Universidade de Brasília

Instituto de Psicologia

Curso de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações

# RELAÇÃO ENTRE CARACTERÍSTICAS DO TESTE EDUCACIONAL E ESTIMATIVA DE HABILIDADE DO ESTUDANTE

Frederico Neves Condé

Brasília, DF

2008

Universidade de Brasília

Instituto de Psicologia

Curso de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações

# Relação entre características do teste educacional e estimativa de habilidade do estudante

Frederico Neves Condé

Brasília, DF

2008

Universidade de Brasília

Instituto de Psicologia

Curso de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações

# Relação entre características do teste educacional e estimativa de habilidade do estudante

Frederico Neves Condé

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações, como requisito parcial à obtenção do grau de Doutor em Psicologia Social e das Organizações

Orientador: Jacob Arie Laros

Brasília, DF

Outubro de 2008

Relação entre características do teste educacional e estimativa de habilidade do estudante

Tese de Doutorado defendida diante e aprovada pela banca examinadora constituída por:

---

Prof. Jacob Arie Laros, Ph.D. (Presidente)

Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações

---

Prof. Bartholomeu Tôrres Tróccoli, Ph.D.

Instituto de Psicologia da Universidade de Brasília.

---

Prof. Dr. Héliton Ribeiro Tavares

Diretoria de Avaliação da Educação Básica do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP

Departamento de Estatística da Universidade Federal do Pará.

---

Prof. Joaquim José Soares Neto, Ph.D.

Núcleo de Pesquisa e Avaliação do Centro de Seleção e de Promoção de Eventos - CESPE

Instituto de Física da Universidade de Brasília.

---

Prof. Luiz Pasquali, Docteur

Instituto de Psicologia da Universidade de Brasília.

---

Dr. Marcos Ruben de Oliveira (Suplente)

Banco Central do Brasil

Dedico o trabalho para

**Arthur**, meu Filhão Flamenguista. Penso em você pela sua presença, não pela falta que me faz.

**Nanda**, minha linda Nanda. Este trabalho é seu. Obrigado por todo amor, apoio e companheirismo nesse período de nossas vidas.

**Mãe, pai, Fabrício e Fabiano**, juntos sempre.

Agradecimentos

Ao Professor e Orientador Jaap Laros, que sempre me incentivou na realização de pesquisas na área de avaliação. Ensinou-me muito desde a época do PROAV, com os estudos sobre a dimensionalidade e forneceu-me o conhecimento e a confiança necessária à realização do mestrado e do doutorado.

Ao Prof. Luiz Pasquali. Só estou nesse ramo hoje em função do Pasquali. A paixão pela área de medidas em psicologia veio nas disciplinas TEP e psicometria na graduação e nas pesquisas que realizei como bolsista do LabPAM na área de avaliação do Temperamento. Seu carisma permitiu transformar o estudo em algo realizador, em função da clareza e da paixão com que trata o conhecimento científico.

Aos membros da banca de doutoramento Bartholomeu Tôrres Tróccoli, Héliton Ribeiro Tavares, Joaquim José Soares Neto e Marcos Ruben de Oliveira por todo apoio oferecido para a consecução de meu doutoramento e, principalmente, pelas oportunidades que tive em atuar profissionalmente com todos eles.

À Professora Amélia Regina Alves, desde a época da TELEBRÁS. Amiga que me possibilitou assimilar um conjunto de conhecimentos e de preceitos éticos que balizaram toda minha formação e desenvolvimento profissional.

Aos amigos Guilherme Coelho Rabello, Eduardo de São Paulo e Robson Medeiros de Araújo, parceiros no desenvolvimento do presente trabalho. Extremamente presentes em minha trajetória acadêmica e profissional, agradeço-lhes toda colaboração e amizade.

Ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, que, não só me disponibilizou as bases de dados, mas forneceu todo o suporte técnico necessário para o desenvolvimento deste estudo. Em especial, agradeço a Héliton Ribeiro Tavares, Amaury Patrick Gremaud, Luiza Massae Uema, Maria Cândida Lacerda Muniz Trigo, Maria Inês Pestana, Iza Locatelli, Maria Alejandra Schulmeyer Iriarte, Teófilo Francisco de Paula e Elaine Cristina Sampaio Castelo Branco Barros, extremamente presentes em minha trajetória de quase dez anos de DAEB.

Ao Instituto de Educação Superior de Brasília - IESB, minha instituição de ensino. Em especial, agradeço à Profª. Eda Coutinho B. Machado, Prof. João Cláudio Todorov, Profª. Gláucia Melasso Garcia de Carvalho, Prof. Teobaldo Rivas, Profª Graziela Furtado Scarpelli Ferreira, Prof. Márcio Borges Moreira e a todos os professores e alunos do curso de Psicologia.

Sumário

Lista de tabelas

Lista de Figuras

Resumo

O presente estudo teve como objetivo verificar a relação de características de testes educacionais de matemática e a validade e a fidedignidade das habilidades de estudantes estimadas por meio da Teoria de Resposta ao Item (TRI). Estudos prévios mostraram que dois testes de matemática aplicados em 2005 a estudantes de 8ª série do Ensino Fundamental, a ANEB, contendo 155 itens, e a Prova Brasil, contendo 81 itens, apresentaram resultados de estimativas de habilidade diferentes para grupos com características semelhantes. Esses resultados não foram os esperados, já que a TRI, teoricamente, permite a estimação das habilidades dos estudantes independentemente das características do teste, uma vez que seus pressupostos são atendidos. O grau de cobertura da matriz de referência e os parâmetros psicométricos dos testes foram analisados para subsidiar a composição de testes simulados. Utilizando o teste ANEB como referência, quatro testes (formas A, B, C e D) foram simulados com diferentes números de itens (104 e 81) itens, a partir da variação de seu grau de dificuldade e de discriminação. As estimativas de habilidade dos estudantes foram comparadas entre os testes originais ANEB e Prova Brasil e entre os quatro testes simulados. Evidências de validade e de fidedignidade foram investigadas. Resultados revelaram que estudantes que responderam à Prova Brasil obtiveram estimativas de habilidade maiores em 0,2 desvios-padrão que estudantes que responderam ao teste ANEB. Essa diferença, significativa ao nível de 5%, não pode ser explicada em função de baixo grau de validade de um dos testes, já que foram encontradas evidências de bom grau de validade para ambos os testes quanto às características: grau de cobertura da matriz de referência, elaboração e revisão de itens, análise pedagógica e análise de Funcionamento Diferencial do Item. Os resultados de análise unidimensionalidade podiam ter sido utilizados para decidir quais itens seriam considerados para estimar as habilidades dos estudantes e serviriam como uma evidência adicional de validade. A comparação ente quatro testes simulados e os testes originais indicaram que o número de itens dos testes respondidos pelos estudantes, a qualidade discriminativa dos itens e a relação do parâmetro *b* com o parâmetro de habilidade são acompanhados de um aumento da fidedignidade dos testes. Os resultados da investigação sugerem que as diferenças observadas quanto às estimativas de habilidade entre ANEB e Prova Brasil estão associadas ao pequeno número de itens discriminativos para estudantes com estimativas baixas e médias. Com base nos resultados do estudo, recomenda-se a inclusão, nos testes de matemática, 8ª série, de um número maior de itens discriminativos para as faixas baixa e média de habilidades estimadas. Os resultados podem auxiliar o

Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) na composição de testes para os próximos processos avaliativos.

Palavras-chave: Construção de testes, Teoria de Resposta ao Item, Validade, Fidedignidade, SAEB, Prova Brasil.

The relation between characteristics of educational tests assessing Mathematics and the estimation of student's proficiency

Frederico Neves Condé

Abstract

The main purpose of this investigation was to verify the relation between psychometric properties of educational tests assessing Mathematics and the validity and reliability of the estimated proficiencies of students using Item Response Theory (IRT). Earlier studies showed that two equalized Mathematics tests applied in 2005 to students attending the 8th grade of basic education, one containing 155 items (*ANEB*), and the other containing 81 items (*Prova Brasil*) presented different outcomes for the estimated proficiencies of groups of students with similar characteristics. These results were not expected considering the fact that IRT theoretically permits the estimation of students' proficiency independent of the characteristics of a test, once the assumptions underlying the IRT model are satisfied. The degree of coverage of the reference matrices and the psychometric parameters of the two Mathematics tests were analyzed in order to obtain a basis for the creation of simulated test forms. Taking the *ANEB* test as point of reference, four tests (forms A, B, C, and D) were simulated with different number of items (104 or 81) and displaying varying degrees of difficulty and discrimination. Estimated students´ proficiencies were compared among the original *ANEB* and *Prova Brasil* tests and among the four simulated test forms. Also indications of validity and reliability were compared. Results from this comparison revealed that students who took the *Prova Brasil* received a higher estimated proficiency than the students who took the *ANEB* test. The difference amounted to .2 standard deviations and was significant at the 5% level. This observed difference can´t be explained by poor validity of one of the measuring instruments because both tests present evidence of good validity based on the following characteristics: degree of coverage of the reference matrices, elaboration and revision of the items, pedagogical analysis and analysis of Differential Item Functioning. Results of the unidimensionality analysis can be used to decide which items should be included to esteem the proficiency of the students, and serve as an additional indication of the validity of the tests. Comparing the four simulated test and the original tests indicated that the number of items answered by the students, the discrimination quality of the items and the relation of the *b* parameter with the estimated proficiency are accompanied by an increase of the reliability of the tests. The results of this investigation suggest that the observed difference in the estimation of students' proficiency of the *ANEB* and *Prova Brasil* test is related to the low number of highly discriminating

items for students with low and medium proficiences. Based on the results of this study it is recommended to include in tests assessing Mathematics for 8th grade students of basic education a greater number of good discriminating items for the low and medium proficiences in Mathematics. The results of this study can assist INEP, the National Institute for Educational Research of Brazil, in the composition and elaboration process of future tests.

Key-words: Test Construction, Item Response Theory, Validity, Reliability, SAEB, Prova Brasil.

# Índice

## 1. Introdução

Programas educacionais de âmbito governamental têm como objetivo promover uma educação com qualidade e eqüidade tendo em vista a demanda da sociedade e a formação de seus cidadãos. Geralmente estão associados a sistemas avaliativos com a função de monitoramento de sua efetividade e eficácia. Segundo esta perspectiva, avaliação é entendida como "qualquer método de obtenção de informações oriundas de testes e de outros instrumentos utilizadas para realizar inferências sobre características de pessoas, objetos e programas" (AERA, APA & NCME, 1999, p. 172).

Barreto e Pinto (2001), após análise da produção acadêmica sobre avaliação da educação básica no Brasil na década de 90, constaram a predominância de produções com foco na discussão sobre teorias e metodologias acerca da avaliação da aprendizagem. Identificaram basicamente ensaios, sem grande pretensão empírica, explorando conceitos, modelos teóricos, pressupostos e alguma produção sobre aspectos técnico-metodológicos relativos à avaliação. Observaram uma evidente preocupação com o significado da avaliação educacional em nosso contexto.

As autoras identificaram um subgrupo dos estudos que abordava os modelos de avaliação em larga escala, sobre trajetória escolar, desenvolvimento cognitivo dos alunos e modelos de monitoramento de redes de ensino e avaliação dos resultados de aprendizagem dos estudantes, denominados avaliação de monitoramento. "Avaliação de monitoramento (...) é entendida como a avaliação padronizada do rendimento escolar dos alunos, realizada no âmbito dos sistemas nacionais ou estaduais de avaliação do ensino básico" (Barreto & Pinto, 2001, p. 49). Sobre o tema, foram identificados artigos que tratavam de diferentes tópicos: (a) medida da qualidade da educação, por meio do estabelecimento de mecanismos de quantificação dos produtos do processo educativo; (b) bases para o desenho de instrumentos de medida da qualidade educativa; (c) mensuração sistemática como meio de fornecer informações para a avaliação, para o desenvolvimento de uma cultura avaliativa e servir de base ao monitoramento do sistema educacional com o objetivo de melhoria de sua qualidade; e (d) gerenciamento do sistema de avaliação e sua implementação.

A partir da década de 1990, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) do Ministério da Educação (MEC) implementou o Sistema Nacional de Avaliação da Educação Básica (SAEB). Trata-se de uma avaliação em larga escala do desempenho dos estudantes brasileiros, bem como de fatores associados a esse desempenho, que impactam na qualidade da educação.

Realizado nos anos 1990, 1993, 1995, 1997, 1999, 2001, 2003, 2005 e 2007, o SAEB 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 avalia, dentre outros aspectos, o nível de desempenho dos estudantes de 4ª e 8ª séries do Ensino Fundamental (EF) e de 3ª série do Ensino Médio (EM) em diversas disciplinas, a partir da aplicação de testes educacionais: língua portuguesa e matemática, para todas as edições do SAEB; ciências da natureza (química, física e biologia), avaliadas pelo SAEB 97 e pelo SAEB 99; história e geografia, avaliadas pelo SAEB 99.

Esse sistema de avaliação de monitoramento, a partir de 1995, assumiu um delineamento de composição dos testes e distribuição de cadernos aos respondentes por Blocos Incompletos Balanceados – BIB (Bekman, 2001; Johnson, 1992). O desenho permite que cada grupo de estudantes responda a cadernos de teste diferentes e que um maior número de itens de teste seja utilizado, de tal forma que o cálculo das habilidades dos estudantes possa contemplar, de forma válida, uma ampla matriz de referência com os conteúdos e os domínios cognitivos avaliados.

O uso do BIB, no caso do SAEB, está associado à estimação das habilidades (do desempenho) dos estudantes de acordo com a Teoria de Resposta ao Item (TRI) sob o modelo logístico de três parâmetros (Baker, 2001; Cronbach, 1996; Hambleton & Jones, 1993; Hambleton, Swaminathan & Rogers, 1991; Lord, 1980; Pasquali, 2003). As habilidades são estimadas e apresentadas em uma escala que varia de 0 a 500 pontos, comum entre anos e séries para cada disciplina, de forma a possibilitar a construção de uma série histórica e permitir a comparação entre as séries. A escala foi construída utilizando-se como grupo de referência a 8ª série do SAEB 1997 de cada disciplina, com média 250 e desvio-padrão (DP) de 50.

De 1995 a 2003, o SAEB tinha caráter amostral e utilizava, geralmente, testes de 169 itens, divididos em 26 cadernos de 39 itens, em que cada estudante respondia a um único caderno composto por três blocos de itens. Em 2005, o SAEB foi dividido em dois processos de avaliação: (a) a Avaliação Nacional da Educação Básica (ANEB) (D.O.U., n.100, Portaria n. 89, de 25 de maio de 2005) e (b) a Avaliação Nacional do Rendimento

Escolar (ANRESC) (D.O.U., n.85, Portaria n. 69, de 4 de maio de 2005), denominada posteriormente de Prova Brasil.

Similarmente ao modelo tradicional do SAEB, a ANEB 2005 foi aplicada em uma amostra de estudantes da 4ª e 8ª séries EF e da 3ª série EM das zonas rural e urbana e das redes federal, estadual, municipal e particular e não emitiu resultados por municípios e escolas. Ademais, utilizou testes compostos por 169 itens, arranjados em 26 cadernos de 39 itens, de forma que cada aluno respondeu a um único caderno de uma única disciplina.

Já a Prova Brasil 2005 emitiu resultados por escola e foi aplicada de uma forma mais universalizada, programada para todos os estudantes das 4ª e 8ª séries EF de escolas públicas e urbanas com mais de 30 alunos. Utilizou 70 itens para 4ª série EF e 84, para 8ª série EF, por disciplina, a partir da combinação de 7 blocos, dois a dois. Cada aluno respondeu a um único caderno composto por 20 itens de língua portuguesa e 20 de matemática, para 4ª série EF, e 24 itens de cada disciplina para a 8ª série EF.

As aplicações da ANEB 2005 e da Prova Brasil 2005 ocorreram praticamente na mesma época, com uma diferença de cerca de um mês, e avaliaram estudantes em comum: uma parcela de alunos de 4ª e 8ª séries EF de escolas públicas e urbanas de escolas com mais de 30 alunos. Esperava-se que os resultados desse grupo de estudantes fossem semelhantes, pois ambas as avaliações utilizaram:

a) mesmo referencial teórico, avaliando um construto igual e sob as mesmas matrizes de referência;

b) idênticas especificações para construção dos itens (múltipla escolha de quatro ou cinco alternativas; mesmas regras para construção e revisão);

c) testes e distribuição baseados no delineamento BIB.

d) itens comuns com o SAEB 2003 para permitir a estimação das habilidades equalizadas na escala do SAEB;

e) procedimentos de aplicação bastante semelhantes: distribuição dos cadernos entre os alunos, instruções e tempo médio de resposta por item em torno de 2 minutos.

f) mesma teoria e modelo de estimação das habilidades (TRI; três parâmetros).

As principais diferenças estruturais entre a ANEB e a Prova Brasil referiram-se à estrutura dos testes:

a) número de itens no teste (ANEB: 169; Prova Brasil: 70 ou 84);

b) número de cadernos (ANEB: 26; Prova Brasil: 21);

c) número de blocos (ANEB: 13; Prova Brasil: 7);

d) número de itens que cada aluno responde (ANEB: 39; Prova Brasil: 40 ou 48);

e) número de disciplinas contempladas em cada caderno (ANEB: 1; Prova Brasil: 2).

Tais elementos, referentes às características dos testes, por sua vez, não deveriam impactar em diferenças de nível de habilidade para grupos com características semelhantes, já que a TRI foi utilizada para a sua estimação. Para o caso em que os dados se ajustam ao modelo, a TRI pressupõe a propriedade de invariância dos parâmetros que afirma que as habilidades dos sujeitos são estimadas independentemente do teste utilizado. Assim como os parâmetros dos itens, independentemente da amostra de examinandos que os responderam (Baker, 2001; Hambleton, Swaminathan & Rogers, 1991).

Condé (2007) e Rabello (2007) compararam as estimativas de habilidade dos estudantes submetidos aos testes da ANEB 2005 e da Prova Brasil 2005. De modo geral, observaram uma proximidade entre as médias estimadas para as avaliações, principalmente para matemática, 4ª série EF, e para língua portuguesa, 8ª Série EF. Para algumas séries e disciplinas e para certos grupos de comparação, no entanto, um conjunto de médias da Prova Brasil se distanciou do limite inferior ou superior do intervalo de confiança de 95% calculado para as médias do SAEB.

Os autores observaram, em nível Nacional, que as médias de língua portuguesa, 4ª série EF, e de matemática, 8ª série EF, da Prova Brasil 2005 extrapolaram o intervalo de confiança de 95% calculado para a ANEB 2005. Implica dizer que, se para a primeira série e disciplina a diferença não foi tão expressiva, para matemática, 8ª série EF, encontrou-se uma diferença superior a cinco pontos da escala do SAEB, o que equivale a 0,10 desvios-padrão (DP) com referência ao limite do intervalo de confiança.

Quando as comparações entre as médias foram realizadas para as Regiões Brasileiras, os resultados de habilidades estimadas dos estudantes foram semelhantes aos encontrados para o Brasil. Em matemática, 8ª série EF, e para todas as Regiões foram observadas diferenças significativas entre os resultados da Prova Brasil e da ANEB, coerentemente aos encontrados em nível Brasil (Condé, 2007; Rabello, 2007). De modo geral as médias da Prova Brasil 2005 para matemática 8ª série EF, tanto em nível Brasil, quanto para Regiões, foram superiores às médias da ANEB 2005.

Que fatores relacionados ao teste teriam influenciado na diferença entre os resultados dos estudantes de 8ª série EF na ANEB e na Prova Brasil? Supõe-se, por meio de uma análise preliminar, que características relacionadas ao teste estejam superestimando os resultados da Prova Brasil, já que, de maneira geral, os demais aspectos envolvidos nas avaliações são semelhantes. Essa suposição, a princípio, é inconsistente com a propriedade

de invariância do parâmetro de habilidade da TRI (Baker, 2001; Hambleton, Swaminathan & Rogers, 1991).

Condé (2002) e Condé e Laros (2007) verificaram que a propriedade de invariância do parâmetro de habilidade procede apenas no caso em que o teste se aproxima da unidimensionalidade, para modelos TRI unidimensionais. Assim, testes com dificuldades diferentes fornecem resultados de habilidade diferentes para grupos equivalentes quando se distanciam da unidimensionalidade. No caso da distância entre os resultados da ANEB 2005 e Prova Brasil 2005, para matemática 8ª série EF, é possível que a diferença de dificuldades de seus itens, associada ao distanciamento do fator único de pelo menos um dos testes, esteja gerando estimativas de habilidade da TRI dependentes da dificuldade.

Laros, Pasquali e Rodrigues (2000) sugeriram que o distanciamento da unidimensionalidade pode ser minimizado caso se excluam itens com baixas cargas fatoriais no fator principal. É possível supor que os resultados diferentes entre ANEB e Prova Brasil decorrem da existência, em algum dos dois testes, de itens com baixas cargas fatoriais associadas ao fator único. A exclusão de itens com essas características e a reestimação das habilidades podem aproximar os resultados entre as avaliações.

A despeito da propriedade de invariância do parâmetro de habilidade, questiona-se, adicionalmente, se diferentes graus de discriminação, tamanhos de teste, número de cadernos, número de blocos, número de itens dentro do caderno e ordenamento desses itens não estariam gerando diferenças nos resultados de matemática, 8ª série EF, entre ANEB 2005 e Prova Brasil 2005.

Todas essas questões referem-se à busca dos fatores que possivelmente estão influenciando na estimativa da habilidade da TRI e que podem impactar na validade e na precisão dos resultados dos testes. Respostas a essas questões fornecerão dicas que subsidiarão decisões relativas ao gerenciamento metodológico do SAEB, principalmente no que tange à mudança ou à manutenção da estrutura dos testes dessa avaliação de monitoramento para os próximos anos.

## 2. Objetivos

### 2.1 Objetivo Geral

O presente estudo tem como objetivo verificar a relação das características dos testes com a validade e a fidedignidade das estimativas de habilidade da TRI. As características dos testes envolvidas no estudo: qualidade pedagógica dos itens e seu alinhamento à matriz de referência, qualidade psicométrica dos itens, tamanho dos testes

(número total e número de itens por caderno) e distribuição dos itens pela escala de estimativas de habilidade.

O estudo é pertinente, pois (a) fornece orientações à elaboração de testes educacionais, ao INEP e a outros interessados no desenvolvimento desses instrumentos; (b) subsidia uma análise da qualidade dos resultados da ANEB e da Prova Brasil divulgados em 2005; e (c) levanta informações sobre fatores associados à diferença dos resultados de estimativas de habilidade entre ANEB 2005 e Prova Brasil 2005.

De acordo com delineamentos específicos, foram comparadas estimativas de habilidade dos estudantes em matemática 8ª série EF obtidos por: (a) teste original ANEB, com 155 itens; (b) tendo por base a seleção de itens do teste ANEB, testes simulados com 104 itens, mas com 24 itens por caderno semelhante ao delineamento da Prova Brasil; (c) a partir da redução de itens do Teste ANEB, teste simulado com 81 itens de forma a equiparar ao número de itens total da Prova Brasil; (d) teste original Prova Brasil, com 81 itens. Os testes ANEB, na prática foram compostos por 155 itens, pois foram excluídos dos 169 originais, aqueles que apresentaram baixa qualidade pedagógica e psicométrica. Esta é a mesma justificativa para a redução de 84 para 81 da Prova Brasil.

## 2.2 Objetivos Específicos

2.2.1 Replicar os estudos de Rabello (2007) e Condé (2007), confirmando as diferenças entre os resultados médios de estimativas de habilidade dos estudantes de escolas públicas urbanas, em matemática 8ª série EF, que responderam aos testes ANEB e Prova Brasil.

2.2.2 Analisar os percentuais de estudantes localizados em cada uma das faixas de habilidades estimadas por meio dos testes ANEB e Prova Brasil. A análise terá a função de complementar os estudos por comparação de médias realizados por Rabello (2007) e Condé (2007).

2.2.3 Analisar os testes ANEB e Prova Brasil em termos da abrangência da cobertura da Matriz de Referência do SAEB (INEP, 2002).

2.2.4 Analisar as características psicométricas dos testes ANEB e Prova Brasil em termos de discriminação e de dificuldade. As funções de informação dos testes (TRI) serão estimadas e analisadas para verificar o grau de fidedignidade de seus resultados.

2.2.5 Estimar as habilidades dos estudantes da ANEB de acordo com os mesmos critérios utilizados pelo INEP.

2.2.6 Estimar as habilidades dos estudantes da ANEB sob novas configurações de teste (Formas A, B, C e D).

2.2.6.1 Estimar as habilidades com base no Teste A, composto por 104 itens de forma que cada estudante tenha respondido a 24 itens, aproximadamente o mesmo número de itens do caderno da Prova Brasil, mantendo a discriminação e a dificuldade próximas da ANEB.

2.2.6.2 Estimar as habilidades com base no Teste B, composto por 104 itens a partir da manutenção dos itens mais discriminativos dos blocos.

2.2.6.3 Estimar as habilidades com base no Teste C, composto por 104 itens a partir da manutenção dos itens mais discriminativos que permitam manter a dificuldade semelhante à da Prova Brasil.

2.2.6.4 Estimar as habilidades com base no Teste D, composto pelos 81 itens mais discriminativos, número total da Prova Brasil.

2.2.7 Verificar a relação entre características dos testes (número de itens, discriminação e dificuldade dos itens) com a validade e a fidedignidade das estimativas de habilidade obtidas por meio dos diferentes modelos de testes.

2.2.7.1 Comparar os resultados de estimativas de habilidade obtidas por meio da ANEB e do Teste A. Como a alteração principal entre os testes foi o número de itens, pode-se fazer inferências sobre o impacto do número de itens nas estimativas de habilidade.

2.2.7.2 Comparar as estimativas de habilidade obtidas por meio dos Testes A, B, C e D com os resultados da Prova Brasil. Comparar a distância entre essas distribuições com a obtida a partir da comparação ANEB e Prova Brasil.

2.2.7.3 Verificar o grau de fidedignidade dos testes e sua associação com o número de itens, com a discriminação e com a dificuldade dos testes. Comparar seus resultados associando o EPM dos estudantes em resposta aos testes e o perfil de informação do teste com os parâmetros *a* e *b* dos itens.

# 3. Revisão da literatura

## 3.1 Avaliação e testagem educacional

### 3.1.1 O construto competência

O termo competência geralmente está associado ao "ser capaz de" realizar alguma tarefa ou um conjunto delas. Nos campos da Psicologia e da Educação, o termo competência é utilizado como definição de um objeto de estudo, embora seu entendimento

não seja consensual. Observam-se similaridades e diferenças de outros termos ou construtos como capacidade, proficiência, desempenho, inteligência, habilidade e conhecimento.

De acordo com Manfredi (1998), competência é um conceito aplicado às seguintes áreas de conhecimento e pesquisa: psicologia do desenvolvimento e da aprendizagem, psicometria e avaliação educacional. A autora apresenta que, historicamente dentro do campo da psicologia, foi estudado no âmbito (a) da psicologia do desenvolvimento, que foca o processo do desenvolvimento psicológico de acordo com as fases do desenvolvimento; (b) da psicologia da aprendizagem, que estuda os processos e as condições em que a aprendizagem humana ocorre em diversos contextos e em situações de ensino-aprendizagem; e (c) da construção de testes psicológicos, com a medida das capacidades e habilidades cognitivas, psicomotoras e afetivo-emocionais.

Nesses campos, identificam-se pelo menos duas linhas epistemológicas, o que remete a diferentes definições de competências e habilidades. A primeira foca que as dimensões objetivas e diretamente observáveis são as únicas passíveis de estudo. "Esta abordagem atribui importância central à construção de instrumentos estatisticamente padronizados de aferição e mensuração dos atributos indicativos da presença ou ausência de determinadas habilidades e ou capacidades." (Manfredi, 1998). Dentro deste contexto, entende-se habilidade como a própria demonstração do comportamento e as competências expressam graus de eficiência no desempenho.

A segunda abordagem utiliza outros conceitos tais como esquemas sensório-motores, ações, operações intelectuais, estruturas cognitivas, funções e representações, baseando-se sempre em uma teoria subjacente no estudo do desenvolvimento cognitivo e aprendizagem humana. Dentro dessa perspectiva, o comportamento é representação visível do traço latente ou do atributo psicológico interno denominado competência. Assim, se pudermos definir competência dentro desse modelo, o objeto de estudo assume um caráter teórico e interno, não observado diretamente, mas apenas indiretamente por meio do comportamento.

Pestana (2006) identificou características comuns entre as diversas definições de competência oriundas de diferentes orientações teórico-conceituais: "(a) há forte tendência em definir a competência por seus atributos, por seus ingredientes; (b) a referência às tarefas, à atividade humana ou à resolução de problemas em circunstâncias identificáveis; (c) o desempenho esperado das pessoas ou grupos diante de tarefas, atividades ou problemas; (d) o caráter estruturado dos processos de mobilização dos saberes, de fazeres e

de atitudes comportamentais que asseguram o desempenho perante a tarefa; e (e) a possibilidade de se fazer predições sobre a capacidade (a competência)" (p. 35-36).

A partir dessas características comuns, Pestana (2006) define competência como "(...) uma característica individual ou coletiva, ligada a uma possibilidade de mobilização e utilização de um conjunto de saberes, de capacidades e de atitudes comportamentais, de forma eficaz em um contexto determinado" (p. 36).

Em consonância com esta definição, Perrenoud (1999) considera que "concreta ou abstrata, comum ou especializada, de acesso fácil ou difícil, uma competência permite afrontar regular e adequadamente uma família de tarefas e de situações, apelando para noções, conhecimentos, informações, procedimentos, métodos, técnicas ou ainda a outras competências, mais específicas." (p. 4).

Para efeitos do presente trabalho, competência é a capacidade de mobilizar recursos (informações, conhecimento, processos psicológicos) para a resolução de problemas. Suas características: (a) trata-se de um construto muito amplo e, por isso, necessita de delimitações: 'competência em que?'; (b) é um traço latente e, portanto, para ser estudado pela ciência, exige uma teoria que associa este construto a uma representação comportamental ou tarefa; (c) pode ser mensurado por meio da manifestação dessa representação; e (d) é objeto de estudo da Psicologia na medida em que é definido em termos de processos psicológicos.

### 3.1.2 Avaliação de competências

Avaliação, no âmbito do senso comum, está relacionada à capacidade do indivíduo de identificar e analisar situações de forma a obter informações para tomar decisões. O termo avaliar tem sua origem no latim, provindo da composição *a valere*, que quer dizer "dar valor a...". O conceito "avaliação" é formulado a partir da conduta de "atribuir um valor ou qualidade a alguma coisa, ato ou curso de ação", que, por si implica um posicionamento positivo ou negativo em relação ao objeto, ou ato ou curso de ação avaliado (Luckesi, 2003, p. 92).

Quando o objeto da avaliação está no âmbito científico, o ato de avaliar é submetido necessariamente às regras e ao método da ciência. Um controle rígido das variáveis do atributo avaliado, do contexto em que está inserido e dos procedimentos como essas variáveis se relacionam são requeridos. Quando o objeto de interesse científico é delimitado no nível de conhecimento apreendido, de habilidade desenvolvida ou de

competência construída, as áreas de avaliação psicológica e educacional encontram rico campo de contribuições.

Maloney e Ward (1976) tratam avaliação psicológica como um processo flexível e não-padronizado que tem por objetivo chegar a uma determinação sustentada a respeito de uma ou mais questões psicológicas através da coleta, avaliação e análise de dados apropriados ao objetivo da questão. De acordo com Alchieri e Cruz (2004), "avaliação psicológica se refere ao modo de conhecer fenômenos e processos psicológicos por meio de procedimentos de diagnóstico e de prognóstico e, ao mesmo tempo, aos procedimentos de exame propriamente ditos para criar as condições de aferição ou dimensionamento dos fenômenos e processos psicológicos conhecidos" (p. 24).

Quando a área de Educação deixa de tratar como seu interesse de estudo os conteúdos aprendidos e passa a investigar os processos psicológicos cognitivos ou as competências, torna-se difícil a distinção do seu objeto com o da Psicologia. Por conseqüência, passa a ser tênue a linha divisória entre o objeto da avaliação educacional e da avaliação psicológica. Trata-se de um diagnóstico de um mesmo fenômeno humano em que são utilizados diversos métodos de coleta com a finalidade de captar informações de naturezas variadas para que a tomada de decisão seja mais eficaz.

### 3.1.3 Teste psicológico em educação

Para a Psicologia, teste "(...) é um procedimento sistemático para a obtenção de amostras de comportamento relevantes para o funcionamento cognitivo ou afetivo e para a avaliação destas amostras de acordo com certos padrões" (Urbina, 2007, p. 12). Trata-se de um instrumento da avaliação e fornece como resultados mais um indicador para tomada de decisão.

De acordo com McIntire e Miller (2000), todos os testes psicológicos apresentam três características fundamentais em comum: (i) avaliam uma amostra representativa de comportamentos que medem atributos pessoais ou predizem outros comportamentos; (ii) a amostra de comportamentos é levantada de acordo com condições padronizadas de aplicação; e (iii) apresentam regras e definições para cálculo de seus escores.

Urbina (2007) ressalta que a denominação teste, em função de sua definição histórica, deveria estar associada apenas àqueles procedimentos que envolvem respostas certas ou erradas e que envolvem a avaliação de algum aspecto do funcionamento cognitivo, conhecimentos, habilidades ou capacidades de uma pessoa. Mas o termo passou a ser utilizado também para a avaliação de construtos como personalidade, preferências,

etc. A autora denomina de teste de habilidades aqueles que avaliam conhecimentos, habilidades ou funções cognitivas.

Novamente, percebe-se uma interseção de objetos ou atributos entre as áreas psicológica e educacional. Um teste que avalia competências em resolução de problemas em matemática é classificado como psicológico, pois busca obter informações sobre os processos cognitivos subjacentes por meio de amostras de comportamento. Como esse tipo de teste é utilizado muitas vezes no contexto educacional, é classificado como teste educacional.

Ferrara (2006) realizou uma revisão da literatura sobre a aplicação da Psicologia Cognitiva para o desenvolvimento de medidas educacionais. Identificou pesquisadores como Snow e Lohman (1989), Mislevy (2006), Camilli (2006) que realizaram estudos sobre as implicações da Psicologia Cognitiva para o delineamento de avaliações educacionais, incluindo análise de itens e validação das inferências de escores de testes sobre a perspectiva do processamento cognitivo e do desenvolvimento de modelos psicométricos cognitivos.

Dada sua utilidade e praticidade, o teste é um instrumento amplamente utilizado no âmbito educativo para diversas finalidades: atribuição de notas em sala de aula, seleção para ingresso em universidade, certificação para exercer uma profissão, verificação do nível de proficiência dos estudantes para tomada de decisão educacional, entre outras. Alguns institutos e associações como o American Educational Research Association (AERA), a American Psychological Association (APA) e o National Concil on Measurement in Education (NCME) têm somado esforços para o desenvolvimento da ciência da testagem de forma a garantir a qualidade técnica dos resultados advindos da testagem, bem como sua utilização de forma ética e inclusiva.

São temas de estudo da área da avaliação que envolvem testes: construção e revisão de itens, procedimentos de administração, metodologias de análises de resultados, desenvolvimento de escalas e de normas, e apresentação e divulgação de resultados. Dentro dos temas de interesse para garantia da qualidade dos resultados obtidos pela avaliação e pelos testes, dois parâmetros são de suma importância e mereceram destaque no estudo dos testes: a validade e a fidedignidade (precisão).

### 3.1.4 Validade

Validade dos resultados de uma testagem é “o grau em que todas as evidências acumuladas corroboram a interpretação pretendida dos escores de um teste para os fins

propostos" (AERA, APA & NCME, 1999, p. 11). Esta definição envolve alguns aspectos a serem discutidos.

Primeiramente, não se pode falar que um teste apresenta ou não validade, e sim que os resultados advindos da testagem possuem um determinado grau de validade. Ainda, o conceito de validade, que foi por muito tempo considerado como um parâmetro do teste, passa a ser atribuído aos escores da testagem. Depois, o grau de validade dos resultados da testagem é relativo ao contexto para o qual o teste foi construído ou teve sua qualidade avaliada. Evidências acumuladas se referem aos estudos empíricos que mostram o grau de validade dos resultados do teste para contextos específicos.

Esta definição contemporânea de validade exige não só do elaborador do teste a tarefa de analisá-la, mas exige do usuário (professor, gestor educacional, etc.) a realização de estudos que possam garantir um bom grau de validade de seus resultados para o contexto de interesse. Percebe-se ser fundamental a realização de estudos para a infinidade de contextos possíveis, incluindo replicações periódicas.

Validade é um conceito único e não é possível falar em tipos de validade, mas em tipos ou fontes de evidência do grau de validade da testagem (AERA, APA & NCME, 1999). As várias fontes de evidência do grau de validade da testagem são as baseadas no conteúdo do teste, nos processos de respostas, na estrutura interna, na relação com outras variáveis e nas conseqüências da testagem (AERA, APA & NCME, 1999). Para instrumentalizar o presente trabalho, duas fontes de evidências serão detalhadas: as baseadas no conteúdo do teste e as baseadas em sua estrutura interna.

