mensuração da habilidade dos estudantes. Para estimação do *theta*, primeiramente são calculadas as proporções para cada um dos níveis de habilidade. Posteriormente, utiliza o procedimento máxima verossimilhança para ajustar uma CCI aos dados e aos valores numéricos dos parâmetros estimados. Sugere-se um aprofundamento da investigação dos aspectos matemáticos deste procedimento de ajuste dos dados ao modelo, para a qual poderá ser realizada uma revisão bibliográfica e definidos procedimentos de análise mais específicos para futuros trabalhos sobre a invariância dos parâmetros.

Por sua vez, quando se busca controlar uma outra condição para a ocorrência da invariância dos parâmetros pela TRI, a unidimensionalidade, argumenta-se como os resultados de *theta* se comportam em associação à dificuldade das provas. O estudo de Laros, Pasquali e Rodrigues (2000) indicou que a prova de Matemática da 8ª série do Saeb 97 pode ser considerada unidimensional após a exclusão dos 26 itens que contribuíam significativamente para o primeiro fator. Se a exclusão desses itens contribui para a unidimensionalidade e se o *theta* estimado tende a ser mais preciso, o que se pode concluir quanto à propriedade de invariância do parâmetro *theta*, para esses novos dados, frutos dos novos índices e estimativas? Uma das condições que Baker (2001) considerou como essencial para que ocorra a invariância do parâmetro de *theta*, com relação aos itens que são utilizados para estimá-la, procede empiricamente?

Com os índices de dificuldade e de habilidades calculados após a exclusão dos 26 itens que praticamente não contribuem para a unidimensionalidade da prova, foram obtidos os resultados de correlação apresentados na tabela 4.2.

Tabela 4.2 – Correlações entre os índices de dificuldade e habilidade, após a exclusão dos itens que não praticamente não contribuem para a unidimensionalidade (*).

		Índices de Dificuldade		**Índices de Habilidade**	
		valor p_d	***parâmetro b_d***	***escore total_d***	***Theta_d***
Índices de Dificuldade	***valor p_d***	-	-0,95	1,00	0,60
	parâmetro b_d	-	-	-0,95	-0,57
Índices de Habilidade	***escore total_d***	-	-	-	0,70
	theta_d	-	-	-	-

(*) Com base em 26 observações.

A correlação entre o *valor p_d* e o *parâmetro b_d* (-0,95) foi a mesma que a encontrada antes da exclusão dos itens que praticamente não contribuem para a mensuração do fator único. Verifica-se que estes índices de dificuldade estão fortemente associados entre

si, ou seja, constantemente quanto maior é o *valor p_d* (menor é a percentagem de estudantes que, em média, acerta os itens), menor é o *parâmetro b_d.* Os resultados destes índices foram muito próximos, mesmo sendo calculados por meio de dois modelos teóricos e metodológicos diferente, apresentando-se bastante coerentes.

A associação entre *valor p_d* e *escore total_d* apresenta coeficiente de correlação, por definição, de valor 1,0. A correlação entre o outro índice de dificuldade, o *parâmetro b_d*, e o *escore total_d* foi de – 0,95, que podemos considerar como próxima de uma correlação perfeita. Verifica-se que o índice de habilidade calculado pela TCT, o *escore total_d*, é extremamente dependente da dificuldade dos itens que compõem os cadernos, após a exclusão dos itens que praticamente não contribuem para a unidimensionalidade.

A associação da habilidade calculada pela TCT e estimada pela TRI caiu de 0,77 para 0,70 após a exclusão dos itens. Embora pareça pequena essa diferença, observa-se que a variância do *theta* associada ao *escore total* caiu de 0,59% para 0,49%. Assim, quando os itens que compõem os cadernos de prova são mais próximos da unidimensionalidade, observa-se um maior distanciamento entre o comportamento dos índices de habilidade calculados pela TCT e pela TRI, aspecto interessante para a investigação da invariância dos parâmetros.