O tipo de evidência de validade baseada no conteúdo do teste é obtido pela relação entre o conteúdo do teste e o construto que se pretende medir. "O conteúdo do teste se refere aos temas, às expressões e ao formato dos itens, tarefas ou questões de um teste, associado às orientações aos procedimentos de administração do teste e de interpretação de seus resultados" (AERA, APA & NCME, 1999, p. 11). O grau de validade de conteúdo dos resultados do teste está intimamente ligado à relação das tarefas com o construto avaliado (domínio de conteúdo, processo cognitivo). Assim, se um teste é construído para avaliar geometria, garante-se um bom grau de validade de conteúdo de seus resultados quando os itens efetivamente estão avaliando conhecimento nesta área e não em outra.

Urbina (2007) considera que os procedimentos de validação para testes de verificação da competência são simples pois "(...) as evidências a partir das quais as inferências serão feitas podem ser defendidas com argumentos lógicos e relações demonstráveis entre o conteúdo do teste e o construto que este pretende representar" (p.

165). Estudos que buscam o grau de evidência de validade baseada no conteúdo do teste necessitam da colaboração de especialistas ou juizes, conhecedores do construto em questão e de técnicas de construção de itens, para duas tarefas: (a) elaborar e revisar as questões do teste orientado pela teoria; e (b) compor o teste, organizando-o de forma equilibrada quanto ao domínio de conteúdo previsto pela teoria (Pasquali 1998). Essa busca pelas evidências da validade de conteúdo do teste é traduzida por Herman, Webb e Zuniga (2002) e por Bhola, Impara e Buchendahl (2003) como a busca pelo alinhamento (*alignment*) entre o teste e o conteúdo ou domínio cognitivo avaliado.

Evidências baseadas na estrutura interna do teste indicam o grau de relação entre os itens e os componentes do teste em conformidade ao construto que o teste se propôs medir (AERA, APA & NCME, 1999, p. 13). Essas evidências têm relação direta com a dimensionalidade do teste. Quanto maior a inter-relação entre as questões de cada dimensão (ou fator), maior o grau de validade dos resultados obtidos.

Uma questão associada à estrutura interna do teste se refere à Função Diferencial do Item (DIF), cujos estudos de consistência interna do teste procuram verificar se um conjunto particular de itens pode funcionar diferentemente para determinados subgrupos de examinandos. No caso de diferentes grupos de examinandos com habilidades similares diferirem em termos de desempenho em um grupo específico de itens, pode estar acontecendo DIF. Os resultados da testagem podem apresentar um baixo grau de validade já que grupos com habilidades semelhantes deveriam apresentar resultados semelhantes.

Toda a argumentação de investigação da validade "(...) pode indicar a necessidade de refinar a definição dos construtos, pode sugerir revisões no teste e em outros aspectos do processo da testagem e podem indicar necessidade de estudos adicionais em determinadas áreas" (AERA, APA & NCME, 1999, p. 17).

### 3.1.5 Fidedignidade

A fidedignidade "(...) é a qualidade dos escores de teste que sugere que eles são suficientemente consistentes e livres de erros de mensuração para serem úteis" (Urbina, 2007, p. 121). Os resultados da testagem apresentam um bom grau de fidedignidade na medida em que o procedimento de testagem é repetido para um mesmo grupo de pessoas e os resultados são consistentes ou semelhantes, em situações que não se esperam alterações na magnitude do construto psicológico avaliado.

Assim, medir de forma fidedigna é medir com um baixo grau de erro. Um erro de mensuração pode ser definido como "(...) qualquer flutuação nos escores resultantes de

fatores relacionados aos processos de mensuração que são irrelevantes ao que está sendo medido" (Urbina, 2007, p. 121).

Da mesma forma que, para o conceito validade, é importante falarmos (a) em grau de fidedignidade e não considerarmos se há ou não há fidedignidade; (b) que o grau de fidedignidade está relacionado aos resultados da testagem e não ao teste; e (c) que depende constantemente de evidências empíricas e sofrem influência das variáveis envolvidas no processo de mensuração em variados contextos. A dissociação da fidedignidade ao teste e associação desta aos seus resultados implicam em relativizar o parâmetro fidedignidade ao contexto em que está sendo aplicado. Assim um teste pode apresentar resultados com excelente fidedignidade para o âmbito de sala de aula, mas uma baixa precisão para avaliações em larga escala.

Pelo menos três teorias são relevantes para o estudo da fidedignidade do teste ou do erro de mensuração: a Teoria do Escore Verdadeiro, a Teoria da Generalizabilidade e a Teoria de Resposta ao Item.

A Teoria do Escore Verdadeiro baseia as conclusões da testagem em um escore ideal livre de erro. Uma das formas de alcançarmos uma proximidade entre o escore observado e o escore verdadeiro é a replicação da testagem inúmeras vezes no mesmo grupo. De acordo com este procedimento, os erros de mensuração tendem a se anular, pois poderemos trabalhar com um único resultado que represente a variabilidade desse erro. Sabe-se, no entanto, que é praticamente inviável coletar inúmeros conjuntos de comportamentos de um mesmo grupo ou pessoa. "Uma vez que a amostra do comportamento é limitada, esse escore observado difere do escore verdadeiro" (Cronbach, 1996, p. 178). Por definição, a diferença entre esses dois escores é o erro de mensuração.

Quando temos várias mensurações de um mesmo evento, observamos erro-padrão de mensuração (EPM). A variância do erro é, portanto, o quadrado de um EPM. "O EPM diz o quão amplamente as medidas de uma mesma pessoa tendem a se distribuir" (Cronbach, 1996, p. 178). A teoria permite estimarmos a proporção de vezes que o escore verdadeiro se encontra dentro de um determinado intervalo de escore observado. Uma definição mais técnica de fidedignidade, que torna mais clara a relação inversa com o erro de mensuração é apresentada por Cohen e Swerdlik (2002): o coeficiente de fidedignidade é "(...) a proporção que indica a razão entre a variância do escore verdadeiro da testagem e a variância total" (p. 128). O coeficiente atinge seu valor máximo (1,0) quando a medida não contém nenhum erro de variável.

Estudo de Embretson (1996) indica que o EPM de acordo com a TCT é constante pelos níveis da escala de escores, mas difere quando a população avaliada, já que essa costuma apresentar variabilidade diferente. O EPM é único para uma população, já que é aplicado a todos os níveis de escores.

A Teoria da Generalizabilidade (Brennan, 1983; Cronbach, Gleser, Rajaratnam & Nanda, 1972), também chamada de Teoria G, procura distinguir as fontes de erro, decompondo o erro em componentes de forma a descobrir a sua magnitude. De acordo com Cronbach (1996) a teoria "(...) nos diz mais sobre um procedimento de mensuração do que a análise tradicional" (p. 180). Diferentemente da Teoria do Escore Verdadeiro (ou da Teoria Clássica dos Testes), que considerava a variância do erro como de um tipo só e de forma que a pessoa tivesse um único escore verdadeiro, a teoria G "reconhece universos alternativos de generalização, e, portanto, muitos escores de universo" (Cronbach, 1996, p. 180).

De acordo com essa teoria, a medida de uma variável pretende generalizar para um domínio ou universo relevante de observações. Daí surge a definição de escores de universo, diferente do escore verdadeiro, que consideram diversas fontes de variância como erro. Pretende responder questões como: quais os erros oriundos de um procedimento de testagem? Quanta variância de erro decorre de cada fonte?

Urbina (2007) considera que "(...) para se aplicarem os delineamentos experimentais requeridos pela teoria G, é necessário obter múltiplas observações do mesmo grupo de indivíduos em todas as variáveis independentes que podem contribuir para a variância de erro em um dado teste (por exemplo, escore em todas as ocasiões, por todos os avaliadores, entre formas alternativas, etc.)" (p. 141-142). Uma ferramenta estatística bastante utilizada quando se quer estimar a força que cada variável contribui para a variância do erro é a análise de variância (ANOVA).

A Teoria de Resposta ao Item (TRI) fornece métodos mais sofisticados para estimar a fidedignidade dos resultados de uma testagem. "(...) As vantagens que esses modelos oferecem, especialmente para a testagem em larga escala e a testagem adaptativa computadorizada, têm estimulado seu desenvolvimento e aplicação nas últimas décadas" (Urbina, 2007, p. 143). De acordo com a autora, os métodos da TRI, a fidedignidade e o erro de mensuração são abordados sob o ponto de vista da função de informação de itens individuais do teste, em oposição ao teste como um todo.

Para a TRI, a função de informação do teste nada mais é que a soma das funções de informação dos itens que compõem o teste. Hambleton, Jones e Rogers (1993) destacam

que o poder de informação do teste influencia na precisão da habilidade estimada, de forma que quanto maior o nível de informação, mais acurada é a estimativa de habilidade. Embretson (1996) abordou que o EPM, no caso da TRI, difere pelos diversos escores, mas mantém-se igual para populações diferentes que respondem a um mesmo teste. Essas conclusões diferem do que a própria autora concluiu para a TCT. Com base nessa evidência, não se pode atribuir um valor único para o EPM, já que varia pelas faixas da escala de estimativas de habilidade, a não ser que os vários EPM possam ser ponderados pela freqüência de estimativas de habilidade para a qual eles correspondem (Embretson, 1996).

Urbina (2007) categoriza os erros que influenciam os escores de teste em três fontes: "(a) o contexto no qual a testagem ocorre (incluindo fatores relacionados ao administrador do teste, ao avaliador e ao ambiente, bem como aos motivos da aplicação do teste), (b) o testando e (c) o teste em si" (p. 125). Essa categorização é semelhante à adotada por Cohen e Swerdlik (2002), que apresentam as seguintes fontes associadas à variância do erro: (a) construção do teste, (b) administração do teste, (c) apuração e interpretação dos resultados do teste.

Os erros associados à construção do teste têm relação com o tamanho da variabilidade entre os itens de um teste. Citam-se dois tipos: os erros de amostragem de conteúdo e os erros por inconsistência entre itens. Urbina (2007) define os erros de amostragem de conteúdo como aqueles que indicam "(...) a variabilidade irrelevante aos traços que pode influenciar os escores de teste como resultado de fatores fortuitos relacionados ao conteúdo de itens específicos" (p. 129). Tem relação com a seleção dos itens que compõem o teste e a adequação da cobertura do conteúdo que o teste pretende avaliar. Quando o erro se manifesta indica o grau de variabilidade dos escores, não relacionados ao nível de competência dos alunos, mas a especificidades do teste. Além de baixo grau de validade, uma supercobertura de um determinado conteúdo ou aspecto do construto em detrimento de outros podem gerar resultados com baixa confiabilidade ou fidedignidade.

Cabe observar que a inconsistência entre itens se refere aos erros nos escores resultantes de flutuações nos itens ao longo do teste, diferentemente do erro de amostragem de conteúdo gerado pela configuração de questões que foram incluídas no instrumento. Correlações baixas entre itens de um teste podem indicar alguns deles não são consistentes com o teste como um todo.

O tamanho do teste tem impacto importante na fidedignidade dos resultados do teste, pelo menos quando são utilizados modelos baseados na TCT e índices de fidedignidade como a fórmula Spearman-Brown (Embretson, 1996, p. 343). Nesse caso, quanto maior a amostra de comportamento, o número de respostas a um teste ou o número de vezes que o teste é aplicado, menor o erro para estimarmos os escores. Assim, os resultados oriundos da aplicação de um teste com muitos itens, de acordo com a TCT, fornecem resultados mais fidedignos que os resultados de testes menores, considerando invariáveis outras fontes de erro. Cronbach (1996) reforça esse aspecto quando afirma que "um teste longo geralmente é melhor do que um curto, porque cada pergunta acrescentada melhora a amostra do desempenho" (p. 189).

Quando a TRI é utilizada, há evidências que o número de itens não necessariamente tem correlação direta com a fidedignidade. Estudo realizado por Embretson (1996, p. 343) indicou que, para uma testagem adaptativa em comparação com uma testagem tradicional, em que é apresentado um número não muito grande de itens, mas apropriados para cada respondente, tende a apresentar baixo EPM para os diversos níveis de estimativas de habilidade.

Por sua vez, a resposta a um teste muito grande pode acarretar fadiga no testando. Trata-se de um aspecto associado ao teste e à sua administração que pode influenciar fatores inerentes à motivação e cansaço dos respondentes. Nesse sentido, para se alcançar uma boa precisão dos resultados da testagem, deve-se procurar compor um teste com um número ótimo de itens, ou seja, o maior número de itens, desde que não afete consideravelmente a motivação e a disposição de responder-lhe de maneira apropriada.

Para minimizar fatores de perturbação e que pode gerar erro dos resultados da testagem, Vianna (1982) sugere que os itens devem ser organizados em ordem crescente de dificuldade e complexidade, componente que considera de ordem psicológica e que pode influenciar na segurança dos testandos. Sugere também que, os itens devem ser organizados em áreas de conteúdo uniforme. O autor reforça que outros fatores relacionados ao teste como sua formatação, legibilidade, construção de itens com linguagem clara também influenciam na fidedignidade dos resultados da testagem.

Os erros associados à administração do teste têm impacto direto na motivação e na atenção dos respondentes, o que gera a diminuição da confiabilidade dos resultados. Para medir com um baixo grau de erro, é fundamental que os avaliadores selecionem os instrumentos mais apropriados à população alvo, preparem ambientes adequados,

estabeleçam um bom *rapport* com os testandos e administrem os testes de acordo com procedimentos padronizados.

A questão da padronização ou da uniformidade nos procedimentos de aplicação merece um cuidado especial. Se existe um procedimento padronizado, com instruções pré-definidas e com tempo limite para resposta às questões, para aplicação em um grupo de respondentes, e esse não for cumprido à risca, a consulta a uma tabela de normas ou a comparação com outro grupo de respondentes fica inviável. Além disso, quando não se cumprem tais procedimentos em um grupo, em aplicação a vários grupos, não se podem comparar de forma precisa os resultados entre eles.

Por sua vez, o tempo disponível para resposta ao teste, mesmo sendo cumprido à risca, pode também ser fonte de erro associado à administração, quando é insuficiente. Questões podem deixar de ser respondidas apenas em função do tempo e não da ausência de competência para tal. Geralmente pré-testes são utilizados para estimar um tempo ótimo que os testandos utilizam para responder todas às questões e para utilizar esse tempo na aplicação final. Vianna (1982) considerou que ao fixar a duração da aplicação de um teste, o examinador deve levar em consideração os elementos: (a) idade e nível de escolaridade dos examinandos; (b) extensão do teste; (c) forma do item; (d) complexidade do conteúdo e dos comportamentos; (e) nível do vocabulário empregado e estrutura das sentenças; e (f) complexidade dos cálculos em testes numéricos.

Os testando precisam estar motivados para responderem ao teste, também uma questão crucial para a fidedignidade dos resultados. Por que os testandos estão respondendo? Porque almejam um cargo no governo, uma vaga na universidade? Porque o Ministério da Educação solicita sua participação e ele está ciente da importância de sua participação para a melhoria da educação brasileira? Esclarecimentos, orientações e outros reforçadores podem ser utilizados para conseguir o comprometimento dos testandos. Sem dúvida a motivação do testando influencia na precisão dos resultados da testagem e cabe à coordenação da avaliação encontrar as melhores estratégias para cada avaliação.

Os erros associados à apuração e à interpretação dos resultados ocorrem quando há diferenças no cálculo ou na interpretação dos resultados da testagem. A fidedignidade é comprometida quando dois apuradores chegam a conclusões diferentes sobre os resultados de um mesmo testando.

Um grau satisfatório de validade e de precisão dos resultados da testagem depende, em grande parte, dos procedimentos de desenvolvimento dos testes. Esses, por sua vez, são guiados pela definição de seu propósito e de acordo com as inferências que se esperam

realizar com base em seus resultados. Adicionalmente, "o processo de desenvolvimento do teste envolve considerações sobre o conteúdo, formato, contexto sob o qual será utilizado e potenciais conseqüências de seu uso" (AERA, APA & NCME, 1999, p. 37). A busca pela qualidade dos resultados inclui também a especificação das condições de administração, dos procedimentos de cálculo e de análise dos resultados de performance dos respondentes e das estratégias de divulgação e de produção de relatórios dos resultados focados nos possíveis usuários.

Se não é possível falar de validade e de precisão exclusivamente do teste, todos os procedimentos acima elencados devem ser considerados no planejamento e na construção do teste. A dissociação do teste do contexto da avaliação, do marco teórico adotado, da tabela de especificações, dos procedimentos de administração, da análise e da produção de materiais de divulgação certamente terá impacto negativo na validade ou na precisão de seus resultados.

O processo de desenvolvimento de testes psicológicos ou educacionais pode ser subdividido em quatro etapas (AERA, APA & NCME, 1999, p. 37): (a) delineamento do propósito do teste e da extensão do construto que será investigado; (b) desenvolvimento da tabela de especificações que orientará a construção do teste; (c) elaboração, avaliação e seleção dos itens e do guia para apuração dos resultados; (d) montagem e avaliação do teste para utilização.

Nas próximas seções do presente trabalho, cada uma das etapas de desenvolvimento de testes será abordada, sempre levando em consideração os fatores associados à validade e à fidedignidade de seus resultados. Pretende-se, ao tratar da elaboração de questões de testes educacionais e da composição de instrumentos, uma abordagem geral com possível aplicação a diversos contextos. No entanto, a aplicabilidade a infinitos contextos não será possível. Como a presente introdução visa fornecer um suporte teórico para análises relacionadas a um sistema de avaliação educacional brasileiro, em larga escala, que procura estimar as habilidades dos estudantes de Ensino Básico nas disciplinas língua portuguesa e matemática, por meio de testes compostos de itens de múltipla escolha, por vezes, a generalização para outros tipos de abordagem não será possível.

## 3.2 Base do desenvolvimento dos testes

### 3.2.1 Documentação da Avaliação

O desenvolvimento de um teste se baseia em decisões como os objetivos da avaliação, o referencial teórico adotado e os domínios cognitivos e conteúdos que serão

abarcados. Essas informações compõem um documento de trabalho que orientará a elaboração dos instrumentos, bem como a seleção de procedimentos de administração, de análise e de divulgação dos resultados.

O documento de trabalho orientador da avaliação (e consequentemente da testagem) tem sido denominado de *framework* (U.S. Department of Education, 1992a, 1992b, 1995a, 1995b, 2002a, 2002b) ou de 'Guia' (U.S. Department of Education, 1996, 1997, 1999) pelos sistemas avaliativos. Muitas vezes assumem nomes-fantasia sem alusão à natureza do documento, como por exemplo: Matrizes Curriculares de Referência do SAEB (1999); *Measuring Student Knowledge and Skills* do PISA (OECD, 2000); Minas Gerais, Avaliação da Educação (UFJF, 2001); SAEB 2001, Novas Perspectivas (INEP, 2002).

Independentemente da denominação utilizada, de modo geral, os documentos orientadores da avaliação em larga escala (que chamaremos de *framework*) apresentam as seguintes informações: (a) contexto Educacional em que o Sistema ou Programa de Avaliação está inserido; (b) apresentação do Sistema ou do Programa Avaliativo, incluindo a instituição responsável por seu planejamento e execução, os objetivos, o histórico das atualizações do documento, o público-alvo da avaliação e os possíveis usuários de seus resultados; (c) marco teórico orientador da avaliação que delimita o construto que se pretende avaliar e perspectivas teóricas; (d) matrizes de referência, indicando os conteúdos e os domínios cognitivos a serem avaliados, elaboradas com relação intrínseca ao marco teórico adotado e que orientam a construção dos itens e do teste. Quando tratados como 'tabelas de especificação' (Pasquali & Alves, 1999; Sant'anna, Enricone, André & Turra, 1996; Tyler, 1950), a organização gráfica dos conteúdos associados a domínios cognitivos trazem o número de itens que comporão a prova para cada uma dessas associações (conteúdo e domínio); e (e) seleção e definição dos instrumentos que serão utilizados para atendimento dos objetivos, incluindo o formato dos itens, as proporções dos testes que serão cobertas com itens de cada conteúdo e domínio, tipologia textual (se for o caso de definição pela utilização desse estímulo).

Alguns *frameworks* também apresentam informações sobre a amostra planejada e critérios de seleção, sobre procedimentos de coleta, de análise e divulgação dos resultados, mas essa observação não é sistemática. Sabe-se, no entanto, que essas informações são fundamentais, pois, como foi verificado anteriormente, não se pode planejar testes, sem que se tenham claros os próximos passos da avaliação.

Embora o termo *framework* esteja sendo utilizado no presente trabalho para se referir ao documento orientador da avaliação, como um todo, em alguns casos encontrados na literatura, referem-se exclusivamente ao marco teórico e aos conteúdos e domínios cognitivos que orientarão a concepção do teste (AERA, APA & NCME, 1999, p. 37; U.S. Department of Education, 1992a, 1992b, 1995a, 1995b, 2002a, 2002b).

Os termos *Stantards* ou *content standards* também são freqüentemente utilizados como documento de referência da avaliação. Apresenta a peculiaridade de indicar "o que deveria ser ensinado aos estudantes e o quão bem eles deveriam ter aprendido" (Herman, Webb, & Zuniga, 2002, p. 1). Cabe ressaltar que esses conceitos são mais amplos, pois incorporam a totalidade dos conteúdos e dos processos que deveriam ter sido adquiridos pelos estudantes no processo educacional. Já uma matriz de referência trabalha geralmente com uma amostra desses conteúdos e domínios, selecionados pelos objetivos da avaliação.

No que tange ao planejamento estrutural da avaliação e à necessidade de documentação, Ferrara e DeMauro (2006) e Ferrara (2006) propuseram quatro características que subsidiam o delineamento de testes e os propósitos da avaliação, no âmbito da Psicologia Cognitiva aplicada às medidas educacionais: (a) especificação do conhecimento do conteúdo, incluindo o que se conhece, como o conhecimento é organizado e quão bem pode ser acessado e utilizado; (b) especificação do conhecimento procedimental que envolve as estratégias específicas, quadro de processos de pensamento e habilidades de comunicação; (c) especificação de um plano de mensuração, incluindo exemplos de tarefas avaliativas e orientações quanto a inferências sobre o que os examinandos sabem e podem fazer e são sustentados pelo teste; e (d) apresentação de hipóteses e de evidências da relação do construto com outros construtos.

Ferrara (2006) sugere ainda a inclusão de três características necessárias à documentação do planejamento da avaliação: "(e) especificação dos caminhos de desenvolvimento dos examinandos sobre todas as facetas do construto, o que descreveria a sua performance em relação ao construto; (f) uma explicação das influências cultural, afetiva, conativa, de linguagem e outras no desempenho no teste; e (g) identificação das fontes de irrelevância de construto na avaliação do próprio construto" (p. 4).

### 3.2.2 Teoria e modelos dos testes

Informações sobre os propósitos da avaliação orientam a seleção do desenho dos testes, bem como da teoria e dos modelos que serão utilizados para seu desenvolvimento e análise de resultados.

Hambleton e Jones (1993) compararam a TCT e a TRI e sua aplicação no desenvolvimento de testes. As teorias dos testes (*test theories*) fornecem uma estrutura geral que vincula variáveis observadas, tais como escores de testes, a variáveis não-observadas, tais como o escore verdadeiro ou a habilidade estimada. Assumir uma opção teórica significa utilizar, necessariamente, seus conceitos, seus pressupostos e as especificidades de seus modelos de testes (*test models*).

Modelos de testes "(...) são formulados no âmbito de uma teoria dos testes e especificam, com consideráveis detalhes, a relação entre um conjunto de conceitos teóricos e um conjunto de pressupostos sobre esses conceitos e relações" (Hambleton & Jones, 1993, p. 39). Estudos empíricos são utilizados, posteriormente à avaliação, para verificar se o modelo adotado é apropriado ao conjunto particular de dados. Conhecendo-se as características, as exigências e os pressupostos de cada modelo, o mais apropriado é selecionado. "Para um teste contendo itens de múltipla escolha, por exemplo, em que é esperado um considerável acerto ao acaso, um modelo de teste com o pressuposto de escores verdadeiros e escores de erro não-correlacionados, pode não ser o mais apropriado" (Hambleton & Jones, 1993, p. 39).

Hambleton e Jones (1993) consideram que uma boa teoria ou um bom modelo de teste (a) ajuda a identificar a influência dos erros de medida na estimação das habilidades, contribuindo a serem minimizados; (b) fornece um conjunto de referências para a elaboração de um desenho de teste; e (c) especifica a relação precisa entre os itens do teste e os escores de habilidade advindos de sua aplicação. Assim, quando o delineamento de uma avaliação e a estrutura dos testes são definidos, devem-se ter claros a teoria e o modelo dos testes que serão utilizados.

A TCT é uma teoria sobre escores de testes que introduz três conceitos: escore observado, escore verdadeiro e escore do erro. Os pressupostos do modelo clássico dos testes são: (a) o escore verdadeiro e o escore do erro não são correlacionados; (b) a média do escore do erro na população é zero; e (c) os escores do erro em testes paralelos não são correlacionados. Testes paralelos são aqueles que medem o mesmo conteúdo, que um mesmo examinando apresenta o mesmo escore verdadeiro e que o tamanho do erro de medida entre as formas é igual. Gulliksen (1950) define as condições para o paralelismo entre testes: igualdade de médias, variâncias e covariâncias entre as formas. Também é um pressuposto considerar que testes paralelos podem ser construídos.

O modelo tem como foco os escores do teste e adota como principais parâmetros dos itens a dificuldade ou a proporção de acertos ($p$) (Hambleton & Jones, 1993; Nunnally

& Bernstein, 1994) e a discriminação (*r*) (Hambleton & Jones, 1993). Esses parâmetros estão associados, sob o modelo clássico dos testes, às estatísticas do teste tais como a média e o desvio-padrão do escore e à sua fidedignidade dentro do processo de desenvolvimento de testes com propriedades estatísticas desejadas. Trata-se de um modelo útil no desenvolvimento de testes quando "(...) a amostra de examinandos é similar à população para qual o teste está sendo desenvolvido" (Hambleton & Jones, 1993, p. 40), já que os parâmetros dos itens dependem da amostra de examinandos utilizada para estimá-los e os escores totais dependem dos parâmetros utilizados para calculá-los.

O cálculo do valor *p* dos itens se dá pela proporção de examinandos que os acertaram. Assim, um item é considerado difícil se esse percentual for baixo, e fácil, se for alto. Por outro lado, 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. Essa dependência circular pode ser minimizada e o modelo se ajustar aos dados quando a amostra de examinandos é similar à população.

Um exemplo de índice *r* é o coeficiente de correlação bisserial ($r_{bis}$) que "(...) é uma medida de associação entre o desempenho no item e o desempenho no teste. O coeficiente bisserial estima a correlação entre a variável de desempenho no teste e uma variável latente (não observável) com distribuição normal que, por hipótese, representa a habilidade que determina o acerto ou erro no item" (CESGRANRIO, 2006, p. 26). Como seus resultados estão atrelados ao desempenho no teste, também é fundamental, para o cálculo do $r_{bis}$ que a amostra de examinandos apresente características similares à da população.

Uma das implicações práticas (ou pouco práticas) dos valores *p* e *r* 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, torna-se um problema de difícil solução quando a amostra não apresenta as mesmas características da população. Assim, se um item foi submetido a um pré-teste e a duas avaliações, por exemplo, recebe três conjuntos de índices TCT. Se o item é o mesmo, como pode apresentar mais de um conjunto de parâmetros, ou seja, mais de uma identidade psicométrica?

A TRI é "uma teoria estatística sobre a performance do examinando no item e no teste e sobre como essa performance relata as habilidades que são mensuradas pelos itens no teste" (Hambleton & Jones, 1993, p. 40). É composta de um conjunto de modelos matemáticos 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, sob um paradigma que especifica uma relação teórica entre as pontuações empíricas dos examinandos em um teste e o traço latente não observável, teorizado como o responsável por tais pontuações.

Hambleton e Jones (1993, p. 40) consideram que vários são os modelos utilizados pela TRI para o estabelecimento da relação entre a resposta ao item com as habilidades subjacentes, sendo que os mais comuns (a) assumem uma habilidade única subjacente à performance ao teste; (b) podem ser aplicados a dados oriundos de testes compostos por itens dicotômicos; e (c) assumem a relação entre a performance no item e a habilidade em função de modelos logísticos de um, dois ou três parâmetros.

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. Requena (1990) ressalta que as funções de resposta ao item estabelecem as relações, matematicamente formalizadas, de como cada resposta depende de certo nível ou grau de habilidade no traço considerado. Quando a Psicometria se apropria desses modelos, percebe-se que seus parâmetros 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, envolve os parâmetros *a*, *b* e o de probabilidade de "acerto ao acaso" (parâmetro *c*) (Cronbach, 1996; Hambleton & Jones, 1993; Hambleton, Swaminathan & Rogers, 1991; Pasquali, 2003). O parâmetro *teta* ($\Theta$) representa a estimativa ou o parâmetro de habilidade dos testandos.

Hambleton, Swaminathan e Rogers (1991) consideram que a TRI é 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 que a Psicometria tradicional, baseada na Teoria Clássicas dos Testes (TCT), contém.

De acordo Hambleton e Jones (1993), tipicamente, dois pressupostos estão relacionados com os modelos da TRI: a estrutura matemática da função ou da Curva Característica do Item (CCI) e a estrutura dimensional dos dados do teste.

A CCI representa graficamente os parâmetros *a*, *b* e *c*, apontando a probabilidade de responder corretamente um determinado item em função da habilidade. Pela variação dos parâmetros do item, várias CCI podem ser geradas para o ajuste aos dados do teste. A

função característica do teste é a soma de todas as funções características dos itens que compõem o teste e pode ser usada para predizer os escores dos examinandos em função dos níveis de habilidade.

As funções de informação do item apresentam a contribuição de cada item para avaliação da habilidade. De modo geral, itens com alto poder discriminativo contribuem mais para a fidedignidade da medida que itens com baixo poder discriminativo. A função de informação do teste, I(Θ), é a soma das funções de informação dos itens (Hambleton & Jones, 1993).

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

Cada item contribui independentemente para a função de informação do teste (Hambleton, Swaminathan & Rogers, 1991) de forma que a contribuição individual de cada item é possível sem o conhecimento das informações dos outros itens. Por sua vez, o índice de fidedignidade da TCT não pode ser determinado independentemente das características do conjunto de itens do teste, já que são considerados, para seu cálculo, os escores totais. O conjunto de informações obtidas por um teste é inversamente relacionada ao EPM e diretamente relacionada à fidedignidade da medida para cada ponto da escala de habilidades (Embretson, 1996; Hambleton, Swaminathan & Rogers, 1991).

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

Consegue-se com essa função avaliar o grau discriminativo das estimativas de habilidade para as diversas magnitudes da escala. Embretson (1996) destaca que o EPM é baixo para as faixas de estimativas de habilidade moderadas e é alto para as faixas extremas. A autora sugere um valor composto de EPM para cada faixa de estimativas de habilidade calculado a partir da média dos EPM individuais das estimativas, ponderado pela freqüência de sujeitos localizados em cada faixa. Como o EPM é inversamente relacionado à raiz quadrada da informação do teste, sugere-se que a função do teste possa ser criada tendo por base exclusivamente o EPM estimado para cada sujeito.

A magnitude do EPM depende, de forma geral (Hambleton, Swaminathan & Rogers, 1991): (a) do número de itens do teste; (b) da qualidade dos itens do teste (EPM pequenos são associados à alta discriminação dos itens); e (c) da relação entre a dificuldade e a habilidade estimada (EPM pequenos são associados a testes com parâmetro *b* aproximadamente igual ao parâmetro de habilidade dos examinandos).

A TRI assume a propriedade de invariância dos parâmetros, considerada como a sua maior distinção da TCT. Esse princípio afirma que as habilidades dos sujeitos são estimadas independentemente do teste utilizado; bem como os parâmetros dos itens, independentemente da amostra de examinandos que os responderam (Baker, 2001; Fan & Ping, 1999; Hambleton, Swaminathan & Rogers, 1991).

Condé e Rabello (2001), Condé (2002) e Condé e Laros (2007), com os dados de aplicação de 26 formas de provas de língua portuguesa do SAEB aplicado em 1997, verificaram o comportamento dos índices de habilidade calculados por meio da TCT e da TRI, quando correlacionados com índices de dificuldades. Os índices de habilidades calculados pela TCT se mostraram mais dependentes da dificuldade das provas que os parâmetros de habilidades estimados pela TRI.

Baker (2001) considerou que a invariância dos parâmetros depende de duas condições: (i) necessidade dos valores de todos os parâmetros dos itens estarem em uma métrica comum; e (ii) necessidade dos itens da prova estarem medindo uma mesma habilidade, ou seja, serem unidimensionais. Assim, se as condições são satisfeitas, os itens tendem a propiciar estimativas de habilidade pela TRI sem dependência com a amostra de examinandos que foi utilizada para estimá-la.

Condé (2002) e Condé e Laros (2007) investigaram se a estimativa de habilidade da TRI depende da dificuldade dos itens utilizados para estimá-la, bem como em que medida a unidimensionalidade do teste influencia a propriedade de invariância da habilidade dos sujeitos. Foram utilizados os dados de 26 formas de teste de matemática de 8ª Série do SAEB 97 respondidas por 18.806 estudantes brasileiros de escolas públicas e particulares de cada uma das 27 Unidades da Federação brasileiras. Essas formas de teste foram respondidas por 26 grupos de estudantes equivalentes em termos de habilidades. Os resultados apontaram para a existência de uma dependência da habilidade em relação à dificuldade dos cadernos (*r* de Pearson = 0,68, com o valor *p*; *r* de Pearson = -0,69 com o parâmetro *b*). A dependência entre a habilidade da TRI e a dificuldade diminui quando são excluídos da prova os itens com cargas fatoriais inferiores a 0,20 no fator principal, que contribuem menos 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. Os autores concluíram que a habilidade estimada pela TRI depende da dificuldade dos itens que são utilizados para estimá-la, não confirmando a propriedade de invariância dos parâmetros. Por sua vez esta estimativa da TRI apresenta uma diminuição da dependência com relação à dificuldade quando a prova se aproxima da unidimensionalidade. O estudo reforça a condição que Baker (2001) coloca para a invariância dos parâmetros: estarem medindo a mesma habilidade, já que, quando o teste se distancia da unidimensionalidade, a propriedade de invariância fica prejudicada.

Hambleton e Jones (1993) ressaltam a condição que "a propriedade de invariância dos parâmetros somente é obtida com modelos que se ajustam aos dados do teste aos quais são aplicados" (p. 42). Após revisão de literatura, Fan e Ping (1999) indicaram que questões relacionadas ao impacto da falta de ajuste do modelo aos dados na propriedade de invariância dos parâmetros da TRI não têm sido adequadamente estudadas. Comparando estimativas dos parâmetros *a* e *b*, para modelos de 1, 2 e 3 parâmetros, para populações diferentes, os autores concluíram: (a) nenhum efeito negativo da falta de ajuste do modelo aos dados na propriedade de invariância do parâmetro *b* foi observado; (b) não se pode afirmar que há efeitos negativos de falta de ajustes do modelo aos dados na invariância do parâmetro *a* estimado (neste caso, estudaram apenas modelos de 2 e 3 parâmetros); e (c) há uma tendência dos resultados na direção da falta de ajuste do modelo aos dados reduzir o grau de invariância do parâmetro de habilidade. No entanto os autores consideraram seus achados pouco conclusivos, mas contribuem para o estudo do tema ajuste dos modelos aos dados e propriedade de invariância da TRI, que consideram de grande relevância. Fan e Ping (1999) utilizaram, para verificação do ajuste do modelo aos dados, a checagem individual do desajuste dos itens a partir da razão entre o qui-quadrado e os graus de liberdade. Assim, caso o item apresente razão inferior a 1,96 ($P < 0,05$), não apresenta um bom ajuste ao modelo.

Pontos positivos podem ser encontrados tanto na TRI, quanto na TCT, cabendo a seleção daquela mais apropriada aos propósitos e ao delineamento da avaliação (Hambleton & Jones, 1993). A TRI apresenta quatro aspectos favoráveis: (a) As estatísticas dos itens são independentes dos grupos de examinandos utilizados para estimá-las; (b) As habilidades dos examinandos não são dependentes da dificuldade dos testes utilizados para estimá-las (desde que o pressuposto da unidimensionalidade seja verificado); (c) Os modelos de teste permitem uma relação entre os itens e os níveis de

habilidade; e (d) Os modelos de teste não requerem a construção de testes paralelos para avaliação da fidedignidade.

Por seu turno, a TCT apresenta as seguintes vantagens: (a) Pequenas amostras são requeridas para as análises; (b) Utiliza análises matemáticas mais simples, se comparadas às utilizadas pela TRI; (c) A estimação dos parâmetros do modelo é conceitualmente clara; e (d) Análises não requerem estudos de ajuste para assegurar um bom ajuste do modelo aos dados.

Tendo em vista os pontos fortes e as limitações de cada teoria e de cada modelo associado, selecionam-se os mais apropriados para orientar o desenvolvimento dos testes educacionais dentro dos propósitos de uma avaliação específica e de seu *framework*. A escolha da teoria e do modelo da avaliação definitiva terá impacto direto na elaboração e revisão de itens, na estruturação e na análise de resultados do pré-teste, na composição do teste definitivo e na análise de dados da avaliação.

## 3.3 Elaboração e análise de itens

### 3.3.1 Elaboração e análise teórica de itens

No âmbito educacional, a mensuração de competências é realizada por meio de amostras de comportamentos ou de tarefas que permitem ao sujeito demonstrar um conjunto de habilidades (observáveis) que, em seu conjunto, as caracterizam. Com base nesses comportamentos, infere-se que desenvolveu uma determinada competência.

O teste educacional se fundamenta nas matrizes de referência da avaliação que, por sua vez, apresentam extensão suficiente na cobertura dos aspectos fundamentais do traço latente, delimitado pela teoria e pelos propósitos da avaliação.

Para a construção de um teste, um conjunto de itens é previamente elaborado de forma alinhada aos conteúdos e habilidades previstas nas matrizes de referência. Herman, Webb e Zuniga (2002) definem alinhamento como a sincronia entre os *standards* (lista de conteúdos/habilidades que se espera que os estudantes tenham desenvolvido) com os testes (p. 1). De uma forma mais ampla, Bhola, Impara e Buckendahl (2003) definem alinhamento como "(...) o grau de concordância entre os conteúdos que os estudantes deveriam adquirir (*content standards*) em uma determinada área e a avaliação usada para mensurar o desempenho dos estudantes com relação a esses conteúdos" (p. 21). O conceito de alinhamento tem relação direta com a questão da validade dos resultados da testagem. Os autores especificam que "alinhamento é um elemento básico com relação ao corpo de

evidências relatadas para a validade das interpretações dos escores do teste" (Bhola, Impara & Buckendahl, 2003, p. 22).

Especialistas nas áreas de interesse da avaliação são chamados a elaborar questões em quantidade suficiente para cada uma das habilidades avaliadas. Sua ação está orientada à busca do alinhamento dos itens aos conteúdos e processos cognitivos apresentados na matriz de referência.

O número total de questões que será construído, o número de itens por conjunto de conteúdos/processos cognitivos, o grau de complexidade dos itens, os tipos de itens que serão utilizados (múltipla escolha, resposta construída, etc.), a forma de aplicação, os recursos que os estudantes terão à disposição para responder às questões, a metodologia de análise dos dados e as estratégias de divulgação dos resultados devem estar em sincronia e alinhados. Esses aspectos, por sua vez, devem ser inerentes ao propósito da avaliação, ao próprio teste, à administração, ao tipo de análise, à publicação e à utilização dos resultados orientarão a elaboração de itens. Para tanto, os especialistas devem possuir um conhecimento aprofundado, não só sobre as técnicas de construção dos itens e sobre a matéria da disciplina para a qual pretende construí-los, mas sobre todos os aspectos de um sistema avaliativo.

Antes do processo de elaboração das questões, é fundamental que os planejadores da avaliação já definam o desenho do teste, número de itens, número de cadernos e os tipos de itens. Hambleton e Jones (1993) sugerem que sejam estabelecidos previamente uma as teoria e os modelos dos testes que orientarão sua construção e as etapas posteriores de análise dos resultados. Tendo por base esse planejamento, parte-se para a elaboração dos itens.

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), para a aquisição de itens de múltipla escolha referentes às disciplinas matemática e língua portuguesa para o Sistema Nacional de Avaliação da Educação Básica (SAEB), elaborou um documento com as especificações técnico-pedagógicas para elaboração de itens (PNUD, 2006). Amparados pelos propósitos, pela cobertura das matrizes de referência e pelo público alvo da avaliação, o documento apresenta 21 critérios gerais para construção de itens, seis para construção dos enunciados e 13 para construção das alternativas dos itens de múltipla escolha.

De acordo com as especificações (PNUD, 2006), os itens devem: (a) ser inéditos; (b) considerar o cotidiano dos alunos; (c) enfocar apenas um problema; (d) utilizar terminologias de caráter universal; (e) apresentar independência local; (f) não apresentar

viés cultural e propagandas; (g) depender pouco ou nada da memorização; (h) evitar expressões duplamente negativas; (i) não conter 'pegadinhas'; (j) considerar o tempo de leitura exigido do aluno; (k) utilizar linguagem apropriada à série; (l) usar linguagem clara e direta; (m) apresentar redação gramaticalmente consistentes e pontuação correta; (n) contemplar um contexto para o problema que atinja a realidade dos estudantes; (o) utilizar distratores plausíveis; e (p) não conter erros conceituais.

Por mais preparados que sejam os especialistas elaboradores de itens, é preciso assegurar que esses itens apresentem boa qualidade técnico-pedagógica. Por isso, dentro de qualquer processo de desenvolvimento de instrumentos, é fundamental a atividade de validação ou de revisão teórica de itens.

A verificação do grau de validade de conteúdo ou de alinhamento entre os itens e as matrizes de referência também é realizada por especialistas na área do construto avaliado, conhecedores do conteúdo e dos processos cognitivos envolvidos, das próprias matrizes e de técnicas de construção de itens.

A revisão dos itens, também chamada de validação teórica, é a etapa de busca *a priori* (antes de qualquer aplicação) da validade dos resultados da testagem. Pasquali (1998) sugere procedimentos para a busca das evidências empíricas que comprovem um bom grau de validade de conteúdo ou de construto das escalas psicológicas. Para verificação da pertinência, os itens construídos com base na teoria devem, por argumentos lógicos e semânticos, avaliar o conteúdo previsto.

Bhola, Impara e Buckendahl (2003) realizaram uma revisão da literatura sobre métodos utilizados para garantir um bom grau de alinhamento entre o sistema avaliativo, incluindo o teste, e os conteúdos/processos cognitivos especificados nas matrizes de referência. Sumarizaram os métodos de alinhamento em três categorias: baixa, moderada e alta complexidade.

No caso dos métodos de baixa complexidade, "especialistas de conteúdo (...) examinam cada item do teste e indicam em que medida os itens apresentam relação com os standards de conteúdo ou aos elementos da tabela de especificação do teste" (Bhola, Impara e Buckendahl, 2003, p. 22). No caso dos métodos de complexidade moderada, os especialistas são questionados sobre a relação entre os standards e os itens do teste sob a perspectiva do conteúdo e da complexidade cognitiva. Como apresenta um critério adicional em comparação ao método de baixa complexidade, ou seja, a avaliação da complexidade cognitiva, os autores alertam para a redução do número de itens considerados alinhados, ou seja, cuja utilização contribuirá para a validade dos resultados

do teste. validade. Ressaltam também que o Council of Basic Education (CBE), dos Estados Unidos, associadamente a esse método, busca avaliar o balanceamento do número de itens por conteúdo e grau de complexidade cognitiva dos testes em fase de montagem, bem como verificar se o tipo de itens (resposta construída, múltipla escolha) fornecerão resultados satisfatórios aos propósitos avaliativos.

São vários os métodos de alinhamento de complexidade alta sumarizados por Bhola, Impara e Buckendahl (2003). Apresenta-se aqui o modelo de La Marca (2000), citado como relevante pelos autores, que busca determinar o quão bem os standards estão sendo mensurados pela avaliação, usando cinco dimensões inter-relacionadas: relação com o conteúdo, profundidade do conteúdo, ênfase, relação com o desempenho e acessibilidade (Bhola, Impara e Buckendahl, 2003, p. 22).

As duas primeiras dimensões são coerentes com o método de moderada complexidade, que contemplam o alinhamento do teste com as matrizes de referência de acordo com o conteúdo e com o grau de complexidade cognitiva. A dimensão denominada 'ênfase' analisa o grau em que a avaliação está alinhada à ênfase teórica da avaliação. A dimensão 'relação com o desempenho' verifica o grau no qual os itens permitem aos estudantes demonstrar seus conhecimentos. A dimensão 'acessibilidade' verifica a extensão em que a avaliação inclui itens cuja dificuldade permite que os estudantes de todos os níveis de proficiência tenham oportunidade de demonstrar seu nível de conhecimento.

Em suma, a busca pelo alinhamento é a função principal da revisão dos itens. Tem como objetivo, por meio da utilização de algumas técnicas, prover os futuros resultados do teste de um bom grau de validade. Essa etapa é capaz de proporcionar um maior aproveitamento do percentual de itens após o pré-teste.

### 3.3.2 Pré-teste e análise empírica de itens

Após a elaboração e a revisão, os itens são submetidos ao pré-teste, ou seja, a uma aplicação, prévia à aplicação definitiva do instrumento, com os objetivos principais de verificar empiricamente a qualidade dos itens e de levantar algumas informações que possibilitem uma tomada de decisão sobre aqueles que entrarão no teste definitivo. Trata-se de mais uma etapa pela busca de um bom grau de validade e fidedignidade dos resultados da avaliação.

### 3.3.2.1 Estrutura do pré-teste

O pré-teste é programado, tendo em vista os propósitos da avaliação, o delineamento do teste e o desenho da amostra de examinandos da aplicação definitiva, associados às teorias e aos modelos que serão adotados para análise de dados. O tamanho e o desenho do teste definitivo, incluindo a cobertura dos conteúdos e domínios cognitivos, orientarão a definição do número e das especificações dos itens a serem pré-testados.

Como o número de itens que apresenta um bom grau de qualidade após a análise dos dados do pré-teste, geralmente, é inferior ao número de itens pré-testados, o número de itens pré-testados deve ser superior ao número que será utilizado. O desenho do teste definitivo planejado antes da estruturação do pré-teste permitirá ainda programar o quantitativo de itens que será pré-testado para cada um dos descritores (conteúdos, domínios cognitivos, etc.) das matrizes de referência.

O planejamento do teste definitivo e os procedimentos que serão utilizados para compô-lo, associados a teorias e a modelos específicos, terão impacto na definição das informações estatísticas que se esperam obter após o pré-teste. Assim, terá relação também com as teorias e os modelos que serão assumidos para a análise dos dados do pré-teste. Se o objetivo é compor o teste definitivo tendo por base uma função de informação meta do teste (*target information function*), selecionando-se itens com base na função de informação de cada um deles (Hambleton, Jones & Rogers, 1993), por exemplo, a estrutura do pré-teste deve permitir que esses parâmetros sejam estimados. Hambleton e Jones (1993) consideram que "(...) em função da TRI requerer tamanhos de amostras grandes para obtenção de boas estimativas dos parâmetros dos itens, o desenvolvedor do teste deve selecionar uma amostra de examinandos com tamanho suficiente para garantir uma calibração acurada dos itens" (p. 44).

Por sua vez, caso se pretenda utilizar a TCT para análise dos resultados do Pré-teste, por considerar que apresenta informações suficientemente claras para um grupo de professores construírem o instrumento definitivo, deve-se preocupar em constituir testes com características de paralelos e delinear a amostra de examinandos representativa da população, pressupostos da teoria (Hambleton e Jones, 1993).

Geralmente, o desenho do pré-teste deve contemplar a inclusão de uma grande quantidade de itens. Assumir um delineamento em que todos os estudantes respondem a uma grande quantidade deles torna-se praticamente inviável. Johnson (1992) alerta para a deteriorização do desempenho dos estudantes em função dos efeitos da fadiga e da decrescente motivação em respostas a testes muito extensos.

Uma solução é a aplicação de instrumentos diferentes para grupos diferentes de examinandos. Pode ser viabilizado pela construção de blocos de itens e combinação por rotação desses para a construção de vários cadernos.

#### 3.3.2.2 Análise de dados do pré-teste

Os resultados do pré-teste podem ser analisados de acordo com a TCT (Hambleton & Jones, 1993; Pasquali, 2003) ou com a TRI (Cronbach, 1996; Hambleton & Jones, 1993; Hambleton, Jones, & Rogers, 1993; Hambleton, Swaminathan, & Rogers, 1991; Pasquali, 2003), considerando-se sempre as limitações quanto aos pressupostos de cada uma das teorias, bem como as vantagens de cada uma delas e de seus modelos associados. De forma geral, ambas fornecem informações relevantes para tomada de decisão dos itens que comporão o teste definitivo, bem como sugerir ajustes na formulação de itens.

Por meio da TCT, os índices p e r orientam a tarefa de desenvolvimento do teste definitivo. O $r_{bis}$ calculado por alternativa de itens de múltipla escolha fornece informações preciosas, pois permite indicar um possível distrator (alternativa incorreta) atrativo para os estudantes que se desempenharam bem no teste, o que não é esperado de um item discriminativo. Esses itens podem ser descartados ou mesmo sofrerem algum ajuste pontual, com base nas informações estatísticas, de forma a serem aproveitados no teste final.

Os parâmetros *a, b* e *c* estimados pela TRI, bem como a CCI e a FCI também orientarão a seleção dos itens do teste definitivo. De acordo com Hambleton, Jones e Rogers (1993) os modelos de resposta ao item traduzem-se em um poderoso método para a descrição e a seleção de itens. Ressalta-se a importância da (a) função de informação do item para a seleção de itens que cubram toda a extensão do traço e (b) da inclinação da curva característica do item para a seleção daqueles mais discriminativos.

Em suma, com base nos resultados do pré-teste, é possível calcular o poder discriminativo e a dificuldade dos itens que orientarão a decisão sobre sua permanência ou não no teste; indicar a existência de algum distrator não-plausível ou que está atraindo indevidamente ao erro alunos com maiores habilidades; indicar problemas de entendimento do enunciado ou das alternativas que impedem um bom desempenho dos estudantes com proficiências mais altas; indicar a chance que alunos com baixa habilidade têm de acertar um item mais difícil sem apresentar habilidade suficiente para tal.

Além de subsidiar a construção do teste, os resultados do pré-teste permitem orientar os procedimentos de aplicação e de padronização, o pré-teste pode orientar a

adoção de um tempo de aplicação adequado ao ritmo dos estudantes, verificar se as instruções previstas para a aplicação final são de claro entendimento, testar os procedimentos operacionais de distribuição de testes aos locais de aplicação, de treinamento dos aplicadores.

### 3.4 Características de testes e efeito nas estimativas de habilidade

De acordo com AERA, APA e NCME (1999), o processo de desenvolvimento de testes educacionais passa por quatro etapas: "(a) delineamento do propósito do teste e do escopo do construto ou extensão do domínio que será mensurado; (b) desenvolvimento e avaliação das especificações do teste; (c) elaboração, testagem de campo, avaliação e seleção dos itens e os procedimentos e guias para pontuação; e (d) montagem e avaliação do teste para utilização" (p. 37).

As três primeiras etapas foram tratadas nas seções anteriores do presente trabalho. Destaque para o papel do planejamento da avaliação e seu impacto em todas as etapas do desenvolvimento do teste. A quarta etapa "montagem e avaliação do teste" será tratada na presente seção.

O teste educacional é estruturado em alinhamento (Bhola, Impara & Buckendahl, 2003, p. 21; Herman, Webb & Zuniga, 2002, p. 1) com os conteúdos e domínios cognitivos selecionados e apresentados no *framework* da avaliação, especificamente nas matrizes de referência. Esses, por sua vez, apresentam relação com os objetivos educacionais e com os propósitos da avaliação. O grau de alinhamento traduz-se em evidências de validade baseada no conteúdo do teste (tradicionalmente denominada de validade de conteúdo).

Se as etapas de elaboração, revisão e de pré-testagem foram realizadas a contento, considerando os propósitos da avaliação, o modelo de análise de dados, os tipos de itens, a busca pela qualidade técnica dos itens e pelo alinhamento de cada questão à habilidade que se pretende avaliar, o teste é desenvolvido.

O teste deve cobrir a extensão do conteúdo ou do construto avaliado. O planejamento do teste, detalhado no *framework* ou nas especificações do teste, deve prever sua estrutura, indicando: os tipos de itens (múltipla escolha, resposta construída, etc.), o número de modelos de teste, o número de questões do teste como um todo e por modelo de teste, a distribuição dos itens pelos modelos e a ordem desses dentro de cada modelo.

De acordo com o planejamento do teste, a seleção dos itens deve considerar "(...) a qualidade e o escopo do construto a ser avaliado, os pesos dos itens e dos subdomínios e o quanto são apropriados para a população que responderá os testes" (AERA, APA &

NCME, 1999, p. 39). Considerando que os itens já apresentam boa qualidade pedagógica, a seleção daqueles que integrarão o teste deve ser orientada em função de seus índices estatísticos.

### 3.4.1 Seleção dos itens com base nas estatísticas

A TCT fornece um conjunto de informações para tomada de decisão dos itens que serão selecionados para o teste (índices *p* e *r*). Qual o percentual de itens fáceis, de dificuldade média e difíceis deve compor o teste? A resposta depende do propósito da avaliação. Se essa tiver o objetivo de discriminar examinandos que apresentam altos escores daqueles com escores mais altos ainda, por exemplo, remete à inclusão de um maior quantitativo de itens difíceis. Para avaliações diagnósticas, com o objetivo de analisar o percentual de estudantes com baixo, médio ou alto grau de habilidade, é fundamental a inserção de itens de dificuldades variadas, de forma a cobrir toda a extensão do traço latente. Pasquali (1996) considera que, utilizando o modelo da TCT, os itens "(...) devem cobrir toda a extensão de magnitude do traço e que os itens de dificuldade 50% são os que produzem maior informação. Pode-se sugerir que uma distribuição dos mesmos mais ou menos dentro de uma curva normal seria o ideal" (p. 83). Sugere a seguinte distribuição de itens por faixa de habilidade:

$0,0 < p < 0,2$: 10% dos itens;

$0,2 < p < 0,4$: 20%;

$0,4 < p < 0,6$: 40%;

$0,6 < p < 0,8$: 20%;

$0,8 < p < 1,0$: 10%.

Com relação à discriminação (*r*), o coeficiente de correlação bisserial (índice $r_{bis}$) calculado para cada uma das alternativas do item é uma poderosa ferramenta para seleção daqueles que farão parte do teste, como foi abordado anteriormente. Um cuidado a ser observado: se em uma população uma minoria de examinandos com altos escores acerta um item, o que pode indicar um conteúdo pouco assimilado, o índice $r_{bis}$ será baixo, mesmo se pedagógica ou tecnicamente o item seja muito bom. Nesse caso, a utilização dos resultados psicométricos pode considerar erroneamente que o item não apresenta boa qualidade técnica. Como segundo cuidado a ser observado, as estatísticas calculadas pela TCT dependem fortemente da aplicação do pré-teste em uma amostra representativa da população (Hambleton & Jones, 1993).

A TRI, por sua vez, fornece um "(...) poderoso método para descrição de itens e de testes e para seleção de itens quando se observa que os dados do teste se ajustam ao modelo" (Hambleton, Jones & Rogers, 1993, p. 144). Para utilização dos modelos de resposta ao item, é fundamental que os itens sejam aplicados em grandes amostras para viabilizar uma calibração adequada.

A curva característica do item (CCI) fornece um conjunto de informações que permite ao desenvolvedor do teste selecionar os itens que farão parte do teste. A partir da CCI, o desenvolvedor tem acesso às informações da discriminação, da dificuldade e da probabilidade de acerto ao acaso (parâmetros *a*, *b* e *c*) (Cronbach, 1996; Hambleton, Swaminathan, & Rogers, 1991; Hambleton & Jones, 1993; Hambleton, Jones & Rogers, 1993; Pasquali, 2003).

Hambleton, Jones e Rogers (1993) e Hambleton e Jones (1993), citando Lord (1980), indicam os passos a serem seguidos para o uso das funções de informação do item para construir testes:

(1) Decidir o formato desejado da função de informação do teste (função de informação meta).
(2) Selecionar itens do banco de itens com funções de informação que se ajustam à função de informação meta.
(3) Calcular a função de informação do teste para os itens selecionados.
(4) Continuar selecionando itens até que a função de informação do teste esteja satisfatoriamente de acordo com a função de informação meta.

A função de informação do teste está associada ao parâmetro *a* ou à inclinação das curvas dos itens. Procura-se selecionar itens com parâmetro *a* alto, sempre que possível, e não utilizar itens com baixa discriminação, ou seja, aqueles com "valor *a* baixo positivo ou negativo" (Hambleton & Jones, 1993, p. 45).

A função de informação meta, sugerida por Lord (1980), reflete o propósito da avaliação. Para avaliações diagnósticas, a função meta tende a se aproximar de uma curva normal. Os itens serão selecionados de acordo com o parâmetro *b*, necessariamente, orientados pela função de informação meta. Deve-se selecionar itens para ao vários intervalos de habilidade. De maneira prática, selecionam-se itens representativos de cada uma das faixas de habilidades, de acordo com a função de informação meta, e avalia-se o quanto a função de informação do teste se aproxima dessa, substituindo-se itens posteriormente se necessário (passos 2 a 4, citados por Lord, 1980).

### 3.4.2 Desenho do teste

O tamanho do teste é definido em função da cobertura da matriz de referência. Cobrir com um item cada habilidade da matriz pode gerar problemas para a validade e fidedignidade das estimativas de habilidade dos examinandos.

Caso algum item não apresente um funcionamento esperado, o teste deixa de cobrir completamente a matriz e sua validade fica comprometida. Esse desenho também não permite a inclusão de itens com complexidades cognitivas diferentes para cada habilidade, com impacto também na validade. Como a fidedignidade é diretamente proporcional ao tamanho do teste (Cronbach, 1996) e ao número de itens por habilidades, cobrir cada habilidade da matriz com apenas um item fornecerá estimativas de habilidade dos examinandos com um grau alto de erro. Uma alternativa para garantir um bom grau de validade e de fidedignidade seria aumentar o tamanho do teste e o número de itens por habilidade avaliada. Esse procedimento, no entanto, pode acarretar em fadiga ao testando por ter que responder a uma grande quantidade de itens.

A solução de testes-âncora pode ser utilizada, permitindo a aplicação de dois ou mais testes compostos por itens diferentes a dois ou mais grupos de examinandos. Um grupo de itens comuns às formas dos testes é aplicado e, por técnicas de equalização, os resultados são estimados na mesma escala, a partir de estatísticas derivadas dos itens comuns (Pasquali, 2003; Urbina, 2007). Sob esse delineamento, Pasquali (2003) faz a menção que o conteúdo do teste de ancoragem (itens comuns) deve ser representativo de todos os modelos de testes, como se fosse um miniteste com as mesmas características dos testes originais. Sob essa estrutura, consegue-se incluir uma maior número de itens, permitindo a cobertura das habilidades com um maior número deles.

O delineamento por Blocos Incompletos Balanceados (BIB) (Bekman, 2001; Johnson, 1992) é um esquema otimizado para o rodízio de blocos cuja utilização se justifica quando dispomos de *b* blocos e só podemos usar *k* deles em cada conjunto. Essa situação é recorrente quando se pretende compor uma prova com um número total de itens maior que o número que um sujeito poderia responder. Nesse caso, o BIB seria útil para que cada sujeito respondesse a apenas alguns blocos de itens. Bekman (2001) apresenta a origem da denominação Blocos Incompletos Balanceados.

a) Distribui-se certo número *b* de *blocos* de itens em um determinado número de cadernos de prova (*c*) de forma que cada caderno não seja composto pela totalidade dos blocos.

b) Como os cadernos não são compostos por todos os blocos, são chamados de *incompletos*. Cada um dos alunos recebe um subconjunto do total de blocos, ou seja, uma fração ($f_u$) do total de blocos.

c) A distribuição dos blocos é feita de forma *balanceada* em que cada caderno contenha o mesmo número de blocos *k;* cada bloco seja utilizado o mesmo número de vezes (*r*) dentro do conjunto total dos cadernos; e cada par de blocos é utilizado o mesmo número de vezes (λ) dentro do conjunto total dos cadernos.

Bekman (2001) utilizou a seguinte convenção (p. 121):

*c* = Número de cadernos.

*b* = Número de blocos.

*k* = Número de blocos em cada caderno.

*r* = Número de repetições de cada bloco no conjunto total dos cadernos.

$\lambda$ = Número de repetições de cada par de blocos no conjunto total dos cadernos.

$f_u$ = Fator de utilização.

Considera que para que haja um esquema solução BIB, é necessário que algumas soluções sejam satisfeitas, a partir da combinação de *c*, *b*, *r*, *k* e $\lambda$..

(i) $c = (r.b)/k$;

(ii) $\lambda = [r.(k-1)]/(b-1)$, em que *c*, *b*, *r*, *k* e $\lambda$, pertençam a N;

Para ser considerado um BIB espiral, as seguintes propriedades devem ser satisfeitas:

(iii) $c = n.b$, em que *n* pertença a N;

(iv) Os blocos devem estar distribuídos em espiral no conjunto dos cadernos de prova.

O fator de utilização ($f_u$) é definido pela razão entre o número de blocos de itens nos cadernos de prova face ao total de blocos de itens do estudo, de forma que:

(v) $f_u = k/b = r/c$

O fator de utilização também pode ser entendido como a proporção de respondentes que é submetida a determinado bloco dentro do total de respondentes. Bekman (2001) apresenta um estudo de oito exemplos de BIB, cujos resultados foram aqui sistematizados e apresentados na tabela 3.1.

Tabela 3.1 - Informações sobre exemplos de delineamentos BIB analisados por Bekman (2001).

| Exemplo | n cadernos (*c*) | n blocos (*b*) | *k* | *r* | *c=(r.b)/ k* | *λ=[r.(k-1) /(b-1)* | *n=c/b* | $f_u$=k/b | $f_u$=r/c | BIB |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 3 | 2 | 2 | 3 | 1 | 1 | 0,67 | 0,67 | Espiral |
| 2 | 7 | 7 | 3 | 3 | 7 | 1 | 1 | 0,43 | 0,43 | Espiral |
| 3 | 10 | 5 | 2 | 4 | 10 | 1 | 2 | 0,40 | 0,40 | Espiral |
| 4 | 13 | 13 | 4 | 4 | 13 | 1 | 1 | 0,31 | 0,31 | Espiral |
| 5 | 20 | 16 | 4 | 5 | 20 | 1 | 1,25 | 0,25 | 0,25 | Não-espiral |
| 6 | 21 | 7 | 2 | 6 | 21 | 1 | 3 | 0,29 | 0,29 | Espiral |
| 7 | 21 | 21 | 5 | 5 | 21 | 1 | 1 | 0,24 | 0,24 | Espiral |
| 8 | 26 | 13 | 3 | 6 | 26 | 1 | 2 | 0,23 | 0,23 | Espiral |

Todos os exemplos acima podem ser considerados BIB, pois atendem os pressupostos *i* e *ii* apresentados. Para ser considerado espiral, o BIB deve apresentar *n* natural (número de vezes que cada bloco aparece em cada posição), o que aconteceu para todos os exemplos apresentados, com exceção para o exemplo 6. Neste caso, *n* foi 1,25, não atendendo ao pressuposto número *iii* apresentado.

Quando é necessário distribuir um grande quantitativo de itens em vários cadernos de prova com poucos itens cada, o BIB é bastante útil. Os esquemas BIB permitem que os itens sejam respondidos aproximadamente pelo mesmo número de alunos da amostra, os respondentes recebam cadernos com o mesmo número de blocos; os cadernos não contenham blocos repetidos; e cada par de blocos seja submetido ao mesmo número de respondentes (Johnson, 1992).

Para escolha do BIB mais adequado, Bekman (2001) sugere que o ideal seria inserir o maior número de itens na prova, mantendo-se uma quantidade aceitável de itens nos cadernos (menor $f_u$ possível). Na prática, isso nem sempre é possível, pois se consideram as seguintes limitações:

a) Para estimar os parâmetros dos itens por meio da TRI, é necessário que cada um deles seja respondido por um número mínimo de alunos. O autor sugere

que cada item seja respondido por pelo menos 200 alunos de forma que: *($f_u$ >200)/número total de respondentes*.

b) Não é interessante que cada caderno contenha muitos blocos k.

c) Não é interessante montar muitos blocos b e muitos cadernos c.

O número de itens inseridos em cada bloco merece destaque, pois tem impacto direto na validade, no que tange à cobertura da matriz de referência, e na fidedignidade dos resultados da avaliação. Johnson (1992) alerta para a relevância da realização de estudos sobre a fidedignidade quando poucos itens são utilizados para a estimação da performance individual dos sujeitos. "Quando muitos sujeitos recebem poucos itens de uma determinada área, resulta uma considerável imprecisão na estimação das proficiências individuais" (p. 105). O autor sugere para esse caso que a tecnologia de valores plausíveis seja utilizada para o alcance de estimativas fidedignas. O aumento do número de itens por bloco e, consequentemente, o aumento do número de itens que cada estudante responde reduz a necessidade de utilizar metodologias como valores plausíveis para estimar a fidedignidade das estimativas de proficiência.

### 3.4.3 Dimensionalidade

Se um conjunto de itens mede um mesmo traço latente, considera-se que apresentam um bom grau de unidimensionalidade. Trata-se de um pressuposto da TCT e da TRI que apresenta impacto na validade dos resultados do teste.

No caso da TCT, um teste com bom grau de unidimensionalidade é aquele cujos itens apresentam uma boa correlação com o escore total. Pasquali (2003) alerta para os problemas de verificação da dimensionalidade utilizando a TCT, pois "o escore total consiste na soma das respostas dadas aos itens; assim, ela faz a suposição que eles são somáveis e isto faz sentido somente se eles referem à mesma coisa (...)" (p. 114). A incoerência ocorre quando um item não contribui significativamente para a unidimensionalidade e é utilizado para o cálculo do escore total.

No âmbito da TRI, unidimensionalidade também é um pressuposto em que apenas uma habilidade é medida por um conjunto de itens em um teste. Praticamente, um teste é unidimensional se apresenta um componente ou fator dominante que influencia o desempenho dos examinandos.

Para a estimação dos parâmetros dos itens e das habilidades pela TRI, a verificação da unidimensionalidade da prova utilizada se torna fundamental. Laros, Pasquali e

Rodrigues (2000) apresentaram quatro efeitos negativos que podem surgir quando é violado o pressuposto da unidimensionalidade dos itens na utilização da TRI: (a) diminuição da validade de construto do teste, dificultando a interpretação dos escores; (b) aumento da função diferencial do item; (c) dificuldade de realização da equalização dos resultados de várias formas de uma prova; e (d) probabilidade do parâmetro de habilidade, dado o padrão de resposta, não é válida e as estimativas e os desvios-padrão do parâmetro podem ser errôneos.

Os autores 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" (p. 12). Concordam com o proposto por Hattie (1985), que os índices baseados na TRI são os mais adequados para a verificação da unidimensionalidade.

Laros, Pasquali & Rodrigues (2000) analisaram ainda a dimensionalidade das provas do SAEB aplicadas em 1997 utilizando esse método e alguns índices complementares porcentagem de variância explicada pelo primeiro fator, a correlação bisserial item-total e a correlação tetracórica entre os itens. Os resultados para a prova de matemática, 8ª série, com 161 itens, 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 apresenta unidimensionalidade. No entanto, nem todos os itens contribuíram igualmente 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 que, após a exclusão destes itens que praticamente não contribuem para a unidimensionalidade, a prova de matemática pode ser considerada unidimensional e pode ser analisada pela TRI, sem a violação do seu pressuposto principal.

Condé (2002) e Condé e Laros (2007) investigaram se a estimativa de habilidade pela TRI independe da dificuldade dos itens utilizados para estimá-la, bem como em que medida a unidimensionalidade da prova influencia na propriedade de invariância da habilidade dos sujeitos. Concluíram que a estimativa de habilidade da TRI apresenta uma diminuição da dependência com relação à dificuldade quando a prova se aproxima da unidimensionalidade. Percebe-se necessário um maior rigor no controle da condição de unidimensionalidade da prova para a obtenção de estimativas de habilidade mais invariantes.

### 3.4.4 Tamanho do teste e tempo de resposta

O tamanho do teste e o número de questões que cada estudante responde são planejados levando em consideração o tempo que terá disponível para conclusão da prova. Testes muito extensos podem levar ao cansaço, à impossibilidade de respondê-lo completa ou adequadamente, gerando baixa confiabilidade de resultados.

Uma série de estudos investigou a influência da velocidade (*speededness*) em função de tempo insuficiente na resposta a testes na validade e na precisão dos resultados (Oshima, 1994; Bolt, Cohen e Wollack, 2002; Sireci, 2005; Sireci, Scarpati e Li, 2005; Lu e Sireci, 2007).

Em muitos contextos avaliativos, observam-se testes com tempo delimitado, mesmo que seu objetivo não seja avaliar a velocidade em que os estudantes os respondem. Geralmente, utilizam-se uma padronização referente ao tempo de aplicação em função da organização, da conveniência e do custo.

Nos testes de velocidade (*speed tests*) pelo menos parte do construto a ser medido deve ser dependente da velocidade nas respostas. Testes de potência (*power tests*), por sua vez, são compostos de questões interessadas exclusivamente na performance do sujeito. Se um teste de potência é utilizado no âmbito de uma aplicação com tempo determinado e esse não é suficiente para pelo menos uma amostra de respondentes, seus resultados podem estar enviesados, já que o construto medido não é exclusivamente o desempenho, mas o desempenho associado à velocidade da resposta.

De acordo com Lu e Sireci (2007), Sireci (2005) e Sireci, Scarpati e Li (2005), quando o tempo limite para administração de teste de potência é estabelecido exclusivamente para propósitos práticos, é desejável analisar se os examinandos possuem tempo suficiente para responder completamente a todos os itens, sob pena da rapidez de resposta prejudicar a validade e a precisão dos resultados da testagem.

A velocidade "(...) introduz uma variância irrelevante do construto no escore do teste, mudando o próprio construto que se pretende medir" (Lu e Sireci, 2007, p. 31). Como não se sabe a partir de que item o examinando passou a responder sem critério, aceitaremos que os erros às questões se devem à ausência do construto e não a outro motivo.

É possível que o examinando responda com atenção a um conjunto de itens, mas deixe outros em branco no final. Neste caso, embora se saiba com mais certeza até qual item o estudante respondeu, corre-se o risco da validade ficar comprometida, já que há uma

perda concentrada de respostas para os últimos itens do teste. Como para garantir a validade de conteúdo dos resultados do teste, os últimos itens do teste são necessários, a perda concentrada de itens gerará uma sub-exploração de alguns conteúdos e habilidades.

Oshima (1994) realizou uma simulação em que estimou, por meio da TRI, os parâmetros *a, b* e *c* dos itens localizados nas últimas posições dos blocos, bem como o parâmetro de habilidade. Encontrou que os parâmetros *a* e *b* foram subestimados e o parâmetro *c*, superestimado na grande maioria dos itens. Como esses parâmetros serão apresentados ao software utilizado para análise dos dados como base para a estimação das habilidades, esperava-se também uma influência nesses resultados. Oshima (1994) concluiu que "(...) a velocidade pode contribuir levemente para a distorção da estimação da habilidade" (p. 214). O autor recomenda que, em situações de velocidade para testes de potência, os itens sejam apresentados em ordem crescente de dificuldade e que a opção "não-apresentados" (*not-presented* ou *not-reached*) do software BILOG (Bock & Zimowski, 1995) seja atribuída aos itens. Itens não-apresentados são muitas vezes identificados como aqueles não respondidos após a última resposta do examinando. Esta estratégia é pautada em uma inferência, já que na prática, fica muito difícil saber a partir de qual item o estudante não teve tempo disponível para responder. Já Lord (1980) sugeriu uma estratégia mais conservadora: os itens em situação de velocidade deveriam ser excluídos da análise se a estimativa dos estudantes fosse estimada. Por sua vez, Oshima (1994) considerou que "excluir itens não respondidos a análise pode gerar sérios efeitos se um grupo étnico particular tende a ter um maior grau de omissão" (p. 214). Os estudos da velocidade em testes de poder indicam também a possibilidade de encontrarmos Função Diferencial dos Itens (DIF) localizados nas últimas posições do teste. Isso ocorre porque dois grupos de mesma habilidade terão diferentes probabilidades de acertar a esses itens (Oshima, 1994).

## 4. Sistema Nacional de Avaliação da Educação Básica - SAEB

O SAEB é uma avaliação de monitoramento em larga escala que tem, da década de noventa até os dias atuais, embasado uma série de estudos na área de psicometria e de educação. Para o desenvolvimento do presente estudo, cabe aqui um detalhamento desse sistema de avaliação.

O Sistema avalia periodicamente estudantes da 4ª e 8ª séries EF e da 3ª série EM para monitorar a qualidade educacional e fornecer aos agentes educacionais e à sociedade informações sobre os resultados dos processos de ensino. De 1990 a 2007 vêm fornecendo

informações sobre o desempenho dos estudantes sobre diversas disciplinas, fundamentalmente, língua portuguesa e matemática.

### 4.1 O que o SAEB avalia?

Falar que o SAEB avalia língua portuguesa e matemática não é muito esclarecedor. Avalia habilidades e competências? Se sim, quais? Em língua portuguesa avalia ortografia, gramática, leitura? De acordo com quais perspectivas teóricas? Com o intuito de esclarecer questões como estas, parte-se para uma exploração do marco teórico que subsidia a construção das provas do SAEB.

Em 2001, foram constituídas as matrizes de referência do SAEB utilizadas pela avaliação de 2001 a 2007. O documento "SAEB 2001: Novas Perspectivas" (INEP, 2002) apresenta as matrizes, bem como os pressupostos teóricos que subsidiaram sua elaboração e a composição dos testes.