De fato, esse distanciamento entre os índices de habilidade é refletido na correlação do *theta* com os índices de dificuldade. Com a exclusão dos 26 itens, a correlação do *theta* com os índices de dificuldade diminuiu de 0,68 para 0,60, no caso da correlação da dificuldade calculada pela TCT, e de – 0,69 para – 0,57, no caso da correlação da dificuldade estimada pela própria TRI, embora ainda se mostrem associadas. Isso significa que a percentagem da variância de *theta* associada à dificuldade calculada pela TCT caiu de 46% para 36% e a associada à dificuldade estimada pela TRI caiu de 48% para 32%, uma significativa oscilação, principalmente para o segundo caso.

O pressuposto de unidimensionalidade dos itens parece ser fundamental para a manifestação da propriedade de invariância dos parâmetros. Esta propriedade funciona melhor quando o atendimento desta condição está mais próximo. Quando itens com cargas fatoriais no fator único inferiores a 0,20 são excluídos, é observada uma queda significativa da percentagem da variância de *theta* associada à dificuldade, podendo-se considerar que é de extrema importância a verificação desse pressuposto antes da estimação das habilidades pela TRI. As estimativas com base em itens que estejam avaliando o mesmo traço latente tendem a ser mais precisas e a propriedade de invariância dos parâmetros tende a funcionar empiricamente melhor.

Esses achados sugerem que, se o controle da unidimensionalidade for ainda mais rigoroso, os resultados de invariância tendem a ser ainda melhores. Desta forma, sugere-se uma nova estimação dos parâmetros dos itens e da habilidade, excluindo, agora, os itens que não contribuíam para o fator único com pelo menos uma carga fatorial de 0,30. Espera-se, a partir, dessa nova análise que as correlações entre *theta* e as dificuldades diminuam ainda mais. O estudo de Laros, Pasquali e Rodrigues (2000) novamente será importante para a o aprofundamento do presente estudo, visto que apresentam as informações de cargas fatoriais desta prova e a indicação dos itens que deveriam ser eliminados da prova.

Não foi realizado um estudo da importância do ajuste dos modelos aos dados como condição para a propriedade de invariância dos parâmetros. Próximos estudos poderão considerar essa variável, com base no que Fernandez (1990) considera em uma de suas obras: "Se o modelo se ajusta estritamente aos dados, os objetivos da invariância dos parâmetros se cumprem".

É certo que os resultados indicam para uma menor dependência de *theta* em relação à dificuldade da prova, em comparação com o *escore total*. Ela diminuiu quando a condição de unidimensionalidade foi, pelo menos em parte, atendida. Poderá diminuir ainda mais se a

observação desta condição for mais rigorosamente controlada. Essas informações, por sua vez, remetem a algumas discussões sobre a metodologia utilizada pelo Saeb para análise de seus resultados.

Cabe ressaltar que esse estudo "forçou" um delineamento em que grupos de estudantes com características semelhantes em termos de habilidades respondem à provas de dificuldades diferentes e apresentam determinadas estimativas médias de *theta*. Na análise dos resultados do Saeb, esta habilidade não é estimada com base em cada um dos cadernos e sim com base em todos os itens da prova (150, por exemplo). Também não é estimada com base em cada um dos 26 grupos de estudantes, mas com base em todos os estudantes concomitantemente.

O Saeb já utilizou procedimentos de coleta de dados a partir de provas clássicas de 30 itens. O instrumento, neste formato, é por demais limitado pois, primeiramente, não se consegue uma ampla cobertura curricular. Quando conseguirmos com esse pequeno número de itens avaliar algumas das principais habilidades e conteúdos em Matemática. Além disso, sabe-se que mesmo com todo o processo de validação de itens, existe uma perda de itens por comportamento inadequado na aplicação final da prova, o que acarretaria em um número ainda mais inferior de itens na prova. Considera-se desta forma, que a opção por utilização de um número aproximado de 150 itens para avaliação das habilidades dos estudantes em Matemática já fornece um avanço significativo à avaliação desse traço latente. Quando o Saeb tem como decisão avaliar um espectro maior da habilidade em uma determinada disciplina, deve propor alternativas de análises para resultados de diferentes estudantes que respondem a diferentes formas de prova. Uma excelente alternativa é a TRI. Como foi verificado neste estudo, embora não se tenha uma independência entre o *theta* e a dificuldade, o Saeb pode contar com uma estimativa mais independente da dificuldade que se fossem utilizados os resultados de *escore total*.