O SAEB busca avaliar o nível de competência dos estudantes em se trabalhar com conteúdos das disciplinas. Adota a concepção de competência apresentada por Perrenoud (1993), que é a "capacidade de agir eficazmente em um determinado tipo de situação, apoiando-se em conhecimentos, mas sem se limitar a eles" em que o estudante se utiliza de vários recursos cognitivos complementares em suas ações, dentre os quais os conhecimentos. As competências cognitivas para o SAEB são "(...) as diferentes modalidades estruturais da inteligência que compreendem determinadas operações que o sujeito utiliza para estabelecer relações com e entre os objetos físicos, conceitos, situações, fenômenos e pessoas uma situação, geralmente, colocam-se em ação vários recursos cognitivos" (INEP, 2002, p. 11).

O SAEB também trabalha com o conceito de habilidades instrumentais que "(...) referem-se especificamente ao plano do saber fazer e decorrem, diretamente, do nível estrutural das competências já adquiridas e que se transformam em habilidades" (INEP 2002, p. 11). A opção teórica, de natureza cognitivista, adotada nas Matrizes de Referência do SAEB para a construção dos descritores, prioriza, portanto, a avaliação de conteúdos na perspectiva das competências e habilidades neles implícitas.

As matrizes de referência do SAEB são compostas por descritores orientados nos conteúdos, competências e habilidades. "(...) Têm como pressuposto epistemológico o fato de que os conteúdos científicos, matemáticos, lingüísticos, históricos, etc., se constituem de princípios, conceitos e informações relacionadas por operações intelectuais

(classificação, seriação, correspondência, causa e efeito, correlação, implicação, etc.)" (INEP, 2002, p. 12).

Em língua portuguesa, a estrutura teórica do SAEB, a partir de 2001, teve por base a concepção dos Parâmetros Curriculares Nacionais (PCN) de que o ensino deve contribuir para o desenvolvimento do uso da linguagem de forma a ampliar as possibilidades dos estudantes na participação social e no exercício da cidadania. Assim, a escola tem o papel de fornecer subsídios para o efetivo desenvolvimento de competências e habilidades fundamentais para o domínio dos usos lingüísticos. A competência no uso da linguagem possibilita a compreensão e a produção de textos orais e escritos adequados às situações de comunicação em que atual. Ainda, "(...) posicionar-se criticamente diante do que lê ou ouve; de ler e escrever produzindo sentido, formulando perguntas e articulando respostas significativas em variadas situações" (INEP, 2002, p. 17).

Embora o ensino da língua portuguesa se paute em práticas de compreensão e de produção de textos, de análise lingüística, fica impraticável para um sistema de avaliação contemplar toda essa amplitude. No SAEB, a partir de 2001, decidiu-se avaliar exclusivamente habilidades de leitura, dentro da concepção que "um bom leitor, além de mobilizar esquemas cognitivos básicos, de ativar conhecimentos prévios partilhados e relevantes ao contexto, recorre a seus conhecimentos lingüísticos para ser capaz de perceber os sentidos, as intenções – implícitas e explícitas – do texto e os recursos que o autor utilizou para significar e atuar verbalmente" (INEP, 2002, p. 18).

Da mesma forma, quando se trabalha com a concepção de competências cognitivas, para matemática, não se pode considerar prioritário o ensino de matemática por meio de memorização de fórmulas, de regras e de técnicas. O SAEB é desenvolvido sobre a perspectiva de ensino da matemática que considera a resolução de problemas como eixo norteador, pois "possibilita o desenvolvimento de capacidades como: observação, estabelecimento de relações, comunicação (diferentes linguagens), argumentação e validação de processos, além de estimular formas de raciocínio como intuição, indução, dedução e estimativa" (INEP, 2002, p. 22). Assim, as matrizes de referência de matemática têm por base as competências em conteúdos matemáticos desenvolvidos na escola (e fora dela) e que são passíveis de serem verificadas por meio de avaliações escritas.

Para as duas disciplinas, as matrizes não podem ser consideradas parâmetros para a elaboração de estratégias de ensino na escola, papel esse dos parâmetros, dos currículos e das diretrizes curriculares, mas exclusivamente um documento que orienta a elaboração da avaliação.

## 4.2 Matrizes de referência

Optando-se como foco da avaliação Leitura, para língua portuguesa, e Resolução de Problemas, para matemática, as Matrizes de Referência do SAEB 2001 (INEP, 2002) foram estruturadas para cada série e disciplina, a partir de listas de habilidades associadas a conteúdos.

As matrizes do SAEB 2001 foram constituídas tendo por base as Matrizes Curriculares de Referência do SAEB, utilizadas em 1999 (INEP, 1999), a Lei de Diretrizes e Bases da Educação (LDB), e os resultados de uma consulta às equipes de ensino e professores regentes de turmas das cinco regiões do País. Esses verificaram a compatibilidade entre as matrizes então vigentes e o currículo proposto pelos sistemas estaduais para cada disciplina.

O menor elemento da matriz é o descritor, que representa uma determinada habilidade ou comportamento. Um exemplo de descritor das matrizes de língua portuguesa é: "D1 – Localizar informações explícitas em um texto". Trata-se de uma habilidade apresentada nas matrizes das três séries. No entanto, não necessariamente um descritor contido em uma matriz será contemplado em todas as séries.

As matrizes de língua portuguesa são constituídas por 15 descritores na 4ª série EF e 21 descritores em 8ª série EF e em 3ª série EM. Os descritores de língua portuguesa estão categorizados em seis tópicos, que representam grandes estruturas de descritores. Os tópicos são: "I – Procedimento de Leitura"; "II – Implicações do Suporte, do Gênero e/ou do Enunciador na Compreensão do Texto"; "III – Relação entre Textos"; "IV – Coerência e Coesão no Processamento do Texto"; "V – Relações entre Recursos Expressivos e Efeito de Sentido"; e "VI – Variação Lingüística" (INEP, 2002, p. 19-22).

Um exemplo de descritor de matemática: "D1 – Identificar a localização/ movimentação de objeto em mapas, croquis e outras representações". Está localizado tanto na matriz de 4ª série EF, quanto na de 8ª série EF. Este descritor não é contemplado no Ensino Médio.

As matrizes de matemática são constituídas por 28 descritores para 4ª série EF, por 37 para 8ª série EF e por 35 para 3ª série EM. Para as três séries, as matrizes categorizam seus descritores em quatro temas: "I – Espaço e forma"; "II – Grandezas e Medidas"; "III – Números e Operações/ Álgebra e Funções"; e "IV – Tratamento da Informação" (INEP, 2002, p. 25-28).

Associada às matrizes do SAEB 2001, uma proposta de hierarquia de prioridades para tópicos/temas e descritores, em função de sua pertinência para cada uma das séries foi elaborada para orientar a construção dos testes (INEP, 2002). A proposta orientaria a construção do teste ao sugerir um número maior de itens os temas ou tópicos e descritores considerados pedagogicamente mais relevantes para cada disciplina e série. Desta forma, quanto mais próximo da prioridade 1 (P1), mais relevante seria o tópico/tema ou o descritor para a série.

Cabe ressaltar a importância da utilização de um modelo de prioridades, no que se refere à validade do teste. Se, pedagogicamente, um aspecto da competência é mais importante que outro para a resolução de problemas matemáticos, o teste deve ser capaz que contemplar essa diferença. Os testes do SAEB 2001, 2003, 2005 e 2007 foram compostos por itens construídos tendo por base as matrizes de 2001.

### 4.3 Testes

Até 1993, o SAEB utilizou provas clássicas para avaliar o desempenho dos estudantes. Esse formato de instrumento é limitado em função da impossibilidade de cobertura de uma matriz que abranja a amplitude do construto, trazendo impacto para a validade de seus resultados. O modelo clássico dificulta ainda a inserção no teste de um número razoável de itens total e por descritor, o que traz impacto para a fidedignidade de seus resultados. Sobre esse aspecto, cabe lembrar que testes com um maior número de itens apresentam resultados mais fidedignos (Cronbach, 1996).

Para corrigir limitações geradas pela instrumentação clássica, o SAEB, a partir de 1995, passou a utilizar um número maior de itens. Nos SAEB 1999, 2001, 2003 e na ANEB 2005 foram aplicados 169 itens, o que possibilitou uma ampla cobertura dos descritores. Não seria viável para um estudante responder a esse quantitativo de itens em função do tempo e do cansaço. Por isso, para viabilizar a utilização desse grande número de itens, o SAEB incorporou a metodologia baseada na amostragem matricial de itens, que utiliza o esquema de montagem e aplicação de provas por BIB (Bekman, 2001; Johnson, 1992).

Do SAEB 1999 ao 2003 e na ANEB 2005, foram montados 26 cadernos (*c*) a partir da composição e combinação de 13 blocos (*b*) de 13 itens, de acordo com a orientação BIB apresentado na tabela 4.1.

Tabela 4.1 - Delineamento de Blocos Incompletos Balanceados (BIB) para 26 cadernos.

| **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 |

Utilizando-se a notação proposta por Bekman (2001), as características desse delineamento de composição de testes são representadas abaixo:

$c = 26$

$b = 13$

$k = 3$

$r = 6$

$\lambda = 1$

$f_u = 0{,}23$

Essa distribuição de itens por blocos e combinação de blocos por cadernos ($k$) permite que um mesmo conjunto de itens esteja localizado na primeira posição (primeiro bloco) em dois cadernos de teste, na segunda posição, em outros dois cadernos e na terceira posição, em outros dois, o que o caracteriza como desenho espiralado. 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.

Os testes do SAEB, a partir de 2001, foram compostos exclusivamente por itens de múltipla escolha, com quatro alternativas e uma resposta correta para as 4ª e 8ª séries EF e

com quatro e cinco alternativas com uma resposta certa para a 3ª série EM, com base nas Matrizes de Referência do SAEB 2001 (INEP, 2002). O número de itens por tópicos/temas e por descritor foi calculado a partir do estudo de prioridades do SAEB 2001 (INEP, 2002). Nos casos, buscou-se reservar um número maior de itens para os descritores com prioridades mais próximas de P1.

Para o desenvolvimento dos testes do SAEB, foram compostos 13 blocos para cada série e disciplina avaliadas, cujas características dividem-nos em dois tipos quanto à sua origem ou utilização ou não em avaliações anteriores: (a) blocos de itens inéditos; e (b) blocos de itens do SAEB do ano anterior, que foram utilizados nos testes atuais para efeitos de comparação dos resultados entre anos.

No caso de todas as séries e disciplinas do SAEB 2003 e da ANEB 2005, onze blocos inéditos e dois do ciclo anterior da avaliação foram utilizados (comuns entre anos). Para 8ª série EF e 3ª série EM, três blocos inéditos são oriundos das séries anteriores (comuns entre séries). O procedimento de utilização de itens comuns permite que os resultados sejam estimados na métrica da escala única do SAEB (1995 a 2007; 4ª e 8ª séries EF e 3ª série EM).

O delineamento do SAEB foi adotado com o objetivo de emissão de resultados para estratos amostrais e não para cada escola ou para cada estudante. Sendo assim, não há necessidade que a dificuldade dos cadernos ou dos blocos que os compõem sejam iguais. Para o menor estrato de divulgação de resultados do SAEB, um mesmo número de estudantes, com os mais variados níveis de habilidades, respondem a cada um dos 26 modelos de cadernos (e a cada um dos 13 blocos), pois sua distribuição é aleatória pela amostra. Se compararmos os resultados dos estudantes de escolas públicas do Pará com os dos estudantes de escolas públicas do Ceará, por exemplo, uma mesma proporção de estudantes respondeu aos cadernos mais fáceis e mais difíceis, aspecto este que minimiza o impacto de uma possível diferença das dificuldades dos blocos.

De toda forma, identificou-se para alguns anos de avaliação, como por exemplo, para o SAEB 2003, a preocupação em se compor blocos com características de dificuldade semelhantes (com pouca variabilidade entre os blocos). Nesse ano, utilizou-se como base os valores *p* dos itens extraídos do pré-teste, no caso de itens inéditos, e do SAEB 2001, no caso de itens já aplicados. Para 4ª e 8ª séries EF, por sua vez, procurou-se compor três blocos de itens com características mais apropriadas para a série posterior, o que geraria blocos mais difíceis.

Para a montagem dos testes do SAEB, além de ter sido considerado o planejamento para o teste como um todo, considerou-se também um planejamento para cada um dos blocos. Os itens foram distribuídos dentro dos blocos de acordo com os seguintes critérios: (a) variedade de descritores, tópicos ou temas para cada bloco de itens; para língua portuguesa, inclui-se o critério de variedade de textos por tipologia textual; e (b) variabilidade e ordenamento dos itens pelo índice *p* de acordo com as informações levantadas pelos pré-testes realizados.

Considerou-se, na composição de cada bloco, uma variedade de descritores e um número de itens por temas ou tópicos coerentes com o planejamento de prioridades. Observa-se na tabela 4.2 o exemplo da composição do bloco 1 do teste de matemática, 4ª Série, do SAEB 2003. Ressalta-se que a tabela está ordenada por temas e por descritores e não reflete a ordem em que os itens foram aplicados.

Tabela 4.2 - Temas e descritores dos itens que compõem o bloco 1 do teste de matemática, 4ª Série EF, do SAEB 2003.

| Item | Tema | Prioridade | Descritor |
|---|---|---|---|
| 1 | I | 3 | 1 |
| 2 | | | 3 |
| 3 | II | 2 | 6 |
| 4 | | | 8 |
| 5 | | | 10 |
| 6 | | | 12 |
| 7 | III | 1 | 13 |
| 8 | | | 14 |
| 9 | | | 15 |
| 10 | | | 17 |
| 11 | | | 19 |
| 12 | | | 23 |
| 13 | IV | 4 | 28 |

Nota-se que: (a) todos os temas da matriz foram incorporados, (b) nenhum descritor foi repetido, (c) um número maior de itens foi utilizado para temas com prioridades maiores.

A montagem de outros blocos dessa série contemplou os mesmos critérios, com o diferencial de incorporarem os outros descritores com itens cujos descritores não tinham sido utilizados nesse bloco. Para os onze blocos inéditos de matemática, 4ª série EF do SAEB 2003, em relação aos temas, os blocos apresentaram quase sempre a mesma

estrutura, caracterizando uma espécie de paralelismo entre eles. Uma grande vantagem da utilização deste modelo de montagem dos blocos é que, utilizando o BIB na composição dos 26 cadernos de teste, todos os cadernos apresentariam estruturas semelhantes. Para o SAEB 2001 e 2003, procurou-se compor os testes, considerando para cada bloco, uma distribuição que contemplasse dificuldades baixas, médias e altas.

Em 2005, o SAEB foi dividido em dois processos de avaliação: a ANEB e a Prova Brasil. Diversas características da ANEB 2005 já foram apresentadas em função da sua semelhança ao SAEB tradicional. A seguir outras características da ANEB e a estrutura da Prova Brasil serão apresentadas como base para o desenvolvimento do presente trabalho.

### 4.4 ANEB 2005

A Avaliação Nacional da Educação Básica (ANEB) é o componente amostral do SAEB 2005. Forneceu informações sobre o desempenho dos estudantes brasileiros de 4ª e 8ª séries EF e de 3ª série EM em língua portuguesa e matemática. As provas foram aplicadas em 194.822 estudantes de 5.940 escolas, públicas ou particulares. A distribuição dos estudantes por série é apresentada na tabela 4.3.

Tabela 4.3 - Número de alunos avaliados na ANEB 2005.

| Série | Número de alunos |
|---|---|
| 4ª EF | 83.929 |
| 8ª EF | 66.353 |
| 3a EM | 44.540 |
| Total | 194.822 |

(Fonte: INEP, 2007a)

Os alunos selecionados compõem amostras aleatórias, probabilísticas e representativas da população de referência. O parâmetro é composto por todos os estudantes matriculados na série. 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 da ANEB 2005 é estratificada, levando-se em conta as variáveis de escolas: zona (rural ou urbana) e dependência administrativa (estadual, municipal ou particular). Os resultados de desempenho dos estudantes podem ser calculados e divulgados (a) para cada grupo de escolas urbanas, estaduais, municipais e particulares por unidade da federação, regiões e Brasil; (b) para o conjunto de escolas rurais, exclusivamente para 4ª série e em nível de Regiões (nunca em nível de unidades da federação); (c) para o conjunto de escolas federais apenas em nível Brasil (INEP, 2007a). Não é possível a apresentação dos resultados de desempenho dos estudantes por escolas ou por município, já que a amostra do SAEB não é preparada para isso.

A estimação das habilidades dos estudantes da ANEB 2005 foi realizada tendo por base a TRI, sob modelo logístico de três parâmetros. Os parâmetros dos itens foram estimados, por série e por disciplina, na métrica da escala SAEB.

### 4.5 Prova Brasil 2005

A Prova Brasil tem como objetivos: (a) avaliar a qualidade do ensino ministrado nas escolas, de forma que cada unidade escolar receba o resultado global; (b) contribuir para o desenvolvimento, em todos os níveis educativos, de uma cultura avaliativa que estimule a melhoria dos padrões de qualidade e eqüidade da educação brasileira e adequados controles sociais de seus resultados; e (c) concorrer para a melhoria da qualidade de ensino, redução das desigualdades e a democratização da gestão do ensino público nos estabelecimentos oficiais, em consonância com as metas e políticas estabelecidas pelas diretrizes da educação nacional (D.O.U., n.85, Portaria n. 69, de 4 de maio de 2005).

A Prova Brasil 2005 teve como universo todos os alunos das Escolas Públicas (estaduais, municipais e federais), 4ª e 8ª séries EF matriculados em escolas situadas na zona urbana e que tenham pelo menos 30 alunos de acordo com o censo preliminar de 2005. Todos os alunos desse universo foram selecionados para realizarem a Prova Brasil.

Exclusivamente, no caso da Rede Estadual de São Paulo foi extraída uma amostra de alunos de cada escola e série pertencente ao universo definido, de acordo com os seguintes critérios: “(a) se a escola tem até 3 turmas, sorteia-se uma; (b) se a escola tem 4 turmas ou mais, sorteiam-se duas.” (CESGRANRIO, 2006, p. 1).

Se na ANEB 2005, cada aluno respondeu a um caderno de uma disciplina, língua portuguesa ou matemática, na Prova Brasil 2005, todos os alunos responderam a testes das duas disciplinas.

De acordo com CESGRANRIO (2006), "(...) o planejamento dos cadernos de teste em cada disciplina seguiu um planejamento em blocos incompletos balanceados (BIB) com 7 blocos, 21 cadernos compostos de 2 blocos, cada bloco aparecendo 3 vezes em cada posição. Cada caderno de teste tinha 4 blocos, 2 de língua portuguesa e 2 de matemática. Os cadernos de prova de número ímpar começaram com língua portuguesa e os de número par com matemática." (p. 1). O esquema dos cadernos da Prova Brasil é apresentado na tabela 4.4.

Tabela 4.4 - Delineamento de Blocos Incompletos Balanceados (BIB) da Prova Brasil.

| Caderno | Disc 1 | Blocos | | Disc 2 | Blocos | |
|---|---|---|---|---|---|---|
| | | Posição 1 | Posição 2 | | Posição 1 | Posição 2 |
| 1 | P | 1 | 2 | M | 1 | 2 |
| 2 | M | 2 | 3 | P | 2 | 3 |
| 3 | P | 3 | 4 | M | 3 | 4 |
| 4 | M | 4 | 5 | P | 4 | 5 |
| 5 | P | 5 | 6 | M | 5 | 6 |
| 6 | M | 6 | 7 | P | 6 | 7 |
| 7 | P | 7 | 1 | M | 7 | 1 |
| 8 | M | 1 | 3 | P | 1 | 3 |
| 9 | P | 2 | 4 | M | 2 | 4 |
| 10 | M | 3 | 5 | P | 3 | 5 |
| 11 | P | 4 | 6 | M | 4 | 6 |
| 12 | M | 5 | 7 | P | 5 | 7 |
| 13 | P | 6 | 1 | M | 6 | 1 |
| 14 | M | 7 | 2 | P | 7 | 2 |
| 15 | P | 1 | 4 | M | 1 | 4 |
| 16 | M | 2 | 5 | P | 2 | 5 |
| 17 | P | 3 | 6 | M | 3 | 6 |
| 18 | M | 4 | 7 | P | 4 | 7 |
| 19 | P | 5 | 1 | M | 5 | 1 |
| 20 | M | 6 | 2 | P | 6 | 2 |
| 21 | P | 7 | 3 | M | 7 | 3 |

(Fonte: CESGRANRIO, 2006)

Desconsiderando a variação entre as posições das disciplinas dentro dos cadernos, observa-se um BIB espiralado já que as soluções apresentadas por bekman foram satisfeitas: (i) c = (r.b)/k; (ii) $\lambda$ = [r.(k-1)]/(b-1); (iii) c = n.b; e (iv) os blocos estão distribuídos em espiral no conjunto dos cadernos de prova.

O delineamento é utilizado para as duas séries, sendo que para a 4ª série EF, cada bloco é composto por 10 itens e para a 8ª série EF, 12 itens. Assim, cada aluno de 4ª série EF responde a um caderno de 40 itens (20 de cada disciplina) e cada aluno de 8ª série EF responde a um caderno de 48 itens (24 de cada disciplina).

Nos testes compostos para um programa educacional específico do Estado do Rio de Janeiro, os blocos 6 e 7 foram substituídos por blocos de itens fornecidos pelo próprio programa. Com exceção dos alunos da rede estadual de São Paulo em que foi extraída uma amostra, não houve ponderação sobre os alunos respondentes da Prova Brasil. Aos alunos amostrados de São Paulo,"(...) foi atribuído o peso igual ao numero de turmas na escola dividido pelo número de turmas sorteadas." (CESGRANRIO, 2006).

Da mesma forma como na ANEB, a estimação das habilidades dos estudantes da Prova Brasil foi realizada tendo por base a TRI, sob modelo logístico de três parâmetros. Na Prova Brasil 2005, os parâmetros dos itens foram estimados, por série e por disciplina, na métrica da escala SAEB. Todas as séries e as disciplinas da Prova Brasil continham itens comuns oriundos do SAEB 2003 para permitir o vínculo com a escala do SAEB. A calibração dos itens da Prova Brasil utilizou, para cada série, os parâmetros desses itens que já tinham sido estimados para o SAEB 2003.

A estimação do parâmetro de habilidade ou a equalização dos resultados foi realizada "(...) utilizando uma amostra seqüencial de 10% dos respondentes, obtida após ordenação por código (do IBGE) de UF, por dependência administrativa (estadual, municipal e federal), por município (em ordem alfabética), por código de escola (do Censo Escolar), por código de turma, por código de aluno" (CESGRANRIO, 2006).

### 4.6 Comparação da ANEB 2005 com a Prova Brasil 2005

A decisão da ampliação do SAEB, por meio do desmembramento do sistema em ANEB e Prova Brasil teve origem política e ancorada pela demanda da sociedade, dos professores e dos gestores educacionais por resultados de desempenho dos estudantes para escolas e municípios. Em 2005, as avaliações foram executadas separadamente, em um espaço de tempo de cerca de um mês, mas envolveram pelo menos um público em comum:

uma parcela de estudantes de 4ª e 8ª séries EF de escolas públicas urbanas com mais de 30 alunos.

A ANEB envolveu uma amostra de estudantes, cujas estimativas de habilidade, por meio de peso amostral, foram expandidas para a população. Já a Prova Brasil pretendeu avaliar o universo por definição, com exceção da Rede Estadual de São Paulo. Assim, exclusivamente para esse estrato, foram considerados pesos amostrais diferentes de 1. A tabela 4.5 apresenta o número de estudantes caracterizados como público-alvo das duas avaliações com características em comum.

Tabela 4.5 - Número de alunos avaliados na ANEB 2005 e na Prova Brasil 2005 de escolas públicas urbanas com mais de 30 alunos.

| Disciplina | Série | ANEB | Prova Brasil |
|---|---|---|---|
| Língua Portuguesa | 4ª | 27.176 | 1.975.635 |
| | 8ª | 22.035 | 1.422.245 |
| Matemática | 4ª | 26.907 | 1.975.635 |
| | 8ª | 22.089 | 1.422.245 |

Um total de 98.207 estudantes de escolas públicas urbanas com mais de 30 alunos foram avaliados pela ANEB. Por pesos amostrais, a expansão para a população representou 2.876.722 e 2.515.730 estudantes em língua portuguesa e 2.876.722 e 2.515.731 estudantes em matemática, 4ª e 8ª séries EF respectivamente.

Para a Prova Brasil, 3.397.880 estudantes foram avaliados, lembrando que um mesmo estudante responde a testes das duas disciplinas. Considerando os pesos amostrais para São Paulo, a expansão para a população da Prova Brasil representou um total de 3.721.631, sendo 2.111.558 para 4ª série EF e 1.610.073 para 8ª série EF.

Com relação ao tempo de aplicação, os estudantes que responderam aos testes da ANEB dispunham de 90 minutos para responder aos três blocos de 13 itens (INEP, 2005b). Para a Prova Brasil, os estudantes de 4ª série responderam aos quatro blocos de 10 itens em um máximo de 80 minutos e os de 8ª série aos quatro blocos de 12 itens em um máximo de 100 minutos (INEP, 2005a). A tabela 4.6 apresenta um detalhamento referente ao tempo de aplicação para ambas as avaliações.

Tabela 4.6 - Tempo de aplicação dos testes da ANEB 2005 e da Prova Brasil 2005.

| Bloco | ANEB | | Prova Brasil - 4ª série | | Prova Brasil - 8ª série | |
|---|---|---|---|---|---|---|
| | N itens | Tempo | N itens | Tempo | N itens | Tempo |
| Bloco1 | 13 | 30 | 10 | 20 | 12 | 25 |
| Bloco2 | 13 | 30 | 10 | 20 | 12 | 25 |
| Bloco3 | 13 | 30 | 10 | 20 | 12 | 25 |
| Bloco4 | - | - | 10 | 20 | 12 | 25 |
| N Total | 39 | 90 | 40 | 80 | 48 | 100 |
| Minutos/item | - | 2,31 | - | 2,00 | - | 2,08 |

Os testes de 4ª e 8ª séries EF da Prova Brasil (40 e 48 itens) foram maiores que os testes da ANEB (39 itens). A ANEB 2005 disponibilizou um maior tempo por item para os respondentes (2,31 minutos por item). O tempo disponibilizado para a Prova Brasil, no entanto, não se distanciou muito desse tempo médio (cerca de 2 minutos por item).

Ambas as avaliações utilizaram a TRI, sob o modelo de três parâmetros, para estimar a proficiência; utilizaram itens comuns com o SAEB 2003, para permitir a equalização e apresentação dos resultados na métrica do SAEB; envolveram disciplinas (língua portuguesa e matemática) e séries (4ª e 8ª séries EF) em comum; tiveram suas provas construídas sob o mesmo enfoque teórico, pautadas na Matriz de Referência do SAEB, compostas pelos mesmos tipos de itens (múltipla escolha, quatro e cinco alternativas); disponibilizaram tempo de resposta por item semelhante.

As duas avaliações utilizaram delineamento de montagem de testes e administração pela amostra por Blocos Incompletos Balanceados – BIB (Bekman, 2001; Johnson, 1992), mas apresentaram diferenças quanto à estrutura dos testes:

(a) A ANEB 2005 manteve a estrutura tradicional do SAEB: composição de 26 cadernos a partir da combinação de 13 blocos de 13 itens, três a três, para todas as séries e disciplinas. Cada caderno foi composto por 39 itens. Cada aluno respondeu a um caderno.

(b) A Prova Brasil 2005: composição de 21 cadernos. Cada aluno respondeu a um caderno composto por dois blocos de língua portuguesa e dois blocos de matemática. Dentro de cada disciplina, houve a rotação de 7 blocos de 10 itens para a 4ª série EF e 12 itens para a 8ª série EF. Cada aluno de 4ª série EF respondeu a 40 itens e cada aluno de 8ª série EF a 48 itens.

Como as habilidades das duas avaliações foram estimadas por meio da TRI, sob modelo de três parâmetros, que assume a propriedade de invariância do parâmetro de habilidade independentemente do teste utilizado, espera-se que os resultados obtidos tenham sido iguais para grupos equivalentes. Ou seja, mesmo que o delineamento dos testes tenha sido diferente para a ANEB e para a Prova Brasil, esse fator não deveria impactar em uma diferenciação nos resultados de habilidades dos estudantes, desde que (a) o modelo adotado se ajuste aos dados (Hambleton & Jones, 1993, p. 42); (b) os valores de todos os parâmetros dos itens utilizados para estimá-los estejam em uma métrica comum (Baker, 2001); e (c) os itens dos testes sejam unidimensionais (Baker, 2001).

Os estudos de Condé (2007) e Rabello (2007) compararam os resultados de habilidades dos estudantes de 4ª e 8ª séries EF de escolas públicas urbanas que responderam aos testes da ANEB 2005 e da Prova Brasil 2005. Observaram uma proximidade entre as médias de habilidades estimadas para as avaliações. Para algumas séries e disciplinas e para certos grupos de comparação, no entanto, um conjunto de médias da Prova Brasil se distanciou das calculadas para a ANEB.

Como pode ser verificado na tabela 4.7, em nível Brasil, as médias de língua portuguesa, 8ª série EF, e de matemática, 4ª série EF, da Prova Brasil 2005 não apresentaram diferenças significativas às médias da ANEB 2005, considerando intervalo de confiança de 95% calculado para a ANEB. Esses resultados apontam para o que seria esperado pela TRI: grupos equivalentes de estudantes que responderam a testes diferentes apresentaram estimativas de habilidade iguais (ou bastante semelhantes).

Tabela 4.7 - Desempenho dos estudantes na ANEB 2005 e na Prova Brasil 2005 – Brasil – língua portuguesa e matemática, 4ª e 8ª séries do EF - Escolas Públicas Urbanas com Federais.

| Disciplina | Série | ANEB 2005 | | | | | Prova Brasil 2005 | | Diferença | Distância do IC 95% |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **Média** | DP | EP | IC 95% (LI ) | IC 95% (LS ) | **Média** | DP | | |
| Língua Portuguesa | 4ª | **170,6** | 42,6 | 1,3 | 168,0 | 173,1 | **173,4** | 41,7 | 2,8 | 0,28 |
| | 8ª | **225,4** | 46,2 | 1,1 | 223,3 | 227,5 | **224,4** | 41,4 | -1,0 | - |
| Matemática | 4ª | **180,1** | 44,6 | 1,1 | 177,9 | 182,3 | **180,6** | 39,9 | 0,5 | - |
| | 8ª | **231,6** | 45,6 | 1,3 | 229,2 | 234,1 | **239,5** | 42,5 | 7,9 | 5,43 |

(Fonte: Condé, 2007; Rabello, 2007).

Já as médias de língua portuguesa 4ª série EF e de matemática 8ª série EF, da Prova Brasil 2005 extrapolaram o intervalo de confiança calculado para a ANEB 2005 (tabela 4.7). Se para língua portuguesa, 4ª série EF, a diferença não é tão expressiva, para matemática, 8ª série EF, observou-se um distanciamento do intervalo de confiança de 95% calculado para a ANEB superior a cinco pontos da escala do SAEB, ilustrado na figura 4.1.

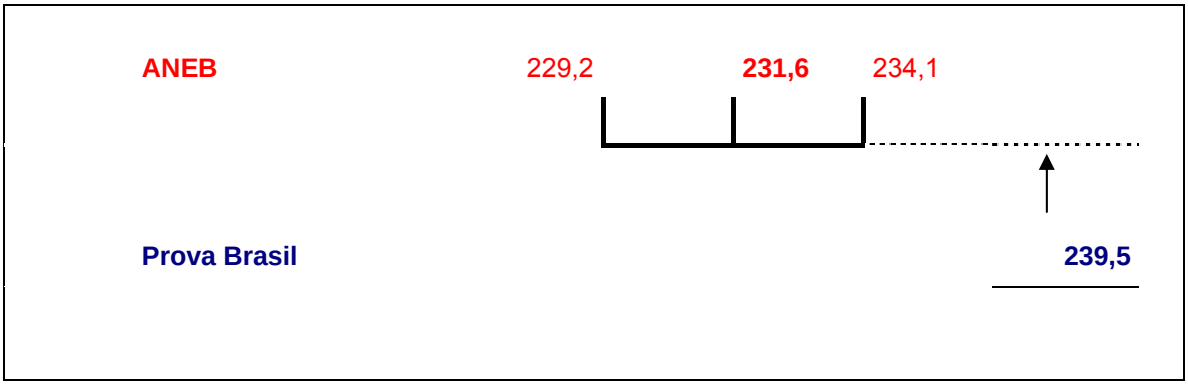


Figura 4.1 - Desempenho dos estudantes na ANEB 2005 e na Prova Brasil 2005 em matemática, 8ª série EF - Escolas Públicas Urbanas com Federais para o Brasil.

Quando as comparações entre as médias foram realizadas em nível de Regiões do Brasil, os resultados foram semelhantes. Para língua portuguesa 8ª série EF, todas as regiões apresentaram médias na Prova Brasil 2005 semelhantes às médias da ANEB 2005, já que apresentavam intersecção com o intervalo de confiança da ANEB (Condé, 2007, p.

7). Para as duas disciplinas em 4ª série EF, as diferenças entre as médias da Prova Brasil 2005 e da ANEB 2005 não foram, de modo geral, significativas, com exceção para a Região Nordeste que apresentou uma distância do limite superior a 4 pontos para as duas disciplinas.

Em matemática 8ª série EF, no entanto, para todas as Regiões foram observadas diferenças significativas entre os resultados da Prova Brasil e da ANEB, coerentemente aos encontrados para a mesma série e disciplina em nível Brasil (Condé, 2007; Rabello, 2007). Para essa série e disciplina, de modo geral, as médias da Prova Brasil 2005, tanto em nível Brasil, quanto para Regiões, foram superiores às médias da ANEB 2005. Uma observação metodológica cabe ser ressaltada: os autores não utilizaram pesos amostrais para expansão dos resultados da Rede Estadual de São Paulo. No entanto, as diferenças entre os resultados não são tão relevantes. Um exemplo é a média Brasil de matemática 8ª série EF que, sem peso, é de 239,50, enquanto que, com peso, de 239,98. Para essa série e disciplina, por exemplo, os resultados entre ANEB e Prova Brasil para estudantes de escolas públicas e urbanas foram significativamente diferentes.

Esses resultados suscitam algumas discussões. O delineamento de montagem, de distribuição dos testes, de composição da amostra e de análises de dados da ANEB 2005 foi realizado sob os mesmos moldes utilizados pelo SAEB 2003 (CESPE, 2007b). Quando há a introdução da Prova Brasil 2005, com a manutenção de uma série de variáveis estruturais da ANEB 2005, mas com a modificação de outras principalmente com relação à estrutura do teste, observou-se entre as avaliações: (a) resultados semelhantes nas estimativas médias de habilidades dos estudantes para as duas disciplinas em 4ª série EF e para língua portuguesa 8ª série EF; (b) resultados diferentes de estimativas de habilidade média dos estudantes em matemática 8ª série EF.

Com relação aos resultados de matemática 8ª série EF da ANEB 2005 e Prova Brasil 2005, questiona-se: que variáveis estão gerando essa diferença de habilidades dos estudantes? Por que os resultados da Prova Brasil foram significativamente superiores aos da ANEB? Se, na Prova Brasil, os testes de matemática 8ª série EF foram aplicados juntamente com os de língua portuguesa 8ª série EF, por que para os de Língua Portuguesa não foram observadas diferenças significativas entre as avaliações?

Atribuir a procedimentos de administração de testes diferenciados da Prova Brasil 2005 pode não justificar satisfatoriamente essa diferença visto que, para 8ª série EF, os mesmos procedimentos de aplicação foram implementados para língua portuguesa e para

matemática (as duas disciplinas compunham cadernos únicos) e apenas os resultados de matemática foram significativamente destoantes entre as avaliações ANEB e Prova Brasil.

A diferença entre os resultados de matemática 8ª série EF da Prova Brasil 2005 e da ANEB 2005, tendo em vista a revisão bibliográfica realizada para o presente estudo, pode ter sido gerada em função do:

a) O alinhamento diferenciado do teste à Matriz de Referência do SAEB. É possível que a cobertura maior ou menor a determinados temas ou descritores tenham influenciado na validade dos resultados.
b) A qualidade e a configuração psicométrica diferente entre os testes.
c) O distanciamento do pressuposto da unidimensionalidade em um dos dois testes influenciou na implementação da TRI.
d) O tamanho dos EPM entre os testes foram diferentes.
e) O número de itens da Prova Brasil 2005 que cada aluno respondeu (24 itens), inferior ao que um aluno respondeu na ANEB 2005 (39 itens), gerou uma erro maior nas estimativas individuais da habilidade para a primeira.
f) A diferença do número de itens com baixo poder discriminativo.

Tendo em vista os aspectos abordados, o presente estudo tem como objetivo verificar a associação entre as características dos testes na validade e na precisão das estimativas de habilidade por meio da TRI. Especificamente, pretende identificar quais fatores relacionados ao teste estão associados às diferenças observadas de resultados de matemática 8ª série EF entre a ANEB 2005 e a Prova Brasil 2005.

## 5. Método

A presente seção apresenta a metodologia utilizada para verificar a relação entre as características do teste e a validade e a precisão das estimativas de habilidade da TRI. Foram realizados quatro estudos com os testes de matemática 8ª série EF da ANEB 2005 e da Prova Brasil 2005.

No Estudo 1, as análises de Rabello (2007) e de Condé (2007) foram replicadas e aprofundadas no que tange à comparação dos resultados entre ANEB e Prova Brasil por faixa de habilidades. O Estudo 2 contemplou a análise das características dos testes quanto à cobertura da matriz, aos seus aspectos psicométricos (TCT e TRI) e à dimensionalidade dos testes. O Estudo 3 estimou as habilidades dos estudantes da ANEB sob novas

configurações de teste para viabilizar comparações com os resultados da Prova Brasil e da própria ANEB. No Estudo 4, a distância entre os resultados da Prova Brasil, da ANEB e de quatro testes simulados (formas A, B, C e D) foi verificada.

## 5.1 Estudo 1: Comparação das estimativas de habilidade dos estudantes da ANEB e da Prova Brasil

Condé (2007) e Rabello (2007) compararam as estimativas de habilidade dos estudantes entre a ANEB e a Prova Brasil, considerando para a primeira avaliação apenas os resultados dos estudantes de escolas públicas (com escolas Federais) e urbanas, já que o universo da segunda envolvia apenas estes níveis. Os autores utilizaram os valores médios pontuais e verificaram se as médias das estimativas de habilidade na Prova Brasil encontravam-se dentro dos intervalos de confiança de 95% calculados para a ANEB. Realizaram a análise para língua portuguesa e matemática, 4ª e 8ª séries EF, para os estratos Brasil, Regiões e Unidades da Federação. Os resultados médios de habilidades dos estudantes na Prova Brasil foram superiores aos da ANEB para matemática 8ª série.

Orientado pelos resultados dessa série e disciplina, o Estudo 1 teve como objetivos replicar as análises realizadas por Rabello (2007) e por Condé (2007) especificamente para matemática 8ª série EF e comparar os percentuais de estudantes por faixa de habilidades da escala.

A base de dados da ANEB foi constituída por estimativas de habilidade de 22.089 estudantes. Considerando-se os pesos amostrais, os resultados foram expandidos para 2.515.731 estudantes. A base da Prova Brasil foi composta por 1.422.245 estudantes. Para a Prova Brasil, com exceção da Rede Estadual de São Paulo, não houve ponderação (CESGRANRIO, 2006). Diferentemente dos estudos de Condé (2007) e de Rabello (2007), para as respostas referentes às escolas estaduais de São Paulo, foram utilizados pesos amostrais. Com a expansão para a Rede Estadual de São Paulo, o número de estudantes envolvidos na Prova Brasil foi de 1.610.073.

As médias e os desvios-padrão das habilidades dos estudantes das bases das avaliações foram calculados separadamente. Utilizando os intervalos de confiança de 95% calculados para a ANEB 2005, foi verificado em que medida as médias da Prova Brasil se aproximaram desse intervalo. Diferentemente dos estudos de Condé (2007) e de Rabello (2007), os resultados foram calculados em escala normalizada para Brasil, Regiões e Unidades da Federação, tendo por sua origem média 0 e desvio-padrão 1,0, referente à media dos estudantes de 8ª série do SAEB 97 nessa mesma disciplina.

Como para as escolas estaduais e municipais do Rio de Janeiro, a Prova Brasil 2005 utilizou dois blocos de itens diferentes das demais Unidades da Federação e para o presente estudo é importante que as provas que foram aplicadas fossem comuns a todos os estudantes, verificou-se o impacto da retirada dos resultados desse estado para o cálculo das médias do Brasil e do Sudeste. Caso os resultados tivessem impacto significativo, seria fundamental que os resultados do Rio de Janeiro fossem retirados da base de dados para realização das próximas análises.

Como complementação à análise por médias, comparações entre os percentuais de estudantes localizados em cada uma das faixas de habilidades de matemática 8ª série EF foram realizadas. Verificou-se se, para grupos diferentes de estimativas de habilidade, os percentuais de estudantes variaram entre as avaliações. Esperavam-se percentuais semelhantes entre elas já que os grupos de estudantes que responderam à ANEB e à Prova Brasil apresentaram características semelhantes. Foram realizadas análises por gráficos de barras e pela comparação das áreas de distâncias entre as distribuições. Esses primeiros resultados subsidiaram argumentos sobre a existência de diferenças entre as distribuições de estudantes entre as avaliações pelos diversos níveis de habilidades da escala.

## 5.2 Estudo 2: Características dos testes ANEB e Prova Brasil

O Estudo 2 teve como objetivo analisar as características dos testes de 8ª série matemática da ANEB e da Prova Brasil no que se refere: (a) à cobertura da matriz de referência; (b) às suas características psicométricas; (c) ao grau de cumprimento do pressuposto da unidimensionalidade.

### 5.2.1 Abrangência da cobertura da matriz de referência

Como é a cobertura da matriz de referência do SAEB, com relação aos testes de 169 itens da ANEB e de 84 itens da Prova Brasil? Como se apresenta o alinhamento (Bhola, Impara & Buckendahl, 2003; Herman, Webb & Zuniga, 2002) do teste com a matriz ou o grau de validade com referência ao conteúdo dos resultados obtidos por sua aplicação? O Estudo 2 teve como um de seus objetivos caracterizar os testes em termos da cobertura da matriz. O número de itens por tema e por descritor para os testes foi calculado e seus resultados comparados entre as avaliações. Possíveis diferenças identificadas indicariam um alinhamento diferenciado entre os testes.

### 5.2.2 Características psicométricas dos testes

A configuração dos parâmetros psicométricos dos testes de matemática 8ª série EF da ANEB e da Prova Brasil, previamente calculados pelo INEP, será estudada. Esses parâmetros foram utilizados como base para a estimação das habilidades analisadas na seção 5.1.

Estatísticas descritivas referentes aos parâmetros *a*, *b* e *c* dos itens foram calculadas para a ANEB e a Prova Brasil, tendo por base os testes como um todo e os blocos. Estatísticas descritivas dos parâmetros *a* e *b* foram calculadas por caderno para ambos os testes. Todos os resultados foram analisados de forma a caracterizá-los.

### 5.2.3 Dimensionalidade dos testes

Procedeu-se à análise dos relatórios técnicos e dos estudos realizados pelo INEP para verificação do pressuposto de unidimensionalidade dos testes de matemática 8ª série da ANEB 2005 e da Prova Brasil 2005 (CESGRANRIO, 2006; CESPE, 2007c). CESPE (2007c) realizou um estudo para verificação da dimensionalidade dos testes tendo por base o método de Análise Fatorial de Informação Plena (*Full-Information Factor Analysis – FIFA)* (Bock, Gibbons & Muraki, 1988; Laros, Pasquali & Rodrigues, 2000), baseada na TRI (Hattie, 1985). O software Testfact 3 (Wilson, Wood & Gibbons, 1991; Wood et al., 2003) foi utilizado. Ressalta-se que o método utiliza padrões distintos de resposta ao item em vez de intercorrelações, utilizando o modelo multifatorial de Thurstone baseado em estimativas de máxima verossimilhança marginal e no algoritmo EM (*expectation – maximization*) (CESPE, 2007b; Wilson, Wood & Gibbons, 1991; Pasquali, 2003). Essa análise é indicada quando a matriz de correlações é do tipo tetracórica, como é o caso dos itens dicotômicos do SAEB. Não foram encontrados estudos de verificação da dimensionalidade para a Prova Brasil.

## 5.3 Estudo 3: Estimação das habilidades dos estudantes da ANEB sob novas configurações de teste

### 5.3.1 Estimação das habilidades de acordo com os critérios utilizados pelo INEP

Considerando os mesmos critérios utilizados para estimar as habilidades dos estudantes da ANEB 2005, rodou-se a Fase 3 do BILOG-MG, versão 1. O objetivo de realização da análise foi o de verificar o alcance dos mesmos resultados obtidos pelo INEP (CESPE, 2007b).

Respostas de 206.453 estudantes de 4ª, 8ª séries do EF e 3ª série do EM, a 792 itens de matemática foram utilizadas para estimar as habilidades. Consideraram-se o modelo logístico de três parâmetros (NPARM=3) sob a métrica de função de resposta normal (NORMAL). Já que a base de dados foi estruturada a partir da utilização de itens de quatro e de cinco alternativas, registrou-se cinco como número máximo de alternativas (NALT=5). O número de formas utilizado foi de 156 (NFORM=156) para seis grupos (NGROUP=6, referentes à 4ª e 8ª EF e 3ª EM para 2003 e 2005). Cada forma de teste era composto por 39 itens, número respondido por cada estudante (LENGHT=39). Para o Grupo 5, ou seja, 8ª série EF da ANEB 2005, 155 itens foram considerados (LENGHT=155).

Os parâmetros a, b e c na escala do SAEB foram inseridos no programa de sorte que quando o item era comum entre séries ou entre anos, o parâmetro era apresentado uma única vez, na série ou ano original. Maiores detalhes sobre os procedimentos de estimação podem ser encontrados em CESPE (2007b).

### 5.3.2 Estimação das habilidades a partir da desvinculação dos itens entre séries para o ano de 2005

Para posteriores manipulações da base da ANEB a partir da retirada de itens, percebeu-se a necessidade de desvincular os itens entre séries (itens de 4ª inseridos no teste de 8ª; itens de 8ª EF inseridos na 3ª EM) para a base referente ao ano de 2005. Isso porque, quando houvesse necessidade de excluir itens de 4ª série EF contidos na 8ª série EF, também seriam retirados da 4ª, já que era sua referência original. Além disso, qualquer exclusão de itens da 8ª série EF comuns à 3ª série EM, esses também seriam retirados da 3ª série EM. Assim, para fornecer mais liberdade à manipulação do teste de matemática, 8ª série, na base de 2005, quando um item era comum entre séries, recebia um nome diferente para cada série (INAMES). A partir desse procedimento, foi necessário repetir na série posterior os parâmetros da série original (em THRESHLD, SLOPE e GUESS).

Após o procedimento, 870 itens foram considerados e não mais 792. Cabe ressaltar que como os parâmetros a, b e c foram repetidos entre as séries de 2005, esperava-se poucas alterações nos resultados de estimativa de habilidades, já que a equalização entre as séries faz referência à proximidade dos parâmetros independentemente da mudança de nomes. No caso os parâmetros dos itens entre série para aqueles que deixaram de ser nominalmente comuns foram iguais. Cabe a realização de uma nova estimação das

habilidades, sendo que os resultados para 2005 não podem se afastar dos resultados originais da ANEB.

O procedimento foi providencial, dado que as próximas análises buscariam delineamentos simulados de testes de forma a aproximar às características da Prova Brasil. A equalização realizada para a Prova Brasil 2005, considera itens comuns com o SAEB 2003, mas não entre séries de 2005, de forma semelhante ao apresentado no presente tópico.

### 5.3.3 Teste A: estimação das habilidades a partir de 104 itens com parâmetros similares aos da ANEB

Após a verificação do alcance dos mesmos resultados obtidos pelo INEP, após os procedimentos empregados nos tópicos 5.3.1 e 5.3.2 do presente método, parte-se para algumas manipulações das bases de dados referentes ao teste de 8ª série EF da ANEB 2005, de forma a manter as médias e os desvios-padrão dos parâmetros *a* e *b* similares aos da ANEB.

Algumas considerações devem ser feitas. Antes da concepção do Teste A, planejou-se compor um teste similar à Prova Brasil 8ª série EF a partir da seleção de itens do teste de matemática 8ª série EF da ANEB 2005. O teste chegou a ser composto. Para um dos 81 itens válidos da Prova Brasil, buscou-se um correlato em termos de descritor ou tema da matriz, parâmetro *a* e parâmetro *b*. Houve a preocupação de manter a ordem dentro do bloco original da ANEB para não impactar nas estimativas em função do efeito posição. Esse procedimento se mostrou inadequado, já que a estrutura da base de dados composta por estudantes como casos e 39 itens como variáveis. Como existem 26 cadernos diferentes de testes, a exclusão 74 itens (155 da ANEB para alcançar 81 itens como a Prova Brasil) impactaria desigualmente nos 26 cadernos. Grupos de alunos responderiam a vários itens contidos no teste simulado, enquanto outros praticamente ficariam sem itens para estimar suas habilidades. Além disso, uma série de mudanças deveria ser realizada na base de dados o que tornaria praticamente inviável para efeitos do presente trabalhos.

Para atingir os objetivos do presente trabalho, decidiu-se que as manipulações fossem realizadas sem alteração das formas de teste (FORM) e com alteração dos itens do teste como um todo (GROUP=5). Dessa forma, a busca de um paralelismo entre as provas não seria realizada, já que não seria possível mudar a ordem dos itens. No entanto, a partir da manipulação dos itens relacionados no Grupo 5, testes com delineamentos diversos puderam ser constituídos.

O primeiro modelo de teste, denominado de Teste A, tema do presente tópico do trabalho, foi composto por 104 itens. Assim, houve uma redução de 51 dos originais 155 itens da ANEB. Como cada caderno da ANEB é comporto por três blocos de 13 itens, foram excluídos exatamente cinco itens de cada bloco. Assim, exatamente 24 itens por forma de teste (FORM) foram considerados no Teste A, número semelhante de itens contidos em cada caderno de matemática 8ª série EF da Prova Brasil. O Teste A se aproximou da Prova Brasil, em comparação à ANEB, em termos de número de itens total e de número de itens por caderno.

Para o Teste A, buscou-se manter os parâmetros *a* e *b* médios constantes, em comparação à ANEB, exclusivamente para verificar o impacto do tamanho do teste nas estimativas de habilidade. Essas foram estimadas e seus resultados analisados posteriormente à luz dos resultados da Prova Brasil e da ANEB.

### 5.3.4 Teste B: estimação das habilidades a partir de 104 itens e da otimização da discriminação da ANEB

Os resultados médios referentes ao parâmetro *a* da Prova Brasil foram muito superiores aos resultados da ANEB (1,87 e 1,24 respectivamente). Questionou-se até que ponto a discriminação do teste poderia influenciar a diferença entre os resultados das avaliações associado a uma aproximação do tamanho dos testes, levando-se em consideração que o parâmetro *a* tem relação direta com a qualidade de testes.

O Teste B foi construído de forma a associar a redução do número de itens (de 155 para 104) a uma elevação da média do parâmetro *a* para 1,46. Esse valor foi o máximo alcançado para 104 itens, já que foram selecionados para compor o Teste B os itens mais discriminativos disponíveis.

Como houve controle exclusivamente do parâmetro *a*, na composição do Teste B, como efeito não esperado houve um aumento da dificuldade média em comparação à ANEB (0,71 da ANEB; 0,79 da Prova Brasil; 0,97 do Teste B). As habilidades foram estimadas e os resultados analisados posteriormente à luz dos resultados da Prova Brasil e da ANEB.

### 5.3.5 Teste C: estimação das habilidades a partir de 104 itens, da otimização da discriminação e do controle da dificuldade da ANEB

O Teste C foi composto de forma a buscar a melhor discriminação para um teste com 104 itens com dificuldade semelhante à da ANEB e que pelo menos não ultrapassasse

a dificuldade da Prova Brasil Assim, foi composto um teste simulado com parâmetro *a* médio de 1,40 e parâmetro *b* médio de 0,73 (lembrando que a ANEB apresentava *a* de 1,24 e *b* de 0,71; a Prova Brasil apresentou *a* de 1,87 e *b* de 0,79).

O Teste C foi composto buscando-se os itens com melhor discriminação, mas considerando a distribuição de dificuldade da Prova Brasil. Se analisarmos a Prova Brasil, verificaremos que em todos os blocos há itens dos mais diversos níveis de dificuldade: com parâmetro *b* inferior a -1, entre -1 e +1 e superior a +1. Procurou-se uma distribuição aproximada. Os resultados foram razoavelmente satisfatórios.

Da mesma forma que para os outros testes, as estimativas das habilidades foram calculadas e os resultados comparados com a ANEB e a Prova Brasil.

### 5.3.6 Teste D: estimação das habilidades a partir de 81 itens e da otimização da discriminação da ANEB

O Teste D foi composto pela exclusão de 74 itens da ANEB , de forma à obtenção de um teste composto por 81 itens, mesmo número total de itens da Prova Brasil. Para os blocos 1 a 11, foram excluídos os sete itens menos discriminativos. Para o bloco 12, foram excluídos cinco itens com os menores parâmetros *a* e, para o Bloco 13, excluíram-se os seis itens menos discriminativos. Sob esse delineamento, os estudantes responderam testes que variaram de 18 a 21 itens. Assim, se para os Testes A, B e C, o número de itens respondidos por aluno foi de 24 itens, de forma semelhante à Prova Brasil, para o Teste D, utilizou-se um número menor de itens por caderno para garantir o número de 81 itens total. Como o critério de exclusão de itens foi exclusivamente o parâmetro *a*, não houve controle do parâmetro *b*.

O Teste D apresentou 15 itens comuns com o SAEB 2003. Da mesma forma que para os outros testes, as estimativas das habilidades foram calculadas e os resultados comparados com a ANEB e a Prova Brasil.

## 5.4 Estudo 4: Comparação entre as estimativas de habilidade dos estudantes para Prova Brasil, ANEB e Testes A a D e sua associação com as características dos testes

O estudo 4 contemplou a comparação dos resultados da Prova Brasil e da ANEB com os resultados dos Testes A a D. As estatísticas de habilidades foram associadas aos parâmetros dos itens e ao número de itens dos testes, já que é objetivo do presente estudo

foi o de verificar a associação entre as características dos testes e os parâmetros de habilidade.

Realizaram-se as seguintes análises (a) as médias e os desvios-padrão do parâmetro de habilidade estimado para cada teste foram associados às estimativas dos parâmetros *a* e *b;* (b) associação entre número de itens e parâmetro *a* médio dos testes com o parâmetro de habilidade; (c) Os percentuais de estudantes por faixa de habilidade estimada de 1 DP foram associados aos percentuais de itens e ao parâmetro *a* médio para essas faixas; (d) O EPM médios dos testes como um todo e por faixa de habilidade foram calculados; (e) Valores pontuais da informações dos itens representantes de cada uma das faixas de habilidades de 1 DP foram calculados e associados aos percentuais de estudantes localizados em cada uma das faixas. Os resultados de informação dos itens representativos das faixas de parâmetro de habilidade foram calculados pelo inverso do quadrado do EPM. Na seguinte seção, os resultados dos quatro estudos propostos são apresentados.

## 6. Resultados

### 6.1 Estudo 1: Comparação das estimativas de habilidade dos estudantes da ANEB e da Prova Brasil

Os resultados médios das estimativas de habilidade dos estudantes de 8ª série EF em Matemática entre a ANEB 2005 e a Prova Brasil 2005, considerando para a primeira avaliação apenas os resultados dos estudantes de escolas públicas (com escolas federais) e urbanas, foram calculados. Os resultados são apresentados na tabela 6.1 em escala normalizada (-3 a +3) para Brasil, Regiões e Unidades da Federação, tendo por sua origem média 0 e DP 1, referente à media dos estudantes de 8ª série do SAEB 97 nessa mesma disciplina. Para a ANEB 2005, foram considerados os pesos amostrais. Para a Prova Brasil foram considerados pesos amostrais especificamente para a Rede Estadual de São Paulo (CESGRANRIO, 2006, p. 2).

Tabela 6.1 - Comparação das médias de estimativas de habilidade dos estudantes em matemática, 8ª série EF, ANEB e Prova Brasil - Brasil, Regiões e UFs.

| UF | ANEB 2005 | | | | Prova Brasil 2005 | Diferença | Sig. |
|---|---|---|---|---|---|---|---|
| | Média | EP | IC | | Média | | |
| | | | LI 95% | LS 95% | | | |
| Brasil | -0,33 | -4,45 | -0,37 | -0,28 | -0,18 | 0,15 | * |
| Norte | -0,50 | -4,45 | -0,54 | -0,45 | -0,31 | 0,19 | * |
| Rondônia | -0,32 | -4,39 | -0,47 | -0,17 | -0,15 | 0,17 | * |
| Acre | -0,48 | -4,43 | -0,56 | -0,40 | -0,34 | 0,14 | * |
| Amazonas | -0,63 | -4,42 | -0,73 | -0,53 | -0,37 | 0,26 | * |
| Roraima | -0,55 | -4,39 | -0,71 | -0,38 | -0,23 | 0,32 | * |
| Pará | -0,42 | -4,43 | -0,51 | -0,34 | -0,31 | 0,11 | * |
| Amapá | -0,48 | -4,43 | -0,57 | -0,39 | -0,36 | 0,12 | * |
| Tocantins | -0,56 | -4,42 | -0,67 | -0,45 | -0,30 | 0,26 | * |
| Nordeste | -0,59 | -4,45 | -0,63 | -0,55 | -0,44 | 0,15 | * |
| Maranhão | -0,66 | -4,44 | -0,72 | -0,59 | -0,43 | 0,23 | * |
| Piauí | -0,57 | -4,40 | -0,71 | -0,43 | -0,33 | 0,24 | * |
| Ceará | -0,60 | -4,42 | -0,70 | -0,50 | -0,44 | 0,16 | * |
| Rio Grande do Norte | -0,57 | -4,43 | -0,66 | -0,49 | -0,40 | 0,17 | * |
| Paraíba | -0,58 | -4,42 | -0,68 | -0,48 | -0,48 | 0,10 | |
| Pernambuco | -0,63 | -4,43 | -0,71 | -0,55 | -0,49 | 0,14 | * |
| Alagoas | -0,67 | -4,40 | -0,80 | -0,53 | -0,50 | 0,17 | * |
| Sergipe | -0,34 | -4,41 | -0,46 | -0,21 | -0,33 | 0,01 | |
| Bahia | -0,54 | -4,41 | -0,66 | -0,41 | -0,42 | 0,12 | |
| Sudeste | -0,23 | -4,43 | -0,32 | -0,14 | -0,09 | 0,14 | * |
| Minas Gerais | -0,02 | -4,38 | -0,20 | 0,16 | -0,03 | -0,01 | |
| Espírito Santo | -0,17 | -4,41 | -0,28 | -0,05 | -0,07 | 0,10 | |
| Rio de Janeiro | -0,31 | -4,43 | -0,40 | -0,21 | -0,12 | 0,19 | * |
| São Paulo | -0,33 | -4,41 | -0,46 | -0,20 | -0,12 | 0,21 | * |
| Sul | -0,07 | -4,43 | -0,16 | 0,01 | 0,04 | 0,11 | * |
| Paraná | -0,20 | -4,37 | -0,40 | -0,01 | 0,04 | 0,24 | * |
| Santa Catarina | -0,04 | -4,43 | -0,13 | 0,05 | 0,00 | 0,04 | |
| Rio Grande do Sul | 0,05 | -4,43 | -0,03 | 0,13 | 0,06 | 0,01 | |
| Centro-Oeste | -0,31 | -4,44 | -0,37 | -0,25 | -0,15 | 0,16 | * |
| Mato Grosso do Sul | -0,19 | -4,43 | -0,28 | -0,09 | -0,02 | 0,17 | * |
| Mato Grosso | -0,40 | -4,42 | -0,51 | -0,29 | -0,22 | 0,18 | * |
| Goiás | -0,39 | -4,43 | -0,47 | -0,32 | -0,21 | 0,18 | * |
| Distrito Federal | -0,03 | -4,36 | -0,25 | 0,20 | 0,03 | 0,06 | |

Os resultados indicaram diferença significativa entre as médias da ANEB e da Prova Brasil para matemática 8ª série EF em nível Brasil, Regiões e para a maioria das Unidades das Federações como indicavam os estudos. Os asteriscos apresentados na tabela indicam que a média calculada para a Prova Brasil extrapolou o limite superior do intervalo de confiança de 95% calculado para a ANEB. Nesses casos, a média da Prova

Brasil foi superior significativamente à da ANEB para grupos com características semelhantes.

Ressalta-se que os estudos de Condé (2007) e de Rabello (2007) não consideraram a ponderação para a Rede Estadual de São Paulo para a Prova Brasil. No entanto o impacto não foi muito significativo. A média Brasil subiu de -0,19 para -0,18, a da Região Sudeste desceu de -0,08 para -0,09 e a de São Paulo se manteve em -0,12.

Como para as escolas estaduais e municipais do Rio de Janeiro, a Prova Brasil 2005 utilizou dois blocos de itens diferentes das demais Unidades da Federação e, para o presente estudo, é importante que as provas que foram aplicadas fossem comuns a todos os estudantes, verificou-se o impacto da retirada dos resultados desse estado para o cálculo das médias do Brasil e do Sudeste. Os resultados não foram alterados, a média para o estrato Brasil manteve-se em -0,18 e para o Sudeste em -0,09. Para efeito de facilitar procedimentos para as próximas etapas do estudo, decidiu-se por considerar os resultados do Rio de Janeiro. Neste trabalho foram utilizadas exclusivamente as estatísticas da ANEB e da Prova Brasil calculadas em nível Brasil. Os resultados são apresentados na tabela 6.2.

Tabela 6.2 - Estatística de estimativas de habilidade dos estudantes em matemática, 8ª série EF, ANEB e Prova Brasil - Brasil.

| Teste | Habilidade | | | | |
|---|---|---|---|---|---|
| | N | Média | DP | Mínimo | Máximo |
| Prova Brasil | 1.610.073 | -0,1786 | 0,7617 | -2,11 | 2,79 |
| ANEB | 2.515.731 | -0,3283 | 0,8157 | -2,79 | 2,98 |

O número de estudantes apresentado na tabela 6.2 refere-se à população. A diferença entre as médias de habilidades estimadas foi de 0,15 pontos. A variabilidade da ANEB foi um pouco superior a da Prova Brasil (0,82 DP e 0,76 DP, respectivamente).

Calculando os percentuais de estudantes por faixa de estimativas de habilidade para ANEB e Prova Brasil de um DP (escala de -3 a +3) em nível nacional, observaram-se diferenças para praticamente todas as faixas. Os resultados estão representados na figura 6.1.

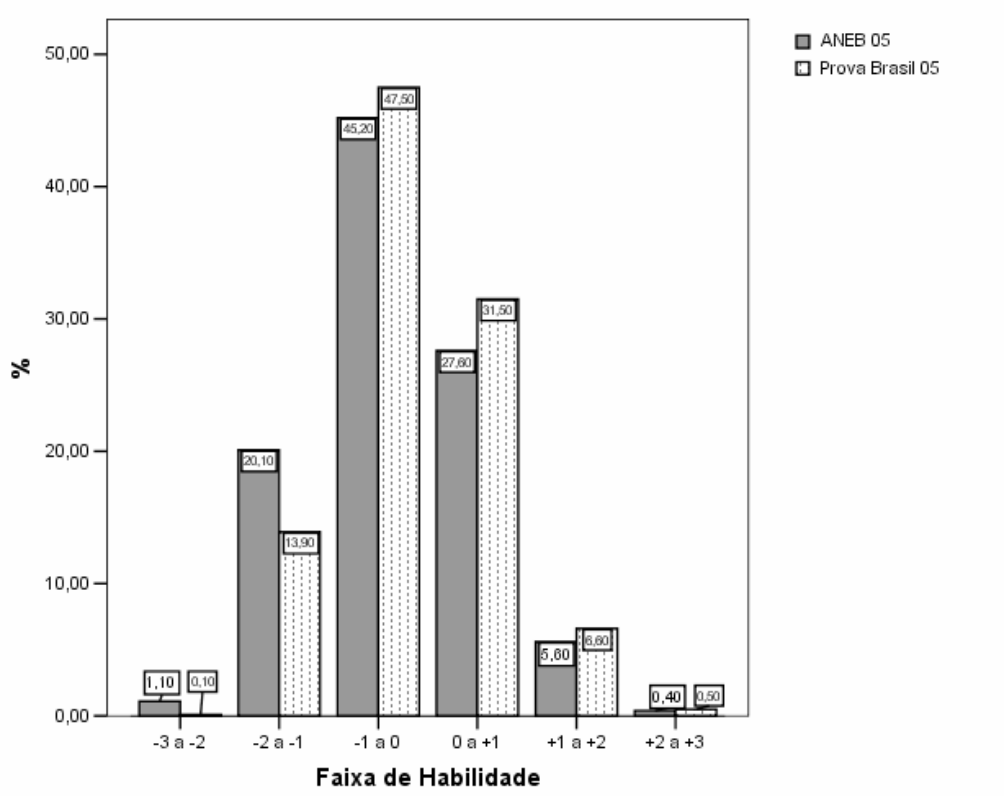


Figura 6.1 - Percentual de estudantes por faixa de estimativa de habilidades em matemática, 8ª série EF, ANEB e Prova Brasil - Brasil.

A distribuição das estimativas de habilidade para a Prova Brasil está deslocada para a direita do gráfico em comparação à da ANEB. Assim, nas faixas -1 a 0, 0 a +1 e +1 a +2, estimativas de um número maior estudantes obtidas pela Prova Brasil que pela ANEB são observadas. Nas faixas -3 a -2 e -2 a -1, um percentual maior de estimativas de habilidade obtidas pela ANEB foi encontrada. Para tornar mais evidente a diferença entre as distribuições, a figura 6.2 apresenta a distribuição de estudantes por faixa de habilidade.

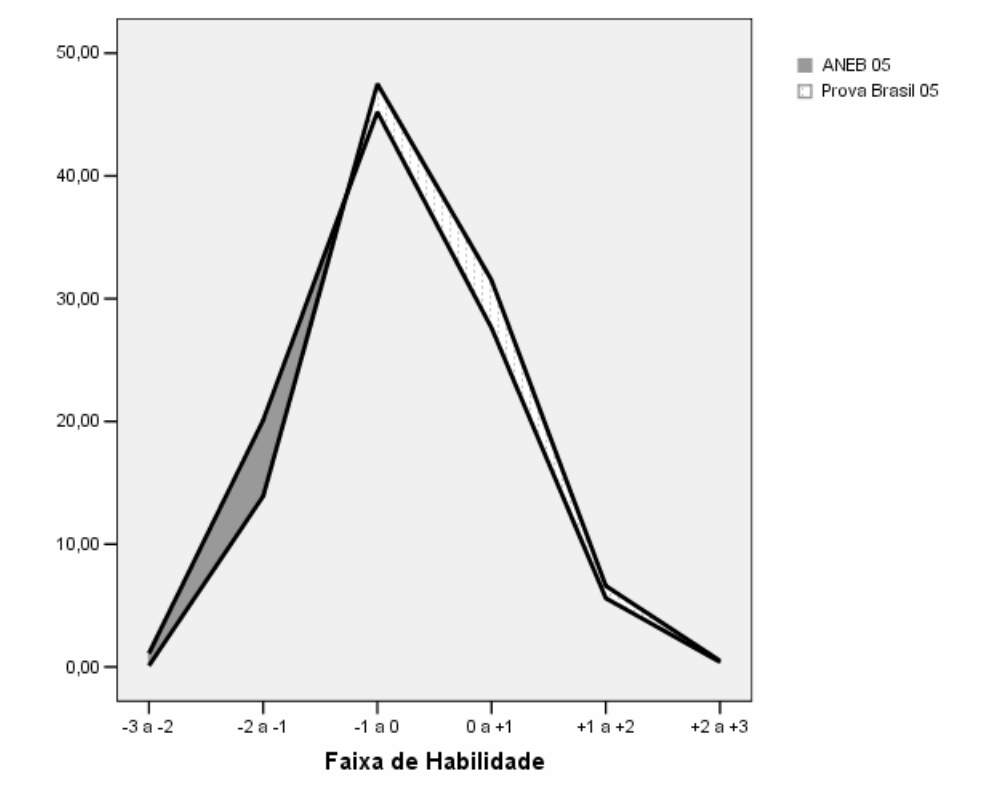


Figura 6.2 - Distâncias entre percentuais de estudantes por faixa de estimativas de habilidade em matemática, 8ª série EF, ANEB e Prova Brasil - Brasil.

A área escura refere-se às faixas de habilidade em que os resultados da ANEB são superiores aos da Prova Brasil. A área clara indica as faixas em que os resultados da Prova Brasil são superiores aos da ANEB. Os resultados indicam que os dois grupos de estudantes de 8ª série EF de escolas públicas urbanas estão obtendo estimativas de habilidade que apresentam distribuições diferentes. Questiona-se, no presente trabalho, até que ponto o teste é responsável por essas diferenças que, à princípio, não deveriam ocorrer caso a propriedade de invariância dos parâmetros da TRI fosse verificada. Os próximos estudos buscarão respostas a esse questionamento.

## 6.2 Estudo 2: Características dos testes ANEB e Prova Brasil

O estudo 2 contempla a análise da cobertura da matriz de referência e a análise psicométrica dos testes da ANEB e da Prova Brasil. Para a ANEB, foram excluídos os itens que não apresentaram parâmetros satisfatórios após a realização das análises TCT, TRI e por apresentarem Função Diferencial (CESPE, 2007a, 2007b). A lista dos itens excluídos é apresentada na tabela 6.3. As habilidades dos estudantes de 8ª série EF da ANEB foram estimadas tendo por base um total de 155 itens.

Tabela 6.3 - Itens excluídos das análises do teste de matemática 8ª série EF da ANEB 2005.

| Excluídos | Bloco | Posição | Descritor |
|---|---|---|---|
| TCT | 3 | 8 | D10 |
| | 4 | 4 | D03 |
| | 4 | 9 | D21 |
| | 4 | 13 | D01 |
| | 5 | 5 | D04 |
| | 5 | 6 | D07 |
| | 5 | 12 | D32 |
| | 6 | 6 | D08 |
| | 7 | 6 | D01 |
| TRI e DIF | 10 | 4 | * |
| | 10 | 5 | * |
| | 10 | 8 | * |
| | 11 | 3 | * |
| | 11 | 6 | * |

* Descritores de 4ª série EF.

Para a Prova Brasil foram excluídos três itens por não apresentarem parâmetros da TCT satisfatórios (CESGRANRIO, 2006), os quais não foram considerados para a análise TRI. A lista dos itens excluídos é apresentada na tabela 6.4. As habilidades dos estudantes de 8ª série EF da Prova Brasil foram estimadas tendo por base um total de 81 itens.

Tabela 6.4 - Itens excluídos das análises do teste de matemática 8ª série EF da Prova Brasil 2005.

| Excluídos | Bloco | Posição | Descritor |
|---|---|---|---|
| TCT | 3 | 8 | D14 |
| | 3 | 9 | D27 |
| | 3 | 12 | D32 |

### 6.2.1 Abrangência da cobertura da matriz de referência

Para conhecimento do alinhamento e da cobertura dos testes com relação à matriz referência, o número e o percentual de itens por tema e por descritor da matriz de

matemática 8ª série EF foram calculados para a ANEB e para a Prova Brasil. Os resultados por tema estão apresentados na tabela 6.5.

Para permitir a comparação da cobertura da matriz de 8ª série EF, os itens de 4ª série EF incluídos no teste de matemática 8ª série EF da ANEB foram desconsiderados. Para a ANEB, foram considerados 121 itens: (a) os itens de 8ª série EF; (b) os itens que não foram excluídos das análises. Para a Prova Brasil, foram considerados 81 itens que não foram excluídos das análises.

Tabela 6.5 - Número e percentual de itens por tema dos testes de matemática, 8ª série EF, ANEB e Prova Brasil.

| Tema | Prioridade | ANEB | | Prova Brasil | |
|---|---|---|---|---|---|
| | | n itens | % | n itens | % |
| I - Espaço e Forma | 2 | 33 | 27,3 | 26 | 32,1 |
| II - Grandezas e Medidas | 3 | 10 | 8,3 | 8 | 9,9 |
| III - Números e Operações/ Álgebra e Funções | 1 | 69 | 57,0 | 43 | 53,1 |
| IV - Tratamento da Informação | 2 | 9 | 7,4 | 4 | 4,9 |
| Total | - | 121 | 100 | 81 | 100 |

De modo geral, no que tange à cobertura da matriz de referência, os testes da ANEB e da Prova Brasil apresentam características semelhantes. Observa-se que os percentuais de itens por tema são bastante próximos entre os testes.

O estudo de prioridades apresentado no *framework* que contém as matrizes de referência do SAEB (INEP, 2002) prevê que a prova deve contemplar mais itens dos temas considerados pedagogicamente mais relevantes para a série. Assim, quanto mais a prioridade é próxima de 1, um número maior de itens deveria contemplar o tema. Como pode ser observado na tabela 6.5, com exceção do tema IV - Tratamento da Informação, um número maior de itens contemplou os temas com maiores prioridades (mais próximas de 1), tanto para o teste da ANEB, quanto para a Prova Brasil.

Cabe uma ressalva com relação ao tema IV - Tratamento da Informação, que apresentava prioridade 2 e foi coberto com um número de itens inferior ao do tema II - Grandezas e Medidas, com prioridade 3. O tema IV é composto por apenas dois

descritores, número inferior aos demais. Se houvesse a previsão que o mesmo apresentasse um número de itens semelhante ao do tema I - Espaço e Forma, por apresentarem mesma prioridade, ou seja, cerca de 30% dos testes, um número muito grande de itens cobriria os descritores 36 e 37 (únicos representantes do tema IV). Os especialistas responsáveis pela elaboração dos testes decidiram por abrir mão desse critério de prioridades especificamente para o tema IV para evitar uma supercobertura desses dos seus descritores.