Além disso, já que o Saeb utiliza uma grande quantidade de itens para estimar a habilidade dos estudantes, um maior rigor na consideração do pressuposto de unidimensionalidade é aqui sugerido como uma excelente oportunidade para propiciar um *theta* mais independente da dificuldade das provas administradas. Isso porque é possível a exclusão de um número razoável de itens que praticamente não contribui para a avaliação do fator único, sem grandes implicações negativas na precisão do instrumento geradas quando se tem disponível um número reduzido de itens. Pelo contrário, certamente a precisão do instrumento melhora quando são considerados apenas aqueles que estão avaliando o fator único.

Considera-se, finalmente, que os resultados apresentados e discutidos na presente dissertação contribuem para o estudo da propriedade de invariância dos parâmetros e colabora para diminuir a escassez de estudos empíricos sobre o tema, como indica Fan (1998).

5. Bibliografia

Baker, F. B. (2001). *The basics of item response theory*. USA: Eric Clearinghouse on Assessement and Evaluation. Second edition.

Bock, R. D.l, & Zimowski, M. F. (1995). Multiple group IRT. In W. van der Linden & R. Hambleton (Eds.), *Handbook of item response theory*. New York: Springer Verlag.

Brogan, D. J. (1997). Pitfalls of using standard statistical software packages for samples survey data. In *Encyclopedia of Biostatistics*. Atlanta: Emory University.

Condé, F. N., & Rabello, G. C. (2001). A invariância dos parâmetros na teoria de resposta ao item: um estudo com os dados do Saeb. *Anais do marco de aprendizagem contínua em avaliação.* Salvador: Dez/2001.

Fan, X. & Pin, Y. (1999). Assessing the effect of model-data misfit on the invariance property of IRT parameter estimates. *Paper presented at the 1999 annual meeting of the american educational research association, april 19-23*, Montreal, Canada (Session # 38.05).

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58,* 357-381.

Fernandez, J. M. (1990). *Teoria de Respuesta a los ítems: un nuevo enfoque en la evolución psicológica y educativa*. Madrid: Ediciones pirâmide.

Hambleton, R.K., Swaminathan, H. e Rogers, H.J. (1991). *Fundamentals of item response theory: measurement methods for the social sciences*. Newbury Park, CA: SAGE publications, Inc.

Hattie, J.A. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Instituto Nacional de Estudos e Pesquisas Educacionais (1998). *Relatório Técnico da Amostra do Saeb 97*. Brasília: INEP.

Instituto Nacional de Estudos e Pesquisas Educacionais (2001). *Guia para elaboração e revisão de itens*. Brasília: INEP.

Instituto Nacional de Estudos e Pesquisas Educacionais (2002). *Saeb 2001: novas perspectivas*. Brasília: INEP.

Klein, R. & Klein, T. S. (1998). *Programa para Teoria Clássica dos Testes*.

Kvanli, A.H., Guynes, C.S., & Pavur, R.J. (1991). *Introduction to business statistics* (4a ed.). USA: West Publishing Company.

Laros, J.A. (2001). *Diferenças entre estados em escores gerais e em escores de temas e tópicos das provas do Saeb 1999 em matemática e português para a 4a série do ensino fundamental*. Brasília: Centro de Pesquisa em Avaliação Educacional – CPAE, UnB.

Laros, J.A., Pasquali, L. & Rodrigues, M.M.M (2000). *Análise da unidimensionalidade das provas do Saeb*. Brasília: Centro de Pesquisa em Avaliação Educacional – CPAE, UnB.