Considerando os testes da ANEB e da Prova Brasil como um todo, a tabela 6.6 apresenta as freqüências e os percentuais de itens por descritor da matriz, bem como a diferença entre percentuais de itens por descritor. Os dados estão ordenados em função das diferenças entre percentuais. Neste caso, quando o valor é negativo, há um percentual menor de itens da ANEB cobrindo o descritor, em comparação à Prova Brasil. Quando o valor é positivo, há um percentual maior de itens da ANEB cobrindo o descritor, em comparação à Prova Brasil.

Tabela 6.6 - Número, percentual de itens por descritor e diferença entre percentuais dos testes de matemática, 8ª série EF, ANEB e Prova Brasil.

| Descritor | ANEB | | Prova Brasil | | Diferença de % |
|---|---|---|---|---|---|
| | n itens | % | n itens | % | |
| D17 | 7 | 5,8 | 0 | 0,0 | 5,8 |
| D18 | 5 | 4,1 | 0 | 0,0 | 4,1 |
| D22 | 5 | 4,1 | 0 | 0,0 | 4,1 |
| D36 | 4 | 3,3 | 0 | 0,0 | 3,3 |
| D13 | 3 | 2,5 | 0 | 0,0 | 2,5 |
| D20 | 5 | 4,1 | 2 | 2,5 | 1,7 |
| D04 | 3 | 2,5 | 1 | 1,2 | 1,2 |
| D06 | 3 | 2,5 | 1 | 1,2 | 1,2 |
| D25 | 3 | 2,5 | 1 | 1,2 | 1,2 |
| D35 | 3 | 2,5 | 1 | 1,2 | 1,2 |
| D23 | 4 | 3,3 | 2 | 2,5 | 0,8 |
| D34 | 4 | 3,3 | 2 | 2,5 | 0,8 |
| D16 | 5 | 4,1 | 3 | 3,7 | 0,4 |
| D05 | 2 | 1,7 | 1 | 1,2 | 0,4 |
| D01 | 3 | 2,5 | 2 | 2,5 | 0,0 |
| D27 | 3 | 2,5 | 2 | 2,5 | 0,0 |
| D02 | 4 | 3,3 | 3 | 3,7 | -0,4 |
| D07 | 4 | 3,3 | 3 | 3,7 | -0,4 |
| D28 | 4 | 3,3 | 3 | 3,7 | -0,4 |
| D29 | 4 | 3,3 | 3 | 3,7 | -0,4 |
| D21 | 1 | 0,8 | 1 | 1,2 | -0,4 |
| D37 | 5 | 4,1 | 4 | 4,9 | -0,8 |
| D03 | 2 | 1,7 | 2 | 2,5 | -0,8 |
| D14 | 2 | 1,7 | 2 | 2,5 | -0,8 |
| D15 | 2 | 1,7 | 2 | 2,5 | -0,8 |
| D26 | 2 | 1,7 | 2 | 2,5 | -0,8 |
| D10 | 3 | 2,5 | 3 | 3,7 | -1,2 |
| D11 | 3 | 2,5 | 3 | 3,7 | -1,2 |
| D24 | 3 | 2,5 | 3 | 3,7 | -1,2 |
| D09 | 4 | 3,3 | 4 | 4,9 | -1,6 |
| D08 | 2 | 1,7 | 3 | 3,7 | -2,1 |
| D30 | 2 | 1,7 | 3 | 3,7 | -2,1 |
| D12 | 3 | 2,5 | 4 | 4,9 | -2,5 |
| D19 | 3 | 2,5 | 4 | 4,9 | -2,5 |
| D33 | 3 | 2,5 | 4 | 4,9 | -2,5 |
| D32 | 1 | 0,8 | 3 | 3,7 | -2,9 |
| D31 | 2 | 1,7 | 4 | 4,9 | -3,3 |
| Total | 121 | 100,0 | 81 | 100,0 | - |

Considerando os testes como um todo e não restringindo a que tema da matriz os descritores se referem, foram observadas diferenças entre o alinhamento dos testes (Bhola, Impara & Buckendahl, 2003; Herman, Webb e Zuniga, 2002) em relação à matriz. O teste da ANEB abarca todos os descritores da matriz, com um quantitativo variando de 1 a 7

itens por descritor. Por sua vez, a Prova Brasil não contemplou a matriz completa. Observou-se que cinco descritores não foram cobertos por nenhum item.

Nos extremos inferior e superior da tabela 6.6 são destacados os descritores com diferenças superiores a 2% entre as avaliações. Os descritores 17, 18, 22, 36 e 13 foram cobertos com um número maior na ANEB que na Prova Brasil. Já os descritores 88, 30, 12, 19, 33, 32 e 31 foram cobertos com um percentual maior de itens na Prova Brasil que na ANEB.

### 6.2.2 Características psicométricas dos testes

Os procedimentos utilizados para a calibração foram razoavelmente semelhantes entre ANEB e Prova Brasil. Com base no BILOG-MG (versão 1), os seguintes procedimentos foram utilizados para a ambas as calibrações: (a) os parâmetros do SAEB 2003 foram mantidos fixos e transformados para que a 8ª série EF de 2003 tivesse media 0 e DP 1; (b) A referência do SAEB foi a 8ª série EF de 1997; (c) foram considerados itens comuns com o SAEB 2003; (d) Valores idênticos referentes aos comandos de calibração NQPT, NEWTON, CRIT, IDIST, NORMAL, READPRI, NOFLOAT; (d) *Prioris* (TPRIOR,SPRIOR,GPRIOR) foram utilizadas;

Os seguintes procedimentos foram diferentes entre as calibrações: (a) para a Prova Brasil, itens comuns entre séries de 2005 não foram utilizados, como na ANEB; (b) para a Prova Brasil, uma amostra dos respondentes foi utilizada para a calibração, enquanto na ANEB, as respostas de todos os respondentes foram consideradas; (c) a Prova Brasil utilizou os valores para os comandos DIAGNOSIS=0 e REFERENCE=1, enquanto a ANEB utilizou DIAGNOSIS=2 e REFERENCE=2. O comando REFERENCE é utilizado para resolver a indeterminância da localização e da escala da variável latente. No caso, quando REFERENCE é maior que 0 (ambas as calibrações), a média e o DP do grupo *i* são 0 e 1, respectivamente.

Tendo por base 155 itens do teste de 8ª série EF da ANEB e 81 itens do teste da mesma série da Prova Brasil, a tabela 6.7 apresenta os resultados de média e de DP dos parâmetros da TRI.

Tabela 6.7 - Parâmetros psicométricos dos itens estimados pela TRI - testes de matemática, 8ª série EF, ANEB e Prova Brasil.

| Teste | n itens | *a* | | *b* | | *c* | |
|---|---|---|---|---|---|---|---|
| | | Média | DP | Média | DP | Média | DP |
| ANEB | 155 | 1,24 | 0,61 | 0,71 | 1,26 | 0,19 | 0,09 |
| Prova Brasil | 81 | 1,87 | 0,70 | 0,79 | 1,03 | 0,19 | 0,08 |

A Prova Brasil apresenta parâmetro *b* médio superior em relação à ANEB (0,79 e 0,71, respectivamente), considerando todos os itens válidos. Cabe ressaltar que a dificuldade inferior da ANEB pode ter sofrido a influência dos itens de 4ª série EF incluídos no teste. Considerando apenas os itens de 4ª série EF, o parâmetro *b* médio passa de 0,71 para -0,49.

A Prova Brasil se mostrou de modo geral mais discriminativa que o teste da ANEB (Parâmetro *a* de 1,87 e 1,24, respectivamente). Os resultados médios referentes ao parâmetro *c* foram iguais entre as avaliações. Calculando-se os parâmetros da TRI por bloco da ANEB 2005, observam-se os resultados contidos na tabela 6.8.

Tabela 6.8 - Parâmetros psicométricos dos itens estimados pela TRI por Bloco - teste de matemática, 8ª série EF, ANEB.

| Bloco | n itens | *a* | | | | *b* | | | | *c* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Média | DP | Mín | Máx | Média | DP | Mín | Máx | Média | DP | Mín | Máx |
| 1 | 13 | 0,9 | 0,5 | 0,4 | 2,2 | 0,8 | 1,4 | -2,0 | 2,6 | 0,2 | 0,1 | 0,0 | 0,3 |
| 2 | 13 | 1,3 | 0,6 | 0,5 | 2,7 | 1,1 | 1,2 | -1,1 | 2,5 | 0,2 | 0,1 | 0,1 | 0,3 |
| 3 | 12 | 1,5 | 0,6 | 0,7 | 2,7 | 1,1 | 1,2 | -1,3 | 2,7 | 0,1 | 0,1 | 0,0 | 0,2 |
| 4 | 10 | 1,2 | 0,7 | 0,5 | 2,7 | 0,9 | 1,0 | -0,8 | 2,1 | 0,2 | 0,1 | 0,0 | 0,4 |
| 5 | 10 | 1,1 | 0,4 | 0,6 | 1,7 | 0,9 | 1,3 | -1,5 | 2,9 | 0,2 | 0,1 | 0,0 | 0,3 |
| 6 | 12 | 1,2 | 0,5 | 0,6 | 1,9 | 1,6 | 1,2 | -1,0 | 2,7 | 0,2 | 0,1 | 0,0 | 0,4 |
| 7 | 12 | 1,0 | 0,2 | 0,8 | 1,6 | 0,9 | 1,2 | -1,6 | 2,8 | 0,2 | 0,1 | 0,0 | 0,3 |
| 8 | 13 | 1,0 | 0,6 | 0,3 | 2,6 | 1,0 | 0,6 | 0,2 | 2,0 | 0,2 | 0,1 | 0,1 | 0,4 |
| 9 | 13 | 1,1 | 0,5 | 0,4 | 2,6 | -0,5 | 1,4 | -2,6 | 1,7 | 0,2 | 0,1 | 0,0 | 0,4 |
| 10 | 10 | 1,0 | 0,4 | 0,4 | 1,8 | -0,3 | 1,2 | -2,9 | 1,0 | 0,2 | 0,1 | 0,0 | 0,3 |
| 11 | 11 | 1,1 | 0,2 | 0,9 | 1,6 | -0,6 | 0,9 | -2,0 | 1,1 | 0,2 | 0,1 | 0,1 | 0,2 |
| 12 | 13 | 1,9 | 0,8 | 1,1 | 3,6 | 1,1 | 1,0 | -1,0 | 2,4 | 0,2 | 0,1 | 0,1 | 0,3 |
| 13 | 13 | 1,7 | 0,8 | 0,9 | 3,7 | 0,9 | 0,5 | 0,2 | 1,8 | 0,2 | 0,1 | 0,1 | 0,3 |
| Média | - | 1,2 | 0,5 | 0,6 | 2,4 | 0,7 | 1,1 | -1,3 | 2,2 | 0,2 | 0,1 | 0,0 | 0,3 |
| DP | - | 0,3 | 0,2 | 0,2 | 0,7 | 0,7 | 0,3 | 0,9 | 0,6 | 0,0 | 0,0 | 0,0 | 0,1 |
| Mínimo | 10 | 0,9 | 0,2 | 0,3 | 1,6 | -0,6 | 0,5 | -2,9 | 1,0 | 0,1 | 0,1 | 0,0 | 0,2 |
| Máximo | 13 | 1,9 | 0,8 | 1,1 | 3,7 | 1,6 | 1,4 | 0,2 | 2,9 | 0,2 | 0,1 | 0,1 | 0,4 |
| Amplitude | 3 | 1,0 | 0,6 | 0,7 | 2,1 | 2,2 | 0,9 | 3,2 | 1,9 | 0,1 | 0,1 | 0,1 | 0,2 |

O número de itens por bloco variou de 10 a 13. Como para os blocos 12 e 13, advindos do SAEB 2003, não houve exclusão de itens, o teste contou com 26 itens comuns entre anos. Contou também com 34 itens oriundos da 4ª série EF da ANEB 2005.

O parâmetro *b* médio dos blocos variou de -0,6 a 1,6 e em média o bloco apresentou dificuldade de 0,7. Os blocos mais fáceis originaram-se de 4ª série EF (9, 10 e 11) e o mais difícil foi o bloco 6 (1,6). Os dois blocos mais discriminativos foram o 12 e o 13 (parâmetros *a* de 1,9 e 1,7), oriundos da 8ª série do SAEB 2003. Esses foram montados propositalmente com itens bastante discriminativos para garantir a equalização entre anos. O bloco 1 foi o que apresentou a menor discriminação média (0,9). Os parâmetros *c* médios dos blocos apresentaram pouca variabilidade com amplitude de 0,1. O bloco 3 se mostrou com menor probabilidade de acerto ao acaso médio (0,1). Estimando-se os

parâmetros da TRI por bloco da Prova Brasil 2005, verificaram-se os resultados apresentados na tabela 6.9.

Tabela 6.9 - Parâmetros psicométricos dos itens estimados pela TRI por Bloco - teste de matemática, 8ª série EF, Prova Brasil.

| Bloco | n itens | *a* | | | | *b* | | | | *c* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Média | DP | Mín | Máx | Média | DP | Mín | Máx | Média | DP | Mín | Máx |
| 1 | 12 | 2,0 | 0,9 | 1,0 | 3,7 | 0,6 | 1,2 | -1,9 | 2,0 | 0,2 | 0,1 | 0,0 | 0,3 |
| 2 | 12 | 2,1 | 0,8 | 0,8 | 3,2 | 0,8 | 1,3 | -1,6 | 2,8 | 0,2 | 0,1 | 0,0 | 0,3 |
| 3 | 9 | 1,5 | 0,3 | 0,8 | 1,9 | 0,6 | 1,1 | -1,5 | 2,2 | 0,2 | 0,1 | 0,0 | 0,3 |
| 4 | 12 | 2,2 | 0,8 | 1,3 | 3,6 | 0,9 | 1,0 | -0,8 | 2,1 | 0,2 | 0,1 | 0,1 | 0,3 |
| 5 | 12 | 1,8 | 0,7 | 1,0 | 3,7 | 1,0 | 0,6 | 0,2 | 1,9 | 0,2 | 0,1 | 0,1 | 0,3 |
| 6 | 12 | 1,7 | 0,4 | 1,1 | 2,5 | 0,7 | 1,1 | -1,6 | 2,7 | 0,2 | 0,1 | 0,0 | 0,3 |
| 7 | 12 | 1,6 | 0,6 | 0,8 | 2,8 | 0,9 | 1,0 | -1,2 | 2,3 | 0,2 | 0,1 | 0,0 | 0,3 |
| Média | - | 1,9 | 0,7 | 1,0 | 3,0 | 0,8 | 1,0 | -1,2 | 2,3 | 0,2 | 0,1 | 0,1 | 0,3 |
| DP | - | 0,2 | 0,2 | 0,2 | 0,7 | 0,1 | 0,2 | 0,7 | 0,4 | 0,0 | 0,0 | 0,0 | 0,0 |
| Mínimo | 9 | 1,5 | 0,3 | 0,8 | 1,9 | 0,6 | 0,6 | -1,9 | 1,9 | 0,2 | 0,1 | 0,0 | 0,3 |
| Máximo | 12 | 2,2 | 0,9 | 1,3 | 3,7 | 1,0 | 1,3 | 0,2 | 2,8 | 0,2 | 0,1 | 0,1 | 0,3 |
| Amplitude | 3 | 0,7 | 0,6 | 0,5 | 1,8 | 0,4 | 0,7 | 2,1 | 0,9 | 0,1 | 0,0 | 0,1 | 0,1 |

Os três itens excluídos da Prova Brasil localizavam-se no bloco 3, passando esse a ficar com 9 itens. Os demais blocos apresentaram 12 itens. O parâmetro *b* médio variou de 0,6 a 1,0 entre os blocos. O bloco com maior dificuldade foi o 5 (1,0) e os de menores dificuldades foram o 1 e o 3 (0,6). Em média, a dificuldade dos blocos foi de 0,8.

O parâmetro *a* médio por bloco variou de 1,5 a 2,2 para a Prova Brasil. Quando esses resultados são comparados com a ANEB, observa-se que a discriminação dos blocos é superior na Prova Brasil. Na ANEB, 10 dos 13 blocos apresentaram parâmetro *a* médio inferior a 1,5. O parâmetro *c* médio apresentou uma baixa variabilidade com amplitude de 0,1, sendo que a menor probabilidade de acerto ao acaso foi de 0,17.

A análise das estatísticas por bloco é importante já que é a base para a construção dos cadernos dos testes. Por sua vez, como os estudantes respondem a cadernos de testes, cabe sua análise psicométrica. Como o parâmetro *c* não apresentou muita variabilidade na análise por blocos, não será considerado na análise por cadernos e para efeito dos demais

estudos. A tabela 6.10 apresenta os parâmetros *a* e *b* da TRI dos cadernos de matemática, 8ª série EF, da ANEB.

Tabela 6.10 - Parâmetros psicométricos dos itens estimados pela TRI por Caderno - teste de matemática, 8ª série EF, ANEB.

| Caderno | n itens | *a* | | *b* | |
|---|---|---|---|---|---|
| | | Média | DP | Média | DP |
| 1 | 36 | 1,1 | 0,5 | 1,0 | 1,3 |
| 2 | 37 | 1,3 | 0,6 | 1,3 | 1,2 |
| 3 | 34 | 1,3 | 0,6 | 1,0 | 1,1 |
| 4 | 33 | 1,1 | 0,6 | 0,9 | 0,9 |
| 5 | 35 | 1,1 | 0,5 | 0,6 | 1,6 |
| 6 | 34 | 1,1 | 0,4 | 0,8 | 1,4 |
| 7 | 36 | 1,0 | 0,4 | 0,5 | 1,1 |
| 8 | 39 | 1,3 | 0,7 | 0,5 | 1,3 |
| 9 | 36 | 1,3 | 0,7 | 0,0 | 1,3 |
| 10 | 34 | 1,0 | 0,4 | 0,0 | 1,3 |
| 11 | 37 | 1,4 | 0,7 | 0,6 | 1,3 |
| 12 | 38 | 1,7 | 0,7 | 1,0 | 0,9 |
| 13 | 36 | 1,3 | 0,7 | 0,9 | 1,0 |
| 14 | 38 | 1,1 | 0,6 | 1,0 | 1,1 |
| 15 | 36 | 1,2 | 0,6 | 0,5 | 1,4 |
| 16 | 32 | 1,2 | 0,5 | 0,6 | 1,4 |
| 17 | 33 | 1,2 | 0,5 | 0,7 | 1,4 |
| 18 | 35 | 1,4 | 0,7 | 1,0 | 1,1 |
| 19 | 38 | 1,3 | 0,7 | 1,1 | 0,8 |
| 20 | 38 | 1,0 | 0,4 | 0,4 | 1,5 |
| 21 | 36 | 1,1 | 0,6 | 0,7 | 1,2 |
| 22 | 36 | 1,2 | 0,5 | 0,0 | 1,4 |
| 23 | 33 | 1,4 | 0,7 | 0,6 | 1,2 |
| 24 | 34 | 1,4 | 0,6 | 0,4 | 1,2 |
| 25 | 38 | 1,3 | 0,7 | 1,2 | 1,2 |
| 26 | 38 | 1,3 | 0,7 | 1,0 | 1,0 |
| Média | - | 1,2 | 0,6 | 0,7 | 1,2 |
| DP | - | 0,2 | 0,1 | 0,4 | 0,2 |
| Mínimo | 32 | 1,0 | 0,4 | 0,0 | 0,8 |
| Máximo | 39 | 1,7 | 0,7 | 1,3 | 1,6 |
| Amplitude | 7 | 0,7 | 0,4 | 1,3 | 0,7 |

O quantitativo de itens por caderno variou de 32 a 39. Em média os cadernos da ANEB apresentaram parâmetro *b* por caderno de 0,7. Há, no entanto, variabilidade em seus índices de dificuldade já que a amplitude é de 1,3 e o desvio padrão de 0,4. O caderno mais difícil foi o 2 (1,3) e os mais fáceis o 9 e o 10 (0,0). O parâmetro *a* médio por caderno, por sua vez, foi de 1,2. Os cadernos menos discriminativos foram o 7, o 10 e o 20 (parâmetro *a* de 1,0). O mais discriminativo foi o caderno 12 (1,7).

A tabela 6.11 apresenta os parâmetros *a* e *b* da TRI dos cadernos de matemática, 8ª série EF, da Prova Brasil.

Tabela 6.11 - Parâmetros psicométricos dos itens estimados pela TRI por Caderno - teste de matemática, 8ª série EF, Prova Brasil.

| Caderno | n itens | *a* | | *b* | |
|---|---|---|---|---|---|
| | | Média | DP | Média | DP |
| 1 | 24 | 2,0 | 0,8 | 0,7 | 1,2 |
| 2 | 21 | 1,8 | 0,7 | 0,7 | 1,2 |
| 3 | 21 | 1,9 | 0,7 | 0,7 | 1,0 |
| 4 | 24 | 2,0 | 0,8 | 0,9 | 0,8 |
| 5 | 24 | 1,8 | 0,6 | 0,8 | 0,9 |
| 6 | 24 | 1,7 | 0,5 | 0,8 | 1,0 |
| 7 | 24 | 1,8 | 0,8 | 0,8 | 1,1 |
| 8 | 21 | 1,8 | 0,7 | 0,6 | 1,2 |
| 9 | 24 | 2,1 | 0,8 | 0,8 | 1,1 |
| 10 | 21 | 1,7 | 0,6 | 0,8 | 0,9 |
| 11 | 24 | 2,0 | 0,6 | 0,8 | 1,0 |
| 12 | 24 | 1,7 | 0,7 | 0,9 | 0,8 |
| 13 | 24 | 1,9 | 0,7 | 0,7 | 1,2 |
| 14 | 24 | 1,9 | 0,7 | 0,9 | 1,1 |
| 15 | 24 | 2,1 | 0,8 | 0,7 | 1,1 |
| 16 | 24 | 1,9 | 0,7 | 0,9 | 1,0 |
| 17 | 21 | 1,7 | 0,4 | 0,7 | 1,1 |
| 18 | 24 | 1,9 | 0,7 | 0,9 | 0,9 |
| 19 | 24 | 1,9 | 0,8 | 0,8 | 1,0 |
| 20 | 24 | 1,9 | 0,6 | 0,8 | 1,2 |
| 21 | 21 | 1,6 | 0,5 | 0,8 | 1,0 |
| Média | - | 1,9 | 0,7 | 0,8 | 1,0 |
| DP | - | 0,1 | 0,1 | 0,1 | 0,1 |
| Mínimo | 21 | 1,6 | 0,4 | 0,6 | 0,8 |
| Máximo | 24 | 2,1 | 0,8 | 0,9 | 1,2 |
| Amplitude | 3 | 0,5 | 0,4 | 0,3 | 0,4 |

O número de itens por caderno variou entre 21 e 24. Em média os cadernos da Prova Brasil apresentaram parâmetro *b* de 0,8. Embora haja variabilidade com relação ao parâmetro *b*, essa não é alta (amplitude de 0,3 e DP de 0,1). O parâmetro *a* médio por caderno, por sua vez, foi de 1,9, superior à discriminação média da ANEB (1,2). O caderno menos discriminativo foi o 21, com parâmetro *a* de 1,6. Os cadernos mais discriminativos foram o 9 e o 15 com parâmetro *a* de 2,1.

### 6.2.3 Dimensionalidade dos testes

Um estudo de verificação da unidimensionalidade do teste de matemática 8ª série EF foi realizado para a ANEB 2005 (CESPE, 2007c). Utilizou-se a Análise Fatorial de Informação Plena por meio do software *Testfact 3* (Wilson, Wood & Gibbons, 1991). Além dos itens desconsiderados nas etapas de análise prévias, dois itens de matemática 8ª série EF foram excluídos para que fosse possível o cálculo das correlações tetracóricas por apresentarem problemas de convergência (CESPE, 2007c).

Após a exclusão dos itens, a análise da dimensionalidade foi realizada. Após a renormalização dos fatores de expansão dos estudantes de forma a somar 2.000, procedeu-se ao cálculo do qui-quadrado com a verificação do ajuste do modelo de 1 e 2 fatores (CESPE, 2007c).

Após o cálculo da mudança do Qui-quadrado entre os modelos de 1 e de 2 fatores, essa foi dividida por uma constante igual a 3 (Laros, Pasquali & Rodrigues, 2000; Wilson, Wood & Gibbons, 1991). Com base nesses resultados, a significância da mudança no Qui-quadrado corrigida (índice de unidimensionalidade) foi avaliada por meio do cálculo da razão entre a mudança corrigida (169,96) e os graus de liberdade (152,00). Obteve-se como resultado 1,1. O resultado positivo e inferior a 2,0, para matemática 8ª série EF, sugere que o modelo de dois fatores se ajusta melhor aos dados que o de um fator, mas sem significância estatística.

De acordo com os critérios de que "(...) o primeiro fator deve apresentar uma explicação da variância maior que a do segundo fator" (CESPEc, 2007) e que no mínimo 20% da variância deve ser explicada para que se possa obter estimativas mais confiáveis dos parâmetros dos itens (Kirisci, Hsu & Yu, 2001), o teste pode ser considerado unidimensional. O percentual de variância explicada para o primeiro fator foi de 44,6 e para o segundo fator, de 4,2. A razão entre os percentuais de variância explicada para o primeiro e o segundo fator foi de 10,6.

Ainda, 16 itens de matemática 8ª série EF da ANEB 2005 apresentaram carga negativa no primeiro fator (CESPE, 2007b). As cargas fatoriais foram reestimadas após a exclusão desses itens. Após nova análise fatorial, obteve-se razão entre a mudança corrigida de 159,81 com 136,00 graus de liberdade. A razão entre os valores foi de 1,2, sugerindo ainda que o modelo de dois fatores se ajusta melhor aos dados que o de um fator, mas sem significância estatística.

Após a nova análise fatorial, a razão entre as variâncias explicadas pelos dois primeiros fatores (11,5) foi superior à encontrada para a primeira análise (10,6), o que

indica uma aproximação à unidimensionalidade quando se excluem os itens com cargas negativas no primeiro fator. Após a exclusão desses itens, as cargas fatoriais variaram de 0,07 a 0,74, com média de 0,43 e DP de 0,13. Encontraram-se ainda 20 itens com carga fatoriais inferiores a 0,30.

CESPE (2007c) conclui que a unidimensionalidade para matemática 8ª série EF da ANEB 2005 foi aceita. Laros, Pasquali & Rodrigues (2000) propuseram a exclusão dos itens com baixas cargas fatoriais no fator principal de forma a propiciar "(...) um aumento da validade do construto das provas do SAEB, numa melhoria das estimativas da proficiência dos alunos e dos parâmetros dos itens, num aperfeiçoamento do processo de equalização e numa diminuição do número de itens com viés" (p.69). Condé (2002) e Condé e Laros (2007) também verificaram que a exclusão dos itens com cargas fatoriais inferiores a 0,20 e a 0,30 influenciam na propriedade de invariância das estimativas de habilidade dos estudantes. Cabe ressaltar que, embora o estudo de dimensionalidade tenha sido realizado no âmbito da ANEB 2005, não foi utilizado para tomada de decisão antes da divulgação dos resultados finais da avaliação. Assim, os dois itens que não convergiram e os 16 com cargas fatoriais negativas no fator principal não foram retirados para efeitos de estimação das habilidades dos sujeitos.

Não foi encontrado na literatura nenhum estudo que avaliasse a unidimensionalidade da Prova Brasil 2005. Assim, não se sabe ao certo o grau de unidimensionalidade desse teste e o quanto influenciou na propriedade de invariância das estimativas de habilidade dos estudantes. Dessa forma, não se pode realizar, no âmbito do presente trabalho, inferências de comparação entre ANEB e Prova Brasil no que tange à unidimensionalidade.

## 6.3 Estudo 3: Estimação das habilidades dos estudantes da ANEB sob novas configurações de teste

O estudo 3 foi composto por seis análises que buscaram estimar as habilidades dos estudantes da ANEB 2005 por meio do software BILOG-MG. Para todas as análises, considerou-se a base de dados completa. As estatísticas calculadas e apresentadas nesta seção, por sua vez, contemplam exclusivamente os estudantes de escolas públicas urbanas. Consideraram-se os pesos para cálculo dessas estatísticas, assim os resultados representam 2.515.731 estudantes da ANEB. Para as duas primeiras análises, foram utilizados os testes da ANEB em sua versão original com 155 itens e buscou-se verificar se atingiam os mesmos resultados do INEP, exclusivamente para efeito de controle. As demais quatro

análises buscaram estimar as habilidades dos estudantes a partir da redução pela seleção de itens da ANEB original.

### 6.3.1 Estimação das habilidades de acordo com os critérios utilizados pelo INEP

Utilizando o programa BILOG-MG (versão 1) e a programação (ou os arquivos *.blm*) disponibilizado pelo INEP e que foi utilizada para estimação das habilidades dos estudantes para a ANEB 2005, procedeu-se a replicação da análise. O objetivo foi atingir os mesmos resultados da fase 3 do software para dar continuidade às outras análises do presente estudo.

Os resultados obtidos foram idênticos aos estimados pelo INEP. Para matemática 8ª série EF da ANEB, em seus estratos de escolas públicas e urbanas do Brasil, a média da estimativa de habilidade obtida foi de -0,3283 com DP de 0,8157. Ainda mínimo de -2,79 e máximo de 2,98. Todos idênticos aos divulgados pelo INEP, como era de se esperar, já que estava sendo utilizada a mesma programação.

### 6.3.2 Estimação das habilidades a partir da desvinculação dos itens entre séries para o ano de 2005

Itens da 4ª série EF incluídos no teste de 8ª série EF da ANEB e itens de 8ª série EF incluídos na 3ª série EM foram desvinculados em relação ao comando de nomes (INAMES, do arquivo *.blm*), parâmetros *a*, *b* e *c* (TEST) e Grupos (GROUP) de tal forma que quaisquer manipulações posteriores do programa, em termos de exclusão de itens para composição de novas estruturas de testes, fossem possíveis. Foi realizada adicionalmente a repetição do conjunto dos parâmetros *a*, *b* e *c* desses itens para a série posterior. Esperavam-se resultados idênticos já que essa modificação mantém um conjunto de itens de 4ª e 8ª e de 8ª e 3ª com parâmetros idênticos.

Os resultados obtidos foram iguais aos obtidos na rodada com a programação original realizada pelo INEP. Para matemática 8ª série da ANEB, em seus estratos de escolas públicas e urbanas do Brasil, considerando-se os pesos amostrais, foi obtida a média de habilidade de -0,3283 com DP de 0,8157; mínimo de -2,79 e máximo de 2,98.

### 6.3.3 Teste A: estimação das habilidades a partir de 104 itens com parâmetros similares aos da ANEB

O teste A foi composto a partir da redução do número de itens do teste de matemática 8ª série EF original da ANEB, com base na programação do BILOG-MG

citada no tópico 6.3.2. Foi constituído a partir da redução de 155 para 104 itens, de forma que fossem considerados 24 itens para cada estudante de 8ª série EF da ANEB, mesmo número (ou aproximadamente o mesmo, já que seis cadernos da Prova Brasil foram compostos por 21 itens) considerado para a Prova Brasil dessa mesma série e disciplina. Reforça-se a informação que há diferença entre o Teste A e a Prova Brasil no total de itens considerados (104 e 81, respectivamente). Os critérios de modificação do teste original foram detalhados no método do presente trabalho.

A redução do número de itens não resultou em um distanciamento de alinhamento à matriz muito significativo com relação à ANEB e à Prova Brasil em termos de percentuais de itens por tema (Tabela 6.12).

Tabela 6.12 - Número e percentual de itens por tema dos testes de matemática, 8ª série EF, Prova Brasil, ANEB e Teste A.

| Tema | Prova Brasil | | ANEB | | Teste A | |
|---|---|---|---|---|---|---|
| | n itens | % | n itens | % | n itens | % |
| I | 26 | 32,1 | 44 | 28,4 | 29 | 27,9 |
| II | 8 | 9,9 | 16 | 10,3 | 12 | 11,5 |
| III | 43 | 53,1 | 86 | 55,5 | 55 | 52,9 |
| IV | 4 | 4,9 | 9 | 5,8 | 8 | 7,7 |
| Total | 81 | 100,0 | 155 | 100,0 | 104 | 100,0 |

O distanciamento maior referiu-se ao tema I, em que o Teste A apresentou 4% a menos de itens do teste como um todo comparado à Prova Brasil. No entanto, o tema foi coberto por um número semelhante de itens para o Teste A e a Prova Brasil (26 e 29, respectivamente). O rigor na manutenção de um número semelhante de itens por tema, quando da redução de itens da ANEB, teve como objetivo não deixar que o desequilíbrio na cobertura entre o Teste A, a Prova Brasil e a ANEB contribuísse nas novas estimativas de habilidade.

Tendo em vista a existência de um razoável desequilíbrio entre os descritores, considerou-se que os resultados obtidos pela análise por tema são suficientes para demonstrar similaridades entre os testes em termos de cobertura. Certamente um refinamento da análise por descritor será bastante útil, mas não foi realizado no presente estudo.

A tabela 6.13 apresenta os parâmetros *a* e *b* médios e as estatísticas das estimativas de habilidade obtidos para o Teste A, bem como os resultados para a Prova Brasil e para a ANEB.

Tabela 6.13 - Parâmetros psicométricos dos itens e habilidades estimadas - teste de matemática, 8ª série EF, Prova Brasil, ANEB e Teste A.

| Teste | n itens | *a* | | *b* | | Habilidade | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Média | DP | Média | DP | N | Média | DP | Mínimo | Máximo |
| Prova Brasil | 81 | 1,87 | 0,70 | 0,79 | 1,03 | 1.610.073 | -0,1786 | 0,7617 | -2,11 | 2,79 |
| ANEB | 155 | 1,24 | 0,61 | 0,71 | 1,26 | 2.515.731 | -0,3283 | 0,8157 | -2,79 | 2,98 |
| Teste A | 104 | 1,25 | 0,60 | 0,71 | 1,06 | 2.515.731 | -0,3072 | 0,7855 | -2,17 | 2,93 |

Mantendo-se praticamente os mesmo valores médios de discriminação e de dificuldade do Teste A com relação à ANEB original, observa-se um pequeno acréscimo da estimativa de habilidade média dos estudantes a partir do Teste A. A média da ANEB de -0,3283, calculada com 155 itens, passou a ser de -0,3072 com base nos 104 itens do Teste A. Essa alteração ocorreu na direção da média da Prova Brasil (-0,1786). Mas, quando se comparam os resultados do Teste A com os da Prova Brasil, observa-se uma distância grande entre as médias. A distância entre as médias da Prova Brasil e ANEB original foi de 0,14 e a distância entre as médias da Prova Brasil e o Teste A foi de 0,12.

O Teste A continuou apresentando discriminação e dificuldade média inferior à da Prova Brasil. A figura 6.3 apresenta a distribuição das estimativas de habilidade por faixa para a Prova Brasil, a ANEB original e o Teste A.

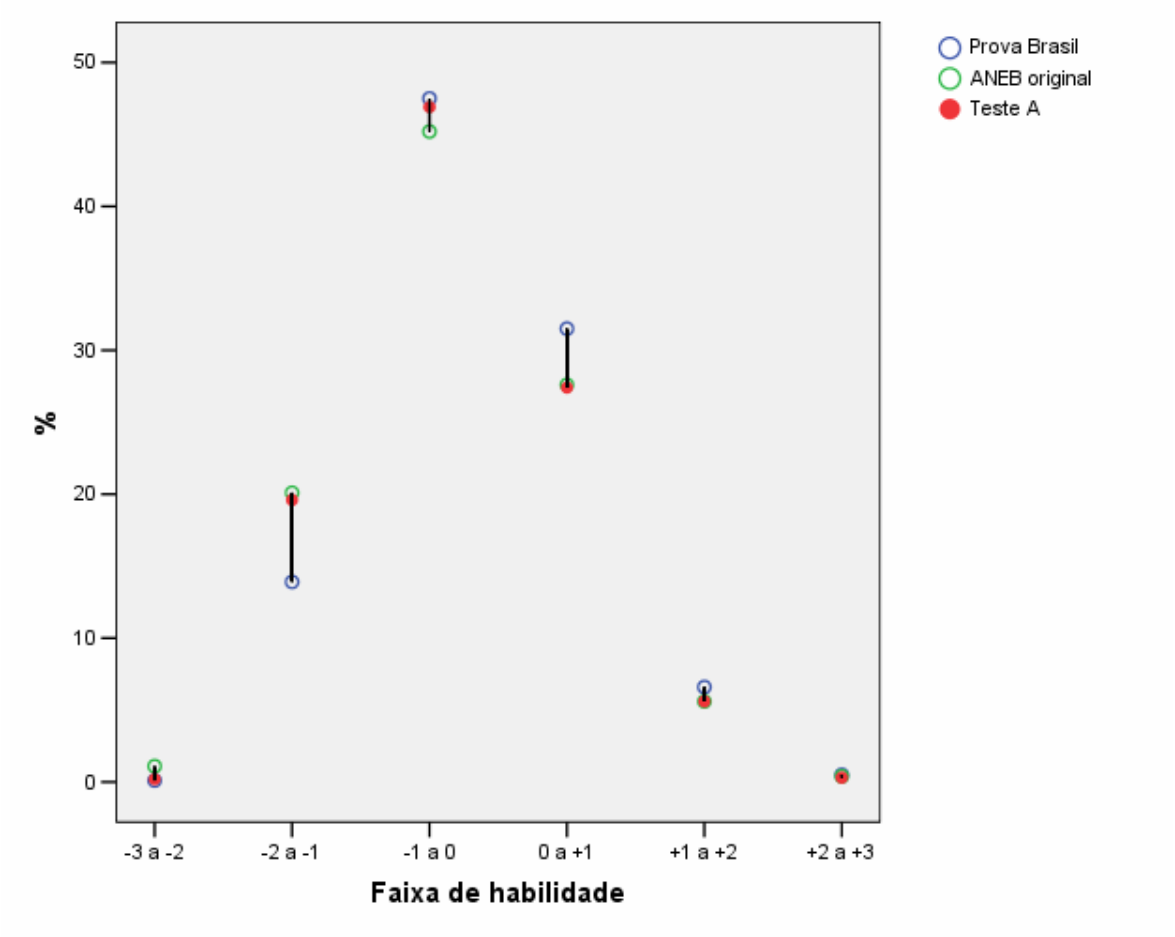


Figura 6.3 - Distâncias entre percentuais de estudantes por faixa de habilidade - matemática, 8ª série EF, Prova Brasil, ANEB original e Teste A.