Lord, F.M. (1980). *Aplications of item response theory to practical testing problems*. Hillsdale (NJ): Lawrence Erlbaum.

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). USA: McGraw-Hill.

Pasquali, L. (1996). *Teoria e métodos de medida em ciências do comportamento*. Brasília: Laboratório de Pesquisa em Avaliação e Medida/Instituto de Psicologia/Universidade de Brasília/Instituto Nacional de Estudos e Pesquisas Educacionais.

Pasquali, L. (1997). *Psicometria: teoria e aplicações*. Brasília: Editora Universidade de Brasília.

Pestana, M.I.G.S. (1997). *Matrizes curriculares de referência para o SAEB*. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais.

Pestana, M.I.G.S. (1999a). *Matrizes curriculares de referência para o SAEB*. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais.

Pestana, M.I.G.S. (1999b). *Saeb 97: primeiros resultados*. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais.

Rabello, G.C. (2001). A técnica de equalização: um estudo comparativo com os dados do SAEB. *Dissertação de mestrado*. Brasília: Universidade de Brasília.

Requena, C.S. (1990). *Psicometria: teoria y práctica en la construcción de tests*. Madrid: Ediciones Norma, S.A.

Riether, M.M. e Rauter, R (2000). A Metodologia de amostragem do Saeb. *Revista brasileira de estudos pedagógicos, 81(197)*, 143-153.

Rodrigues, M.M.M. (2002). Instrumentos de avaliação educacional: uma visão pedagógica e psicométrica integradas: estudos das provas do Saeb, matemática 8ª série, 1997 e 1999. *Dissertação de mestrado*. Brasília: Universidade de Brasília.

SAS Institute Inc. (1993). *SAS companion for the microsoft windows environment, version 6*. SAS Institute Inc., Cary, N.C.

Shah. B.V., Barnwell B.G. and Bieler G.S. (1996). *SUDAAN user's manual: release 7.0*, Research Triangle Institute, Research Triangle Park, N.C.

Shaughnessy, J.J., Zechmeister E.B., & Zechmeister, J.S. (2000). *Research methods in Psychology.* Boston: McGraw-Hill Companies.

Siegel, S. (1975). *Estatística não-paramétrica para ciências do comportamento*. São Paulo: McGraw-Hill.

SPSS (1999). *SPSS base 10.0 applications guide*. USA: SPSS Inc.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *BILOG-MG: multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software International (SSI).

6. Anexo I: Diferenças entre os índices de escore total dos examinandos por bloco

Bloco	Posição	Caderno	N	*escore total*				Blocos de mesma posição		Blocos independentemente da posição	
				Média	d.p.	Mín	Máx	Diferença entre médias	Diferença entre médias normalizada	Diferença entre médias	Diferença entre médias normalizada
1	1	1	757	7,39	2,22	0	11				
		14	688	7,31	2,19	0	11	0,08	0,04		
	2	13	725	6,80	2,33	0	11				
		25	727	6,84	2,19	0	11	0,04	0,02		
	3	10	771	6,71	2,23	0	11				
		20	680	6,95	2,18	0	11	0,24	0,11	0,68	0,30
2	1	2	731	3,73	2,18	0	10				
		15	707	3,81	2,11	0	10	0,08	0,04		
	2	1	757	3,86	2,21	0	10				
		26	761	3,44	1,75	0	10	0,42	0,21		
	3	11	722	3,66	1,99	0	10				
		21	693	3,36	2,03	0	10	0,30	0,15	0,50	0,24
3	1	3	705	6,26	2,31	0	11				
		16	698	6,59	2,32	0	11	0,33	0,14		
	2	2	731	6,18	2,34	0	11				
		14	688	6,59	2,47	0	11	0,41	0,17		
	3	12	757	5,63	2,55	0	11				
		22	691	5,30	2,57	0	11	0,32	0,13	1,29	0,53
4	1	4	742	7,83	2,41	0	11				
		17	717	7,57	2,56	0	11	0,26	0,11		
	2	3	705	7,71	2,54	0	11				
		15	707	7,81	2,51	0	11	0,11	0,04		
	3	13	725	6,97	2,66	0	11				
		23	714	6,91	2,62	0	11	0,06	0,02	0,92	0,36
5	1	5	762	4,47	2,29	0	12				
		18	706	4,49	2,57	0	12	0,02	0,01		
	2	4	742	4,54	2,13	0	12				
		16	698	4,73	2,60	0	12	0,18	0,08		
	3	1	757	4,56	2,15	0	12				
		24	704	3,84	2,09	0	11	0,72	0,34	0,89	0,39
6	1	6	752	4,13	2,93	0	12				
		19	686	4,66	2,61	0	13	0,52	0,19		
	2	5	762	4,06	2,78	0	13				
		17	717	4,76	2,78	0	13	0,70	0,25		
	3	2	731	4,19	2,77	0	13				
		25	727	4,32	2,42	0	13	0,13	0,05	0,70	0,26