Observa-se uma ligeira aproximação dos percentuais de estudantes para as faixas -2 a -1 e -1 a 0 e um pequeno afastamento para a faixa de 0 a +1 do Teste A com a Prova Brasil, em comparação com os percentuais previamente observados para a ANEB original. Assim, houve uma pequena aproximação entre as distribuições para os segmentos da esquerda do gráfico. As distâncias referentes às faixas centrais continuaram grandes.

### 6.3.4 Teste B: estimação das habilidades a partir de 104 itens e da otimização da discriminação da ANEB

O Teste B foi composto a partir da redução do número de itens do teste da ANEB para 104 itens por meio da manutenção de seus itens mais discriminativos, com base na programação do BILOG-MG citada no tópico 6.3.2. A redução de 155 para 104 itens, permitiu que fossem considerados 24 itens para cada estudante de 8ª série EF da ANEB.

O Teste B apresentou parâmetro *a* médio de 1,46, valor máximo permitido a partir da seleção de oito itens por bloco e 24 por caderno da ANEB. Valor mais próximo do parâmetro *a* médio da Prova Brasil (1,87). Questionou-se até que ponto a discriminação do teste associada à aproximação do tamanho dos testes influenciaria a diferença entre os resultados das avaliações. A tabela 6.14 apresenta os resultados de habilidades obtidos para o Teste B, bem como os resultados para a Prova Brasil e para a ANEB.

Tabela 6.14 - Parâmetros psicométricos dos itens e habilidades estimadas - teste de matemática, 8ª série EF, Prova Brasil, ANEB original, Teste A e Teste B.

| Teste | n itens | *a* | | *b* | | Habilidade | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Média | DP | Média | DP | N | Média | DP | Mínimo | Máximo |
| Prova Brasil | 81 | 1,87 | 0,70 | 0,79 | 1,03 | 1.610.073 | -0,1786 | 0,7617 | -2,11 | 2,79 |
| ANEB | 155 | 1,24 | 0,61 | 0,71 | 1,26 | 2.515.731 | -0,3283 | 0,8157 | -2,79 | 2,98 |
| Teste B | 104 | 1,46 | 0,62 | 0,97 | 1,10 | 2.515.731 | -0,2906 | 0,7506 | -2,59 | 2,97 |

Aumentando-se os valores médios de parâmetro *a*, associado a um aumento do parâmetro *b* médio, e diminuindo-se o tamanho do teste, com relação à ANEB, observou-se um pequeno acréscimo da estimativa de habilidade média dos estudantes a partir do Teste B. A média da ANEB de -0,3283, calculada com 155 itens e que tinha passado para -0,3072 para o Teste A, aumentou um pouco mais para o Teste B (-0,2906). Essa alteração ocorreu na direção da média da Prova Brasil, que foi de -0,1786. No entanto, quando se comparam os resultados do Teste B com os da Prova Brasil, observa-se ainda uma distância grande entre as médias. A distância entre as médias da Prova Brasil e ANEB (0,14) e a distância entre as médias da Prova Brasil e o Teste A (0,12) foram superiores à distância entre as médias da Prova Brasil para o Teste B (0,10).

Esse acréscimo na média foi acompanhado de uma redução da variabilidade. Se na ANEB original o DP foi de 0,82 e no Teste A de 0,79, para o Teste B, observou-se o DP de 0,75, mais próximo do DP da Prova Brasil (0,76). Isso sugere que a redução do tamanho do teste associado ao aumento de sua discriminação propiciou uma maior igualdade do desempenho dos estudantes.

A figura 6.4 apresenta a distribuição das estimativas de habilidade por faixa para a Prova Brasil, a ANEB e o Teste B.

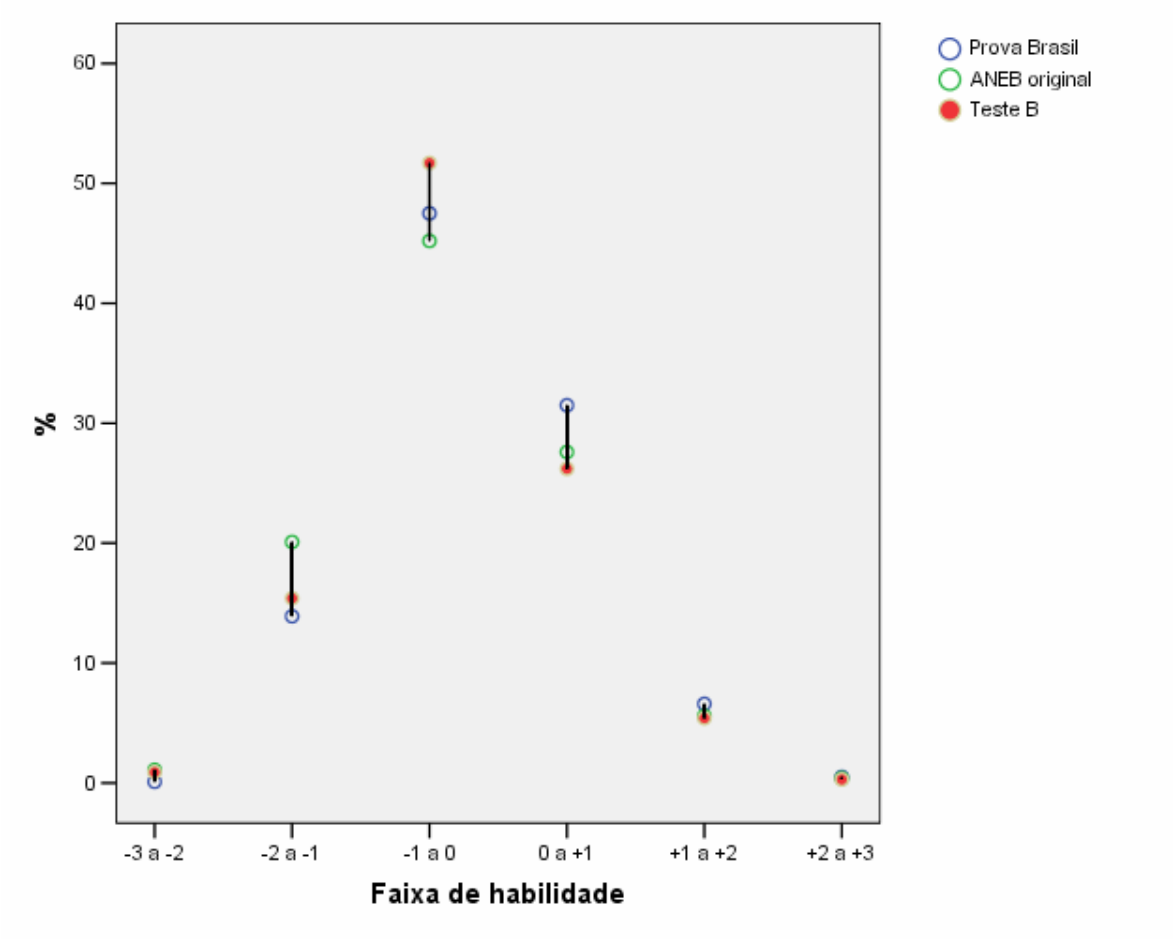


Figura 6.4 - Distâncias entre percentuais de estudantes por faixa de habilidade - matemática, 8ª série EF, Prova Brasil, ANEB e Teste B.

Para a faixa -2 a -1, houve uma aproximação significativa do percentual de estudantes se considerarmos o Teste B e Prova Brasil. Refletindo um pouco a diminuição da variabilidade propiciada pelo Teste B em comparação à ANEB, observa-se para o Teste B que, para a faixa de -1 a 0, houve um grande aumento do percentual de estudantes comparando-se com a ANEB, extrapolando, inclusive, o percentual obtido nessa faixa para a Prova Brasil. Associado a isso, observou-se uma diminuição do percentual de estudantes com estimativas de habilidade localizadas na faixa de 0 a +1 com base no Teste B e um aumento da distância para o percentual da Prova Brasil para essa faixa.

Os percentuais de estudantes, o número e o percentual de itens por faixa de habilidade e a média do parâmetro *a* para a Prova Brasil e o Teste B são apresentados nas tabelas 6.15 e 6.16.

Tabela 6.15 - Percentual de estudantes por faixa de habilidade estimada, número e percentual de itens, média e DP do parâmetro *a* - teste de matemática, 8ª série EF, Prova Brasil.

| Faixa | Habilidade | Itens | | *a* | |
|---|---|---|---|---|---|
| | % | n | % | Média | DP |
| -3 a -2 | 0,1 | 0 | 0,00 | - | - |
| -2 a -1 | 13,9 | 6 | 7,41 | 1,30 | 0,37 |
| -1 a 0 | 47,5 | 8 | 9,88 | 1,52 | 0,32 |
| 0 a +1 | 31,5 | 31 | 38,27 | 1,63 | 0,56 |
| +1 a +2 | 6,6 | 30 | 37,04 | 2,32 | 0,76 |
| +2 a +3 | 0,5 | 6 | 7,41 | 1,85 | 0,36 |
| Total | 100,0 | 81 | 100 | - | - |

Tabela 6.16 - Percentual de estudantes por faixa de habilidade estimada, número e percentual de itens, média e DP do parâmetro *a* - teste de matemática, 8ª série EF, Teste B.

| Faixa | Habilidade | Itens | | *a* | |
|---|---|---|---|---|---|
| | % | n | % | Média | DP |
| -3 a -2 | 0,9 | 2 | 1,92 | 1,10 | 0,07 |
| -2 a -1 | 15,4 | 4 | 3,85 | 1,01 | 0,04 |
| -1 a 0 | 51,7 | 12 | 11,54 | 1,12 | 0,26 |
| 0 a +1 | 26,2 | 34 | 32,69 | 1,39 | 0,63 |
| +1 a +2 | 5,4 | 33 | 31,73 | 1,71 | 0,72 |
| +2 a +3 | 0,3 | 19 | 18,27 | 1,53 | 0,47 |
| Total | 100,0 | 104 | 100 | - | - |

Na região central da escala das estimativas de habilidade, faixas de -1 a 0 e de 0 a +1, o percentual é invertido entre as avaliações. Observam-se 48% dos estudantes da Prova Brasil e 52% do Teste B para a faixa de -1 a 0. A Prova Brasil apresentou 32% dos estudantes de 0 a +1 enquanto o Teste B, 26%. Quatro por cento a mais do primeiro teste

para a primeira faixa e seis por cento a mais do segundo teste para a segunda faixa. Total de cada faixa para cada teste: 79% para a Prova Brasil e 78% para o Teste B. Considerou-se como hipótese a possibilidade de aproximadamente os mesmos estudantes estarem entre as faixas.

A Prova Brasil foi mais discriminativa que o Teste B para as faixas -1 a 0 (1,52 a 1,12) e 0 a +1 (1,63 a 1,39). Para o Teste B, foi o máximo que se conseguiu em termos discriminação para 104 itens e 24 respostas por estudantes. O Teste B não se distanciou muito em termos de percentuais de itens por tema em comparação à Prova Brasil e à ANEB (Tabela 6.17).

Tabela 6.17 - Número e percentual de itens por tema dos testes de matemática, 8ª série EF, para Prova Brasil, ANEB e Teste B.

| Tema | Prova Brasil | | ANEB | | Teste B | |
|---|---|---|---|---|---|---|
| | n itens | % | n itens | % | n itens | % |
| I | 26 | 32,1 | 44 | 28,4 | 31 | 29,8 |
| II | 8 | 9,9 | 16 | 10,3 | 11 | 10,6 |
| III | 43 | 53,1 | 86 | 55,5 | 61 | 58,7 |
| IV | 4 | 4,9 | 9 | 5,8 | 1 | 1,0 |
| Total | 81 | 100,0 | 155 | 100,0 | 104 | 100,0 |

As maiores discrepâncias ocorreram com o Tema III, quando se observaram 5% a mais de itens para o Teste B em comparação com a Prova Brasil; e com o Tema IV, quando foram observados 4% a mais de itens para a Prova Brasil e para a ANEB em comparação com o Teste B. Cabe uma atenção especial para essas diferenças na cobertura da matriz quanto da análise dos resultados obtidos pelo Teste B e seu distanciamento da Prova Brasil e ANEB.

### 6.3.5 Teste C: estimação das habilidades a partir de 104 itens, da otimização da discriminação e do controle da dificuldade da ANEB

A redução de itens realizada no Teste B associada ao aumento da discriminação tornou o teste mais difícil, ou seja, com uma concentração de itens para posições mais elevadas da escala. Tendo em vista o controle do parâmetro *b*, de forma a aproximar-se da

dificuldade média da Prova Brasil (0,79) e da ANEB (0,71), associado à otimização do parâmetro *a*, propôs-se o Teste C.

O Teste C foi composto por 104 itens, 24 por caderno. Foram excluídos cinco itens de cada um dos 13 blocos do teste da ANEB, procurando-se manter itens para os diversos níveis de dificuldade com atenção especial aos itens de menores parâmetros *b*. Obteve-se com resultado um teste com parâmetro *b* médio de 0,73, indicado na tabela 6.18. A tabela apresenta ainda o parâmetro *a* médio e os resultados da habilidade estimada para o Teste C e os resultados para a Prova Brasil e ANEB.

Tabela 6.18 - Parâmetros psicométricos dos itens e habilidades estimadas - teste de matemática, 8ª série EF, Prova Brasil, ANEB e Teste C.

| Teste | n itens | *a* | | *b* | | Habilidade | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Média | DP | Média | DP | N | Média | DP | Mínimo | Máximo |
| Prova Brasil | 81 | 1,87 | 0,70 | 0,79 | 1,03 | 1.610.073 | -0,1786 | 0,7617 | -2,11 | 2,79 |
| ANEB | 155 | 1,24 | 0,61 | 0,71 | 1,26 | 2.515.731 | -0,3283 | 0,8157 | -2,79 | 2,98 |
| Teste C | 104 | 1,40 | 0,64 | 0,73 | 1,16 | 2.515.731 | -0,3066 | 0,7795 | -2,43 | 2,92 |

Controlando-se o parâmetro *b*, para 104 itens, foi possível obter um parâmetro *a* médio de 1,40, aquém ainda da discriminação da Prova Brasil (1,87), mas superior à da ANEB. A estimativa de habilidade média com base no Teste C foi de -0,31, distante da habilidade estimada para a Prova Brasil (-0,1786), mas um pouco superior à habilidade estimada para a ANEB (-0,3283).

Com relação à variabilidade das estimativas de habilidade, se na ANEB o DP foi de 0,8157, para o Teste C, observou-se DP de 0,7795, mais próximo do DP da Prova Brasil que foi de 0,7617. Isso sugere que a aproximação do número de itens do teste associado ao aumento de sua discriminação propiciou uma maior igualdade do desempenho dos estudantes.

A figura 6.5 apresenta a distribuição das estimativas de habilidade por faixa para a Prova Brasil, a ANEB e o Teste C.

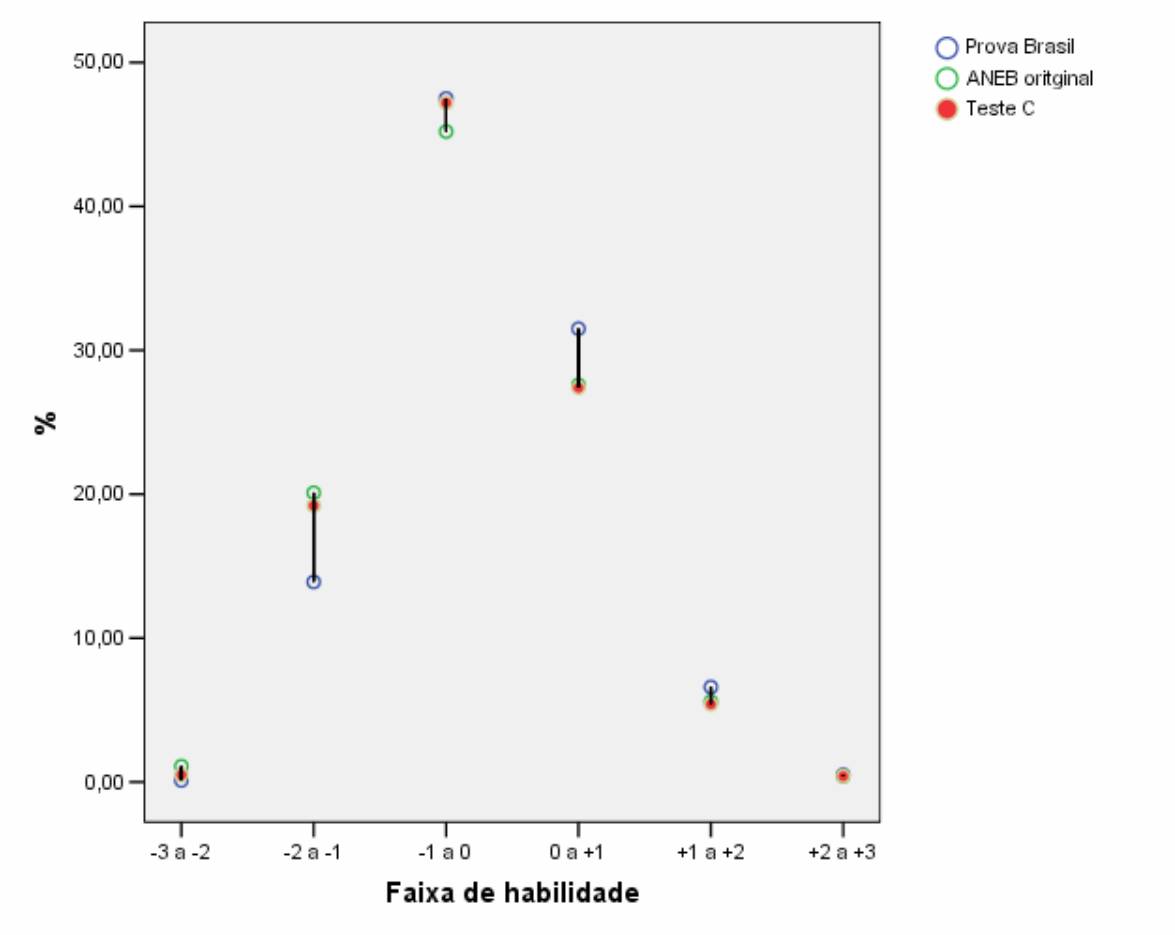


Figura 6.5 - Distâncias entre percentuais de estudantes por faixa de habilidade - matemática, 8ª série EF, Prova Brasil, ANEB e Teste C.

Houve uma aproximação do percentual de estudantes localizados na faixa de -1 a 0 cujas habilidades foram estimadas pelo Teste C, se compararmos com os resultados da Prova Brasil. Um aproximação também foi observada para a faixa de -2 a -1, no entanto, não muito relevante. Para as demais faixas, não foram observadas diferenças entre os percentuais de estudantes cujas habilidades foram estimadas pelo Teste C em comparação com a ANEB. A tabela 6.19 apresenta a distribuição de percentuais de estudantes, quantitativo de itens e parâmetro *a* médio para cada faixa de habilidade, considerando-se o Teste C.

Tabela 6.19 - Percentual de estudantes por faixa de habilidade estimada, número e percentual de itens, média e DP do parâmetro *a* - teste de matemática, 8ª série EF, Teste C.

| Faixa | Habilidade | Itens | | *a* | |
|---|---|---|---|---|---|
| | % | n | % | Média | DP |
| -3 a -2 | 0,50 | 1 | 0,96 | 1,15 | - |
| -2 a -1 | 19,17 | 6 | 5,77 | 0,87 | 0,18 |
| -1 a 0 | 47,18 | 19 | 18,27 | 1,02 | 0,31 |
| 0 a +1 | 27,39 | 34 | 32,69 | 1,32 | 0,61 |
| +1 a +2 | 5,39 | 29 | 27,88 | 1,76 | 0,73 |
| +2 a +3 | 0,37 | 15 | 14,42 | 1,58 | 0,51 |
| Total | 100,0 | 104 | 100 | - | - |

O número de itens para as faixas -3 a -2, -2 a -1, 0 a +1 e +1 a +2 para a Prova Brasil (Ver tabela 6.15) e o Teste C foi semelhante. Para a faixa -1 a 0, o Teste C apresentou mais itens (19) que a Prova Brasil (8). Para todas as faixas de habilidades estimadas, a discriminação da Prova Brasil foi superior à do Teste C. O Teste C se distanciou pouco em termos de percentuais de itens por tema em comparação à Prova Brasil e à ANEB (Tabela 6.20).

Tabela 6.20 - Número e percentual de itens por tema dos testes de matemática, 8ª série EF, para Prova Brasil, ANEB e Teste C.

| Tema | Prova Brasil | | ANEB | | Teste C | |
|---|---|---|---|---|---|---|
| | n itens | % | n itens | % | n itens | % |
| I | 26 | 32,1 | 44 | 28,4 | 32 | 30,8 |
| II | 8 | 9,9 | 16 | 10,3 | 10 | 9,6 |
| III | 43 | 53,1 | 86 | 55,5 | 55 | 52,9 |
| IV | 4 | 4,9 | 9 | 5,8 | 7 | 6,7 |
| Total | 81 | 100,0 | 155 | 100,0 | 104 | 100,0 |

As maiores diferenças por tema observadas não ultrapassaram 3% entre Teste C e Prova Brasil e entre Teste C e ANEB.

### 6.3.6 Teste D: estimação das habilidades a partir de 81 itens e da otimização da discriminação da ANEB

Com base em um teste composto por 81 itens selecionados da ANEB, o Teste D foi constituído e utilizado para estimação das habilidades dos estudantes. Como apresentado no método do presente trabalho, foram selecionados seis a oito itens por bloco com melhores parâmetros *a*.

O Teste D apresentou parâmetro *a* médio de 1,61, um pouco aquém do apresentado pela Prova Brasil (1,87); e parâmetro *b* médio de 0,99, superior à dificuldade da Prova Brasil (0,79) e ANEB (0,71) (Tabela 6.21).

Tabela 6.21 - Parâmetros psicométricos dos itens e habilidades estimadas - teste de matemática, 8ª série EF, Prova Brasil, ANEB e Teste D.

| Teste | n itens | *a* | | *b* | | Habilidade | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Média | DP | Média | DP | N | Média | DP | Mínimo | Máximo |
| Prova Brasil | 81 | 1,87 | 0,70 | 0,79 | 1,03 | 1.610.073 | -0,1786 | 0,7617 | -2,11 | 2,79 |
| ANEB | 155 | 1,24 | 0,61 | 0,71 | 1,26 | 2.515.731 | -0,3283 | 0,8157 | -2,79 | 2,98 |
| Teste D | 81 | 1,61 | 0,62 | 0,99 | 1,17 | 2.515.731 | -0,2708 | 0,7292 | -2,53 | 2,94 |

Sob essa nova configuração, obteve-se habilidade média de -0,2708, superior à ANEB original (-0,3283), mas inferior e distante ainda da Prova Brasil (-0,1786). Se a distância entre as médias da Prova Brasil e ANEB original foi de 0,14, a distância entre as médias da Prova Brasil e o Teste D foi de 0,09.

O DP da habilidade estimada com base no Teste D (0,73) não só foi inferior ao da ANEB (0,82), mas inferior ao da Prova Brasil (0,76). A redução do número de itens associado ao aumento dos parâmetros *a* e *b* médios levou a uma menor variabilidade das estimativas.

A figura 6.6 apresenta a distribuição das estimativas de habilidade por faixa para a Prova Brasil, a ANEB e o Teste D.

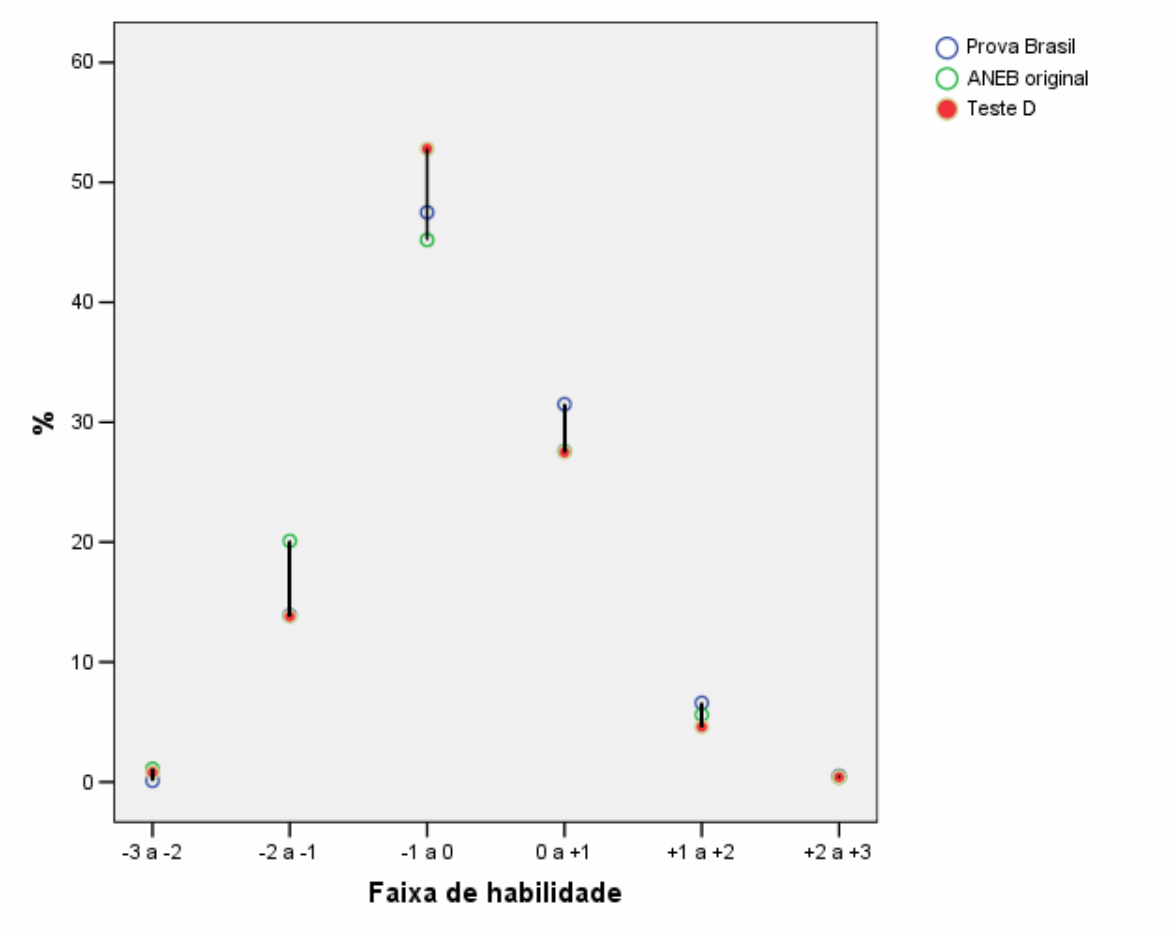


Figura 6.6 - Distâncias entre percentuais de estudantes por faixa de habilidade - matemática, 8ª série EF, Prova Brasil, ANEB e Teste D.

Para a faixa -2 a -1, houve uma aproximação significativa do percentual de estudantes se considerarmos o Teste D e Prova Brasil. Refletindo um pouco a diminuição da variabilidade, observa-se para o Teste D que, para a faixa de -1 a 0, houve um grande aumento do percentual de estudantes comparando-se com a ANEB, extrapolando, inclusive, o percentual obtido nessa faixa para a Prova Brasil. Associado a isso, observou-se uma manutenção do percentual de estudantes com estimativas de habilidade localizadas na faixa de 0 a +1 comparando Teste D e ANEB. O percentual de estimativas foi inferior ao da Prova Brasil.

O número de itens total utilizado para estimar as habilidades da Prova Brasil e do Teste D foi o mesmo. Para as faixas 0 a +1 e +1 e +2, observou-se um maior número de itens da Prova Brasil (61) em comparação ao Teste D (49) (Tabelas 6.15 e 6.22). Para a faixa de +2 a +3, o número de itens do Teste D (17) foi bem superior ao da Prova Brasil (6).

Tabela 6.22 - Percentual de estudantes por faixa de habilidade estimada, número e percentual de itens, média e DP do parâmetro *a* - teste de matemática, 8ª série EF, Teste D.

| Nível | Habilidade | Itens | | *a* | |
|---|---|---|---|---|---|
| | % | n | % | Média | DP |
| -3 a -2 | 0,82 | 2 | 2,47 | 1,10 | 0,07 |
| -2 a -1 | 13,82 | 3 | 3,70 | 1,03 | 0,02 |
| -1 a 0 | 52,83 | 10 | 12,35 | 1,19 | 0,24 |
| 0 a +1 | 27,48 | 23 | 28,40 | 1,58 | 0,68 |
| +1 a +2 | 4,63 | 26 | 32,10 | 1,91 | 0,67 |
| +2 a +3 | 0,43 | 17 | 20,99 | 1,58 | 0,47 |
| Total | 100,0 | 81 | 100 | - | - |

Com exceção da faixa -3 a -2, todas as outras apresentaram itens com parâmetro *a* médio superior para a Prova Brasil. Mesmo com todos os esforços na tentativa de tornar Teste D e Prova Brasil similares, não foi possível, já que o número de itens com um bom grau de discriminação foi superior para a Prova Brasil. O Teste D não se distanciou muito em termos de percentuais de itens por tema em comparação à Prova Brasil e à ANEB (Tabela 6.23).

Tabela 6.23 - Número e percentual de itens por tema dos testes de matemática, 8ª série EF, para Prova Brasil, ANEB e Teste D.

| Tema | Prova Brasil | | ANEB | | Teste D | |
|---|---|---|---|---|---|---|
| | n itens | % | n itens | % | n itens | % |
| I | 26 | 32,1 | 44 | 28,4 | 27 | 33,3 |
| II | 8 | 9,9 | 16 | 10,3 | 7 | 8,6 |
| III | 43 | 53,1 | 86 | 55,5 | 46 | 56,8 |
| IV | 4 | 4,9 | 9 | 5,8 | 1 | 1,2 |
| Total | 81 | 100,0 | 155 | 100,0 | 81 | 100,0 |

As maiores discrepâncias do Teste D com a Prova Brasil referem-se aos temas III e IV (4%). Cabe uma atenção especial para essas diferenças na cobertura da matriz quanto da análise dos resultados obtidos pelo Teste D e sua aproximação da Prova Brasil.

## 6.4 Estudo 4: Comparação entre as estimativas de habilidade dos estudantes para Prova Brasil, ANEB e Testes A a D e sua associação com as características dos testes

A tabela 6.24 apresenta estatísticas de tendência central e de variabilidade dos parâmetros *a* e *b* e das estimativas de habilidade dos estudantes com base em todos os testes envolvidos.

Tabela 6.24 - Parâmetros psicométricos dos itens e habilidades estimadas - teste de matemática, 8ª série EF, Prova Brasil, ANEB original, Testes A a D.

| Teste | n itens | *a* | | *b* | | Habilidade | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Média | DP | Média | DP | N | Média | DP | Mínimo | Máximo |
| Prova Brasil | 81 | 1,87 | 0,70 | 0,79 | 1,03 | 1.610.073 | -0,1786 | 0,7617 | -2,11 | 2,79 |
| ANEB | 155 | 1,24 | 0,61 | 0,71 | 1,26 | 2.515.731 | -0,3283 | 0,8157 | -2,79 | 2,98 |
| Teste A | 104 | 1,25 | 0,60 | 0,71 | 1,06 | 2.515.731 | -0,3072 | 0,7855 | -2,17 | 2,93 |
| Teste B | 104 | 1,46 | 0,62 | 0,97 | 1,10 | 2.515.731 | -0,2906 | 0,7506 | -2,59 | 2,97 |
| Teste C | 104 | 1,40 | 0,64 | 0,73 | 1,16 | 2.515.731 | -0,3066 | 0,7795 | -2,43 | 2,92 |
| Teste D | 81 | 1,61 | 0,62 | 0,99 | 1,17 | 2.515.731 | -0,2708 | 0,7292 | -2,53 | 2,94 |

Em síntese, O Teste A foi construído para apresentar estatísticas semelhantes à ANEB com um número de itens (104) mais aproximado ao da Prova Brasil (81) em que cada estudante responde a 24 itens como se dá aproximadamente na Prova Brasil. O impacto da modificação nas estimativas médias de habilidades foi pequeno (de -0,33 a -0,31), mas na direção da Prova Brasil (-0,18).

Também com a redução para 104 itens, um aumento da média do parâmetro *a* em comparação à ANEB propiciado pelo Teste B (de 1,24 para 1,46), mas associado a um aumento do parâmetro *b*, promoveu um aumento das estimativas médias de habilidade (-0,29) e uma aproximação maior da Prova Brasil que a encontrada pelo Teste A.

Com 104 itens, o Teste C foi constituído de forma a controlar o parâmetro *b*, elevando o parâmetro *a* ao máximo possível (1,40 contra 1,87 da Prova Brasil). A estimativa média de habilidade também subiu em comparação à ANEB (-0,31 contra -0,33), mas ficou aquém da aproximação da estimativa média dos estudantes para a Prova Brasil apresentada pelo Teste B.

O Teste D, composto por 81 itens, não permitiu que cada estudante respondesse ao mesmo número de itens que cada estudante da Prova Brasil, o que pode ter prejudicado a fidedignidade da estimativa individual. No entanto, permitiu que fosse considerado o mesmo número total de itens da Prova Brasil (81). Com a exclusão dos itens menos discriminativos por bloco, obteve-se o parâmetro *a* máximo permitido (1,61), o maior valor de discriminação de todos os testes propostos, mas ainda aquém do parâmetro *a* médio da Prova Brasil (1,87). Sob essa configuração, obteve-se o maior valor médio de estimativa de habilidade (-0,27), comparando-se com os Testes A a C, abaixo ainda em 0,09 DP da média da Prova Brasil (-0,18).

A análise dos resultados por médias, embora não permita explicar a variabilidade dos resultados por faixa de habilidade, fornece informações relevantes. Sistematicamente a redução do número de itens, com ou sem a variação do parâmetro *a* e *b*, propiciou um aumento das médias das estimativas de habilidade em comparação à ANEB. A figura 6.7 representa a dispersão dos testes em função do número de itens (155, 104, 81) e habilidade estimada média dos estudantes.

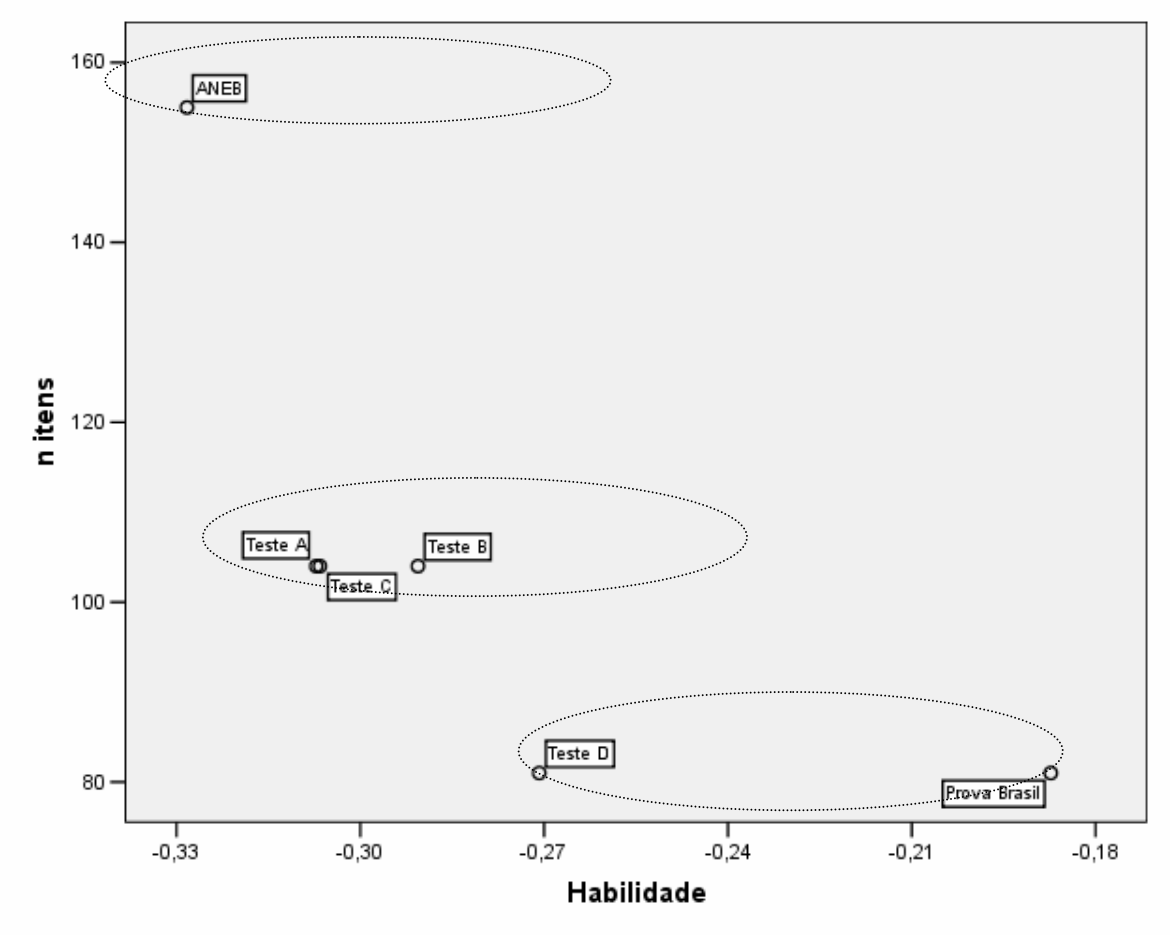


Figura 6.7 - Gráfico de dispersão entre número de itens no teste e habilidade estimada média - matemática, 8ª série EF, Prova Brasil, ANEB, Testes A a D.

Observam-se três grupos de estimativas de habilidade. A ANEB, com maior número de itens (155) apresentou a maior estimativa de habilidade média. Os Testes A, B, C (104 itens), apresentaram estimativas de habilidade superiores à da ANEB e próximas entre si. O Teste D e a Prova Brasil (81 itens) apresentaram estimativas de habilidade superiores à ANEB e aos Testes A a C. Nesse caso, os resultados de habilidades estimadas para a Prova Brasil foram superiores aos do Teste D. Os resultados sugerem que o número de itens dos testes como um todo está inversamente associado à média das habilidades estimadas.

Após a simulação dos testes, o aumento das estimativas médias de habilidade foi mais evidente para os Testes B e D, exatamente os testes com maior poder discriminativo. A figura 6.8 ilustra que quanto maior o parâmetro *a* médio do teste, maior a habilidade estimada para os estudantes.

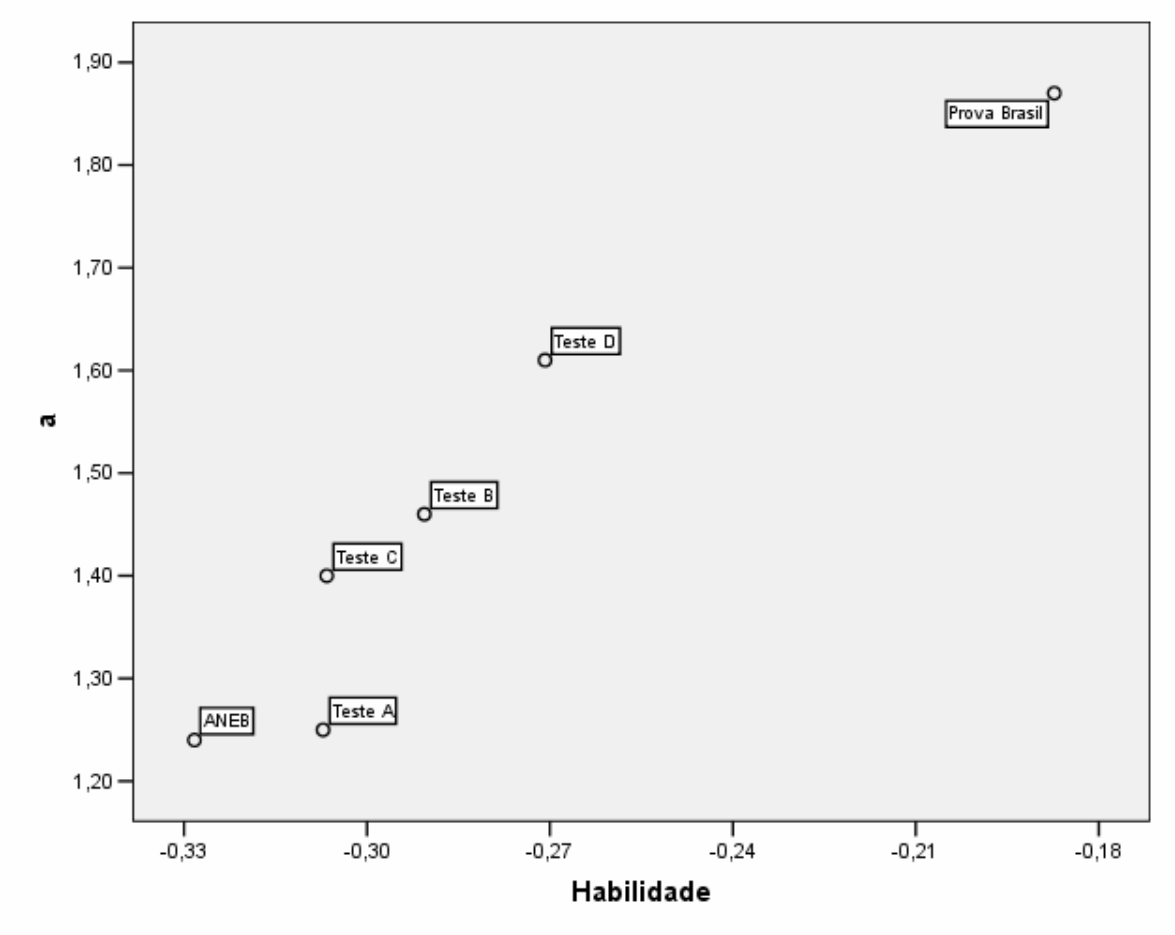


Figura 6.8 - Gráfico de dispersão entre parâmetro *a* médio e habilidade estimada média - matemática, 8ª série EF, Prova Brasil, ANEB, Testes A a D.

A redução do número de itens promoveu estimativas de habilidade com DP inferiores ao encontrado para a ANEB. Se na ANEB (155 itens), observou-se DP de 0,82, para os Testes A, B e C (104 itens), obtiveram-se DP variando entre 0,75 e 0,79. Para o Teste D (81 itens), o DP foi menor ainda (0,73). A Prova Brasil, também com 81 itens, apresentou DP de 0,76 com relação às estimativas de habilidade.

Uma análise mais superficial indica que o tamanho do teste pode ter certa influência na variabilidade das estimativas. No entanto, se são considerados os critérios utilizados para exclusão de itens, geralmente foram retirados da análise os itens menos discriminativos. Esses estavam mais concentrados nas faixas inferiores da escala e, quando foram excluídos, levaram a uma concentração de itens nas faixas médias, como na Prova Brasil. A variabilidade das estimativas de habilidade dos quatro testes (A a D) se aproximou da variabilidade da Prova Brasil. Esses achados indicam que a variabilidade das estimativas está associada à distribuição de itens pelas faixas da escala.

Com o objetivo de entender como se deu a variabilidade das estimativas de habilidade, a figura 6.9 apresenta os percentuais de estudantes por faixa de habilidade estimada tendo por base a Prova Brasil, a ANEB e os Testes A a D.

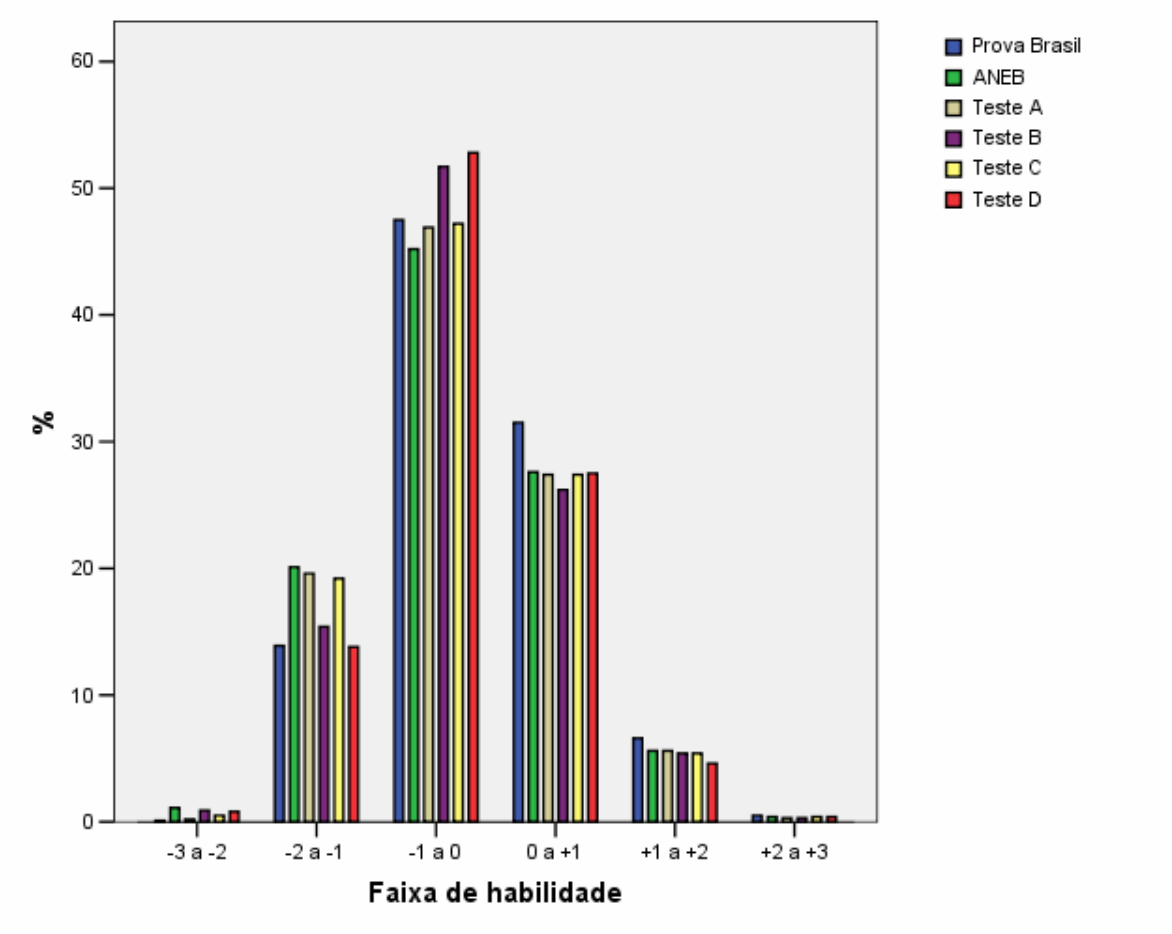


Figura 6.9 - Percentuais de estudantes por faixa de estimativas de habilidade - matemática, 8ª série EF, Prova Brasil, ANEB, Teste A a D.

Os Testes B e D levaram a estimativas de habilidade semelhantes para todas as faixas. Independentemente se os estudantes responderam a 104 ou a 81 itens no total, se responderam a 24 ou a 18 itens, observou-se distribuição de percentual semelhante para as faixas. A mesma análise pode ser realizada comparando a distribuição de estudantes pelas faixas para os Testes A e C. Que aspectos diferenciam os Testes B e D dos Testes A e C? Os Testes B e D apresentaram parâmetro *a* médios (1,46 e 1,61) maiores que os dos Testes A e C (1,25 e 1,40). Que aspectos assemelham os Testes A e C e os Testes B e D? Os Testes A e C apresentaram parâmetro *b* médio (0,71 e 0,76) menores que os dos Testes B e D (0,97 e 99).

A exclusão dos itens menos discriminativos sem controle da dificuldade para os Testes B e D retiraram dos testes os itens com parâmetro *b* localizados nas faixas inferiores da escala. O parâmetro *b* médio subiu. Quando se controlou a dificuldade, não se obteve um parâmetro *a* tão alto para os Testes A e C. Essa configuração explica o comportamento semelhante entre Testes A e C e Testes B e D.

A Tabela 6.25 mostra a distribuição de percentuais de itens por faixa de parâmetro *b* ou de estimativas de habilidade.

Tabela 6.25 - Percentual de itens por faixa de habilidades estimadas - teste de matemática, 8ª série EF, Prova Brasil, ANEB, Testes A a D.

| Teste | Habilidade | | | | | | Total |
|---|---|---|---|---|---|---|---|
| | -3 a -2 | -2 a -1 | -1 a 0 | 0 a +1 | +1 a +2 | +2 a +3 | |
| Prova Brasil | 0,0 | 7,4 | 9,9 | 38,3 | 37,0 | 7,4 | 100,0 |
| ANEB | 3,2 | 7,1 | 15,5 | 30,3 | 29,0 | 14,8 | 100,0 |
| Teste A | 0,0 | 5,8 | 19,2 | 33,7 | 28,8 | 12,5 | 100,0 |
| Teste B | 1,9 | 3,8 | 11,5 | 32,7 | 31,7 | 18,3 | 100,0 |
| Teste C | 1,0 | 5,8 | 18,3 | 32,7 | 27,9 | 14,4 | 100,0 |
| Teste D | 2,5 | 3,7 | 12,3 | 28,4 | 32,1 | 21,0 | 100,0 |

Observa-se uma maior variabilidade dos itens em termos de parâmetro *b* ou pelas faixas de estimativas de habilidade da ANEB em comparação com a Prova Brasil. Na Prova Brasil, as faixas extremas de estimativas de habilidade apresentaram poucos itens que as representassem. Já a ANEB apresentou uma boa variabilidade através das faixas, incluindo itens para as faixas extremas, situação apropriada para discriminar estudantes aí localizados. Ambos os testes apresentaram uma maior concentração de itens com parâmetro *b* associados às faixas da direita da escala. Sem considerar ainda o grau de discriminação dos itens, pode-se inferir que a ANEB e a Prova Brasil discriminam melhor os estudantes localizados da faixa central à superior, já que exploram melhor essas faixas ao incluir um razoável percentual de itens com parâmetro *b* nas faixas central e superior.

Uma maior variabilidade entre os percentuais de itens por faixa da ANEB em comparação à Prova Brasil é refletida em uma concentração de itens nas faixas 0 a +1 e +1 a +2 para a Prova Brasil (75%) em comparação à ANEB (59%). Os itens da Prova Brasil são mais concentrados nas faixas de habilidades estimadas centrais da escala.

Os Testes A a D apresentaram características razoavelmente semelhantes entre si. Observaram-se itens representativos das diversas faixas de habilidade, com uma tendência às posições moderadas a altas da escala. Desses testes, os que mais se aproximaram da ANEB, teste a partir do qual foram simulados, em termos de distribuição de itens pelas faixas foram o A e o C. Ressalta-se que a distribuição de estudantes pelas faixas para a ANEB foi semelhante às distribuições encontradas para os Testes A e C.

Os Testes que mais se aproximaram da Prova Brasil, em termos de distribuição de itens pelas faixas de habilidades estimadas, foram o B e o D, mas sem muita similaridade, já que apresentaram (a) um percentual maior de itens no extremo superior da escala (18,3 para o Teste B e 21,0 para o Teste D) que a Prova Brasil; e (b) maior variabilidade que a Prova Brasil. Em termos de distribuição dos itens pelas faixas, de acordo com os critérios adotados para a simulação dos testes, não foi possível construir testes completamente equiparáveis ao da Prova Brasil em termos da distribuição de itens pelas faixas de habilidades estimadas.

As médias do parâmetro *a* foram calculadas para cada faixa de habilidades estimadas, apresentadas na Tabela 6.26.

Tabela 6.26 - Parâmetro *a* médio por faixa de habilidades estimadas - teste de matemática, 8ª série EF, Prova Brasil, ANEB, Testes A a D.

| Teste | Habilidade | | | | | |
|---|---|---|---|---|---|---|
| | -3 a -2 | -2 a -1 | -1 a 0 | 0 a +1 | +1 a +2 | +2 a +3 |
| Prova Brasil | - | 1,3 | 1,5 | 1,6 | 2,3 | 1,8 |
| ANEB | 0,7 | 0,9 | 1,0 | 1,3 | 1,4 | 1,4 |
| Teste A | - | 0,8 | 0,9 | 1,2 | 1,6 | 1,4 |
| Teste B | 1,1 | 1,0 | 1,1 | 1,4 | 1,7 | 1,5 |
| Teste C | 1,2 | 0,9 | 1,0 | 1,3 | 1,8 | 1,6 |
| Teste D | 1,1 | 1,0 | 1,2 | 1,6 | 1,9 | 1,6 |

Todos os testes são mais discriminativos para metade superior da escala de estimativas de habilidade. Essa constatação pode estar associada ao fato de estarmos tratando de resultados de 4ª , 8ª EF e 3ª EM inseridos na mesma escala.

A distribuição do parâmetro *a* médio da Prova Brasil é excelente, variando de 1,3 a 2,3 para as faixas de estimativas de habilidade. Sofre a influência de uma distribuição mais

concentrada que a da ANEB. Os maiores valores médios de parâmetros *a* para as faixas da ANEB (1,4) são iguais aos menores valores para a Prova Brasil (1,3 e 1,5). A Prova Brasil é muito mais discriminativa que o teste da ANEB para todas as faixas, com exceção da faixa extrema negativa.

Os Testes B e D foram os que mais se aproximaram da discriminação da Prova Brasil, mesmo assim ficaram aquém. No entanto, esses testes promoveram a melhor discriminação possível com o número de itens disponíveis (104 e 81). Os testes são muito discriminativos para as faixas superiores da escala. Uma atenção especial merece ser dada para o Teste D. Apresenta bons resultados de distribuição de itens pelas faixas de habilidades estimadas e de médias de parâmetro *a*. Para todos os testes, as faixas com maiores percentuais de itens apresentaram também os maiores parâmetro *a* médios.

As médias ponderadas dos valores de EPM estimados por estudante apresentam-se a tabela 6.27.

Tabela 6.27 - Erro-padrão de mensuração médio ponderado pelo número de estimativas de habilidade - teste de matemática, 8ª série EF, Prova Brasil, ANEB, Testes A a D.

| Teste | EPM |
|---|---|
| Prova Brasil | 0,48 |
| ANEB | 0,42 |
| Teste A | 0,48 |
| Teste B | 0,51 |
| Teste C | 0,47 |
| Teste D | 0,55 |

O EPM funciona como um índice de fidedignidade dos testes, de forma que, quanto maior o valor, menor a fidedignidade. A ANEB apresentou o menor EPM médio (0,42). O Teste D apresentou o menor índice de fidedignidade (EPM médio de 0,55). Observou-se relação inversa entre o tamanho do erro e o número de questões contidas em cada um dos cadernos de teste (aproximadamente, 39 para a ANEB, 24 para os testes A, B e C e 18 para o Teste D), como pode ser verificado na Figura 6.10.

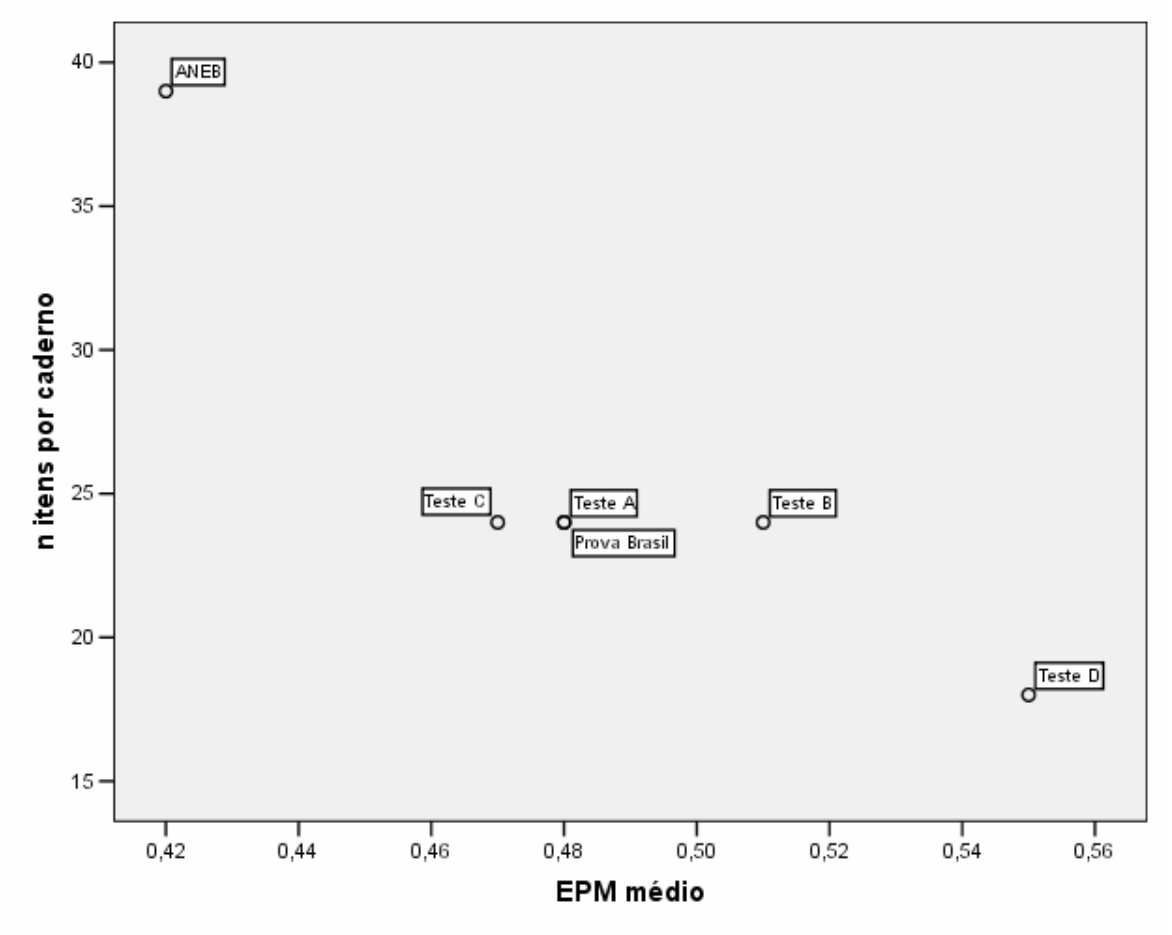


Figura 6.10 - Percentuais de estudantes por faixa de estimativas de habilidade - matemática, 8ª série EF, Prova Brasil, ANEB, Teste A a D.

Identificam-se efetivamente três grupos, de sorte que quanto maior o número de itens em cada caderno de teste, maior a fidedignidade, pois menor é o EPM médio. A distribuição do EPM médio por faixa de habilidade estimada é apresentada na figura 6.11.

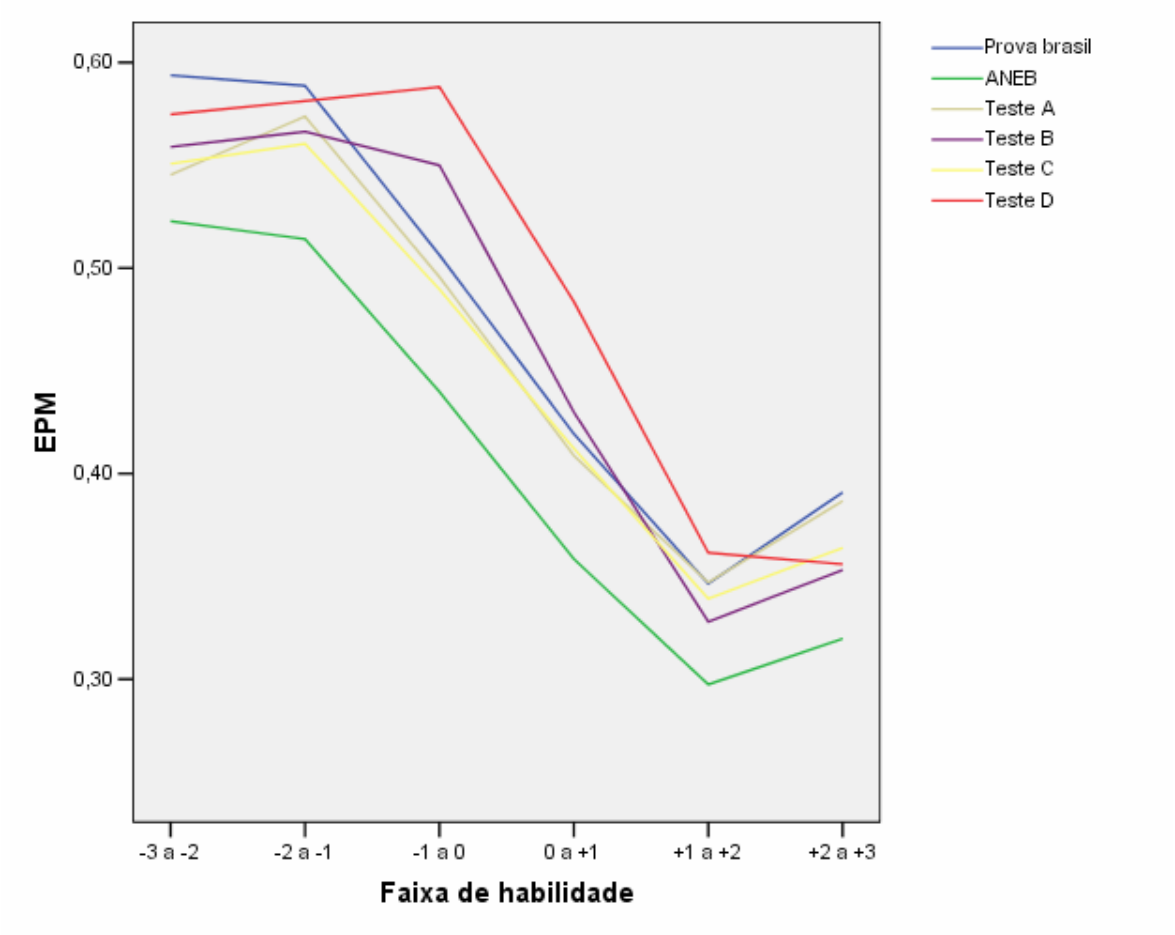


Figura 6.11 - EPM médio por faixa de habilidade estimada - matemática, 8ª série EF, Prova Brasil, ANEB, Teste A a D.

Os resultados indicam que todos os testes apresentam resultados mais fidedignos quando avaliam estudantes localizados na faixa de habilidade estimada de +1 a +2. Além disso, que o EPM vai aumentando para as faixas de habilidades menores. O Teste ANEB apresenta os menores EPM médios para todas as faixas e o Teste D, os maiores, confirmando os resultados do índice de fidedignidade apresentados anteriormente. Os resultados indicam, novamente, que o EPM médio por faixa está associado ao número de itens por caderno utilizados para estimar as habilidades.

A função de informação do teste é inversa ao EPM. Para o presente estudo, a informação dos itens foi calculada pontualmente pelo inverso da média de EPM por faixa de habilidade de 1DP. A figura 6.12 apresenta os resultados de informação para cada uma das faixas de estimativas de habilidade e para todos os Testes. Sabe-se que o gráfico mais apropriado para representar valores pontuais para as faixas não é o de linhas. No entanto, considerou-se visualmente clara a sua utilização.

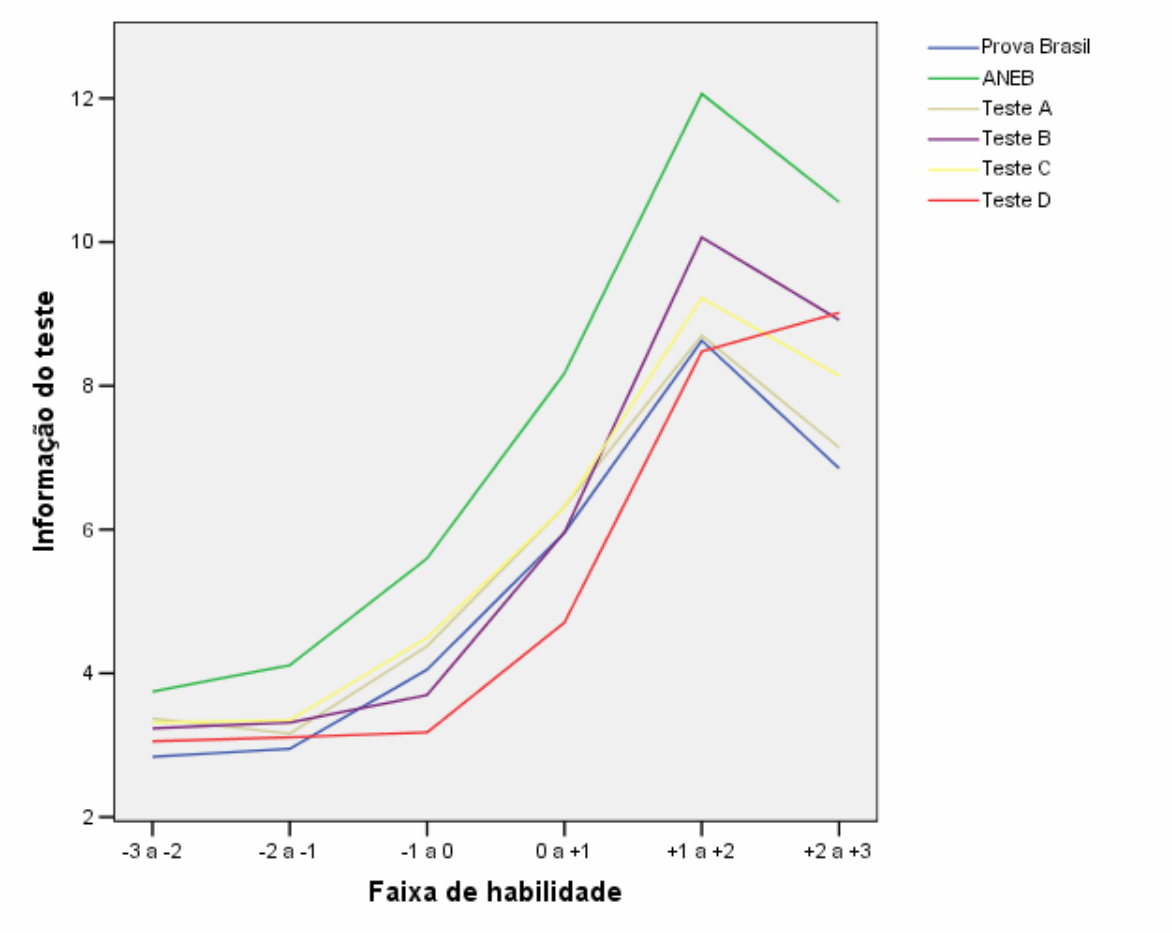


Figura 6.12 - Informação por faixa de habilidade estimada - matemática, 8ª série EF, Prova Brasil, ANEB, Teste A a D.

Todos os testes são mais informativos para a faixa de estimativas de habilidade +1 a +2, sendo a ANEB o mais informativo e o Teste D o menos informativo. O Teste D atingiu seu maior nível de informação na faixa de +2 a +3, em que 23% de seus itens estavam aí localizados com parâmetro *a* médio muito alto.

A informação do teste parece estar associada diretamente (a) ao número de itens total no teste; (b) ao número de itens em cada caderno; e (c) ao índice de discriminação dos itens. Essa última constatação remete a observações anteriores que os itens mais discriminativos estavam localizados na parte superior da escala de habilidades estimadas.

## 7. Discussão

O estudo para verificação da relação entre características do teste e a validade e a fidedignidade das estimativas de habilidade é relevante ao contexto atual do SAEB. Mesmo que os testes tenham se pautado na mesma matriz de referência e na mesma estrutura de itens, em 2005 e 2007, o sistema de avaliação foi modificado em alguns aspectos referentes ao seu delineamento. A expansão do público avaliado foi acompanhada da mudança do tamanho do teste e do número de itens e de disciplinas que cada estudante respondeu entre 2003 e 2005 e, novamente, de 2005 para 2007.

Qual a limitação da TRI em fornecer estimativas de habilidade independentemente do teste utilizado (Baker, 2001; Hambleton, Swaminathan & Rogers, 1991)? De que forma características de tamanho do teste, discriminação, dificuldade e poder de informação estão associados às estimativas de habilidade? O presente trabalho buscou evidências para esclarecer tais questões.

Tais temas não podem ser abordados sem considerar os fatores que extrapolam o âmbito do teste e que estão associados com a avaliação. Se o grau de validade e de fidedignidade dos resultados da avaliação sofre influência de um conjunto de fatores externos ao teste e se está interessado na parcela da variância explicada pelos testes, cabe o controle dessas variáveis. Identificaram-se na literatura os seguintes fatores que extrapolam o teste, mas que tem efeito na variância da habilidade estimada: número de estudantes avaliados, procedimentos de aplicação, método de equalização, influências motivacionais e tempo de aplicação. Citam-se alguns aspectos referentes ao contexto da avaliação e que podem influenciar seu grau de validade e de fidedignidade: mobilização de agentes educacionais para a execução da avaliação e a previsão de utilidade de seus resultados; fatores motivacionais que influenciam no empenho dos estudantes em responder aos itens; público avaliado.

O SAEB sempre contou com o apoio das Secretarias de Estado da Educação para conferência da amostra, treinamento dos aplicadores, estocagem, conferência e distribuição de materiais e administração dos testes. Muitos dos recursos humanos envolvidos na aplicação dos testes foram oriundos das Secretarias. Isso sempre tornou o trabalho do INEP próximo às Secretarias que, com apoio maior ou menor, viabilizavam o trabalho de campo. Em todos os ciclos do SAEB, houve divulgação, em maior ou menor grau aos agentes educacionais e à população.

No entanto, por vezes, observam-se questionamentos e críticas de agentes educacionais quanto à pertinência da avaliação, principalmente sob o formato

exclusivamente amostral (de 1995 a 2003), e quanto às estratégias de divulgação. A aparente pouca utilidade direta dos resultados pela escola pode ter trazido na história do SAEB impacto no empenho de agentes para a divulgação e a administração do teste nas escolas. A realização de estudos específicos, sobre a percepção da pertinência do SAEB pelos diversos agentes educacionais é necessária e pode esclarecer melhor o tema, de extrema relevância para a qualidade dos resultados, já que está relacionado ao contexto que influi significativamente na validade (AERA, APA & NCME, 1999) e na Fidedignidade (Cohen & Swerdlik, 2002; Urbina, 2007).

A divisão do SAEB em ANEB e Prova Brasil modificou o contexto no que se refere à divulgação dos resultados. Se os resultados eram apresentados exclusivamente por estrato (unidades da federação, rede, dependência administrativa, etc.), passaram a ser apresentados para cada unidade escolar pública e urbana. A criação e divulgação, nos últimos anos, do Índice de Desenvolvimento da Educação Básica (IDEB) (Fernandes, 2007), que tem como um de seus componentes os resultados de desempenho dos estudantes na Prova Brasil, vêm tornando os resultados dessa avaliação mais próximos da escola. Isso porque cada escola pública e urbana possui um índice geral comparável com as demais, e com detalhamento de indicadores, sobre os quais podem atuar para melhoria de seus resultados de qualidade e de fatores associados. O IDEB superou, em termos de relevância e proximidade para a escola, os resultados do Índice de Qualidade da Educação Fundamental (IQE) (Araújo, Condé & Luzio, 2004) como relatado no estudo de Condé (2007b), já que o IQE tinha por base os resultados de desempenho dos estudantes por estrato do SAEB.

No entanto, o IDEB foi apresentado à sociedade apenas em 2007. Na história do SAEB, é possível que tenha havido, no decorrer dos ciclos, variação no empenho dos agentes educacionais e, conseqüentemente, na validade e na fidedignidade dos resultados. Alterações substanciais no empenho e no apoio à avaliação gerados pelo IDEB, caso tenham ocorrido, só fazem sentido a partir do SAEB 2007. Não há indícios que pode ter havido diferença significativa na motivação dos estudantes em função de alguma interferência escolar significativa para o SAEB 2005. Isso torna não plausível a hipótese que os examinandos da Prova Brasil 2005 tenham um desempenho maior em função das características de incentivo e de motivação diferentes propiciados pelos agentes educacionais.

Como para o presente estudo, foram utilizados os resultados dos testes de matemática 8ª série EF da ANEB 2005 e da Prova Brasil, a discussão a seguir é focada

apenas nesse escopo. Os fatores relacionados ao contexto em que a testagem ocorre e aos testandos foram bastante semelhantes entre as avaliações. Os estudantes de escolas públicas e urbanas avaliados pela ANEB apresentam características semelhantes ao grupo de estudantes avaliado pela Prova Brasil. Questionou-se se a grande diferença no número de pessoas entre os grupos gerou influências sobre as estimativas de habilidade entre as avaliações, propiciando resultados significativamente diferentes para matemática 8ª série EF. Os estudos de Condé (2007) e de Rabello (2007) observaram praticamente os mesmos resultados de estimativas de habilidade médias entre ANEB e Prova Brasil para língua portuguesa 8ª série EF. Como