Bloco	Posição	Caderno	N	*escore total*				Blocos de mesma posição		Blocos independentemente da posição	
				Média	d.p.	Mín	Máx	Diferença entre médias	Diferença entre médias normalizada	Diferença entre médias	Diferença entre médias normalizada
7	1	7	735	4,61	2,27	0	13				
		20	680	5,53	2,60	0	13	0,93	0,38		
	2	6	752	5,30	2,48	0	13				
		18	706	4,72	2,52	0	13	0,58	0,23		
	3	3	705	5,06	2,44	0	13				
		26	761	4,96	2,36	0	13	0,10	0,04	0,93	0,38
8	1	8	744	3,76	2,20	0	12				
		21	693	4,00	2,31	0	12	0,24	0,11		
	2	7	735	3,47	2,31	0	12				
		19	686	3,85	2,17	0	11	0,38	0,17		
	3	4	742	3,68	2,22	0	12				
		14	688	4,05	2,22	0	12	0,37	0,17	0,57	0,26
9	1	9	729	3,36	2,24	0	11				
		22	691	3,43	2,20	0	11	0,07	0,03		
	2	8	744	3,32	1,95	0	10				
		20	680	3,73	2,23	0	11	0,41	0,19		
	3	5	762	3,02	2,22	0	11				
		15	707	3,71	2,20	0	11	0,70	0,32	0,71	0,33
10	1	10	771	4,25	2,91	0	13				
		23	714	4,37	2,83	0	13	0,11	0,04		
	2	9	729	3,86	2,71	0	13				
		21	693	4,14	2,73	0	13	0,28	0,10		
	3	6	752	3,88	2,98	0	13				
		16	698	4,12	3,11	0	13	0,24	0,08	0,50	0,17
11	1	11	722	2,57	1,78	0	12				
		24	704	2,51	1,56	0	11	0,06	0,04		
	2	10	771	2,74	1,73	0	12				
		22	691	2,65	1,79	0	10	0,09	0,05		
	3	7	735	2,60	1,56	0	10				
		17	717	2,51	1,70	0	10	0,09	0,05	0,23	0,14
12	1	12	757	3,40	2,00	0	11				
		25	727	3,33	1,97	0	12	0,07	0,03		
	2	11	722	3,23	1,92	0	13				
		23	714	3,30	1,97	0	11	0,07	0,04		
	3	8	744	3,39	1,93	0	11				
		18	706	3,04	2,08	0	12	0,35	0,17	0,35	0,18
13	1	13	725	3,08	1,93	0	10				
		26	761	2,77	1,56	0	11	0,32	0,18		
	2	12	757	3,09	1,75	0	10				
		24	704	2,87	1,74	0	11	0,22	0,13		
	3	9	729	2,48	1,60	0	10				
		19	686	2,49	1,66	0	10	0,01	0,01	0,61	0,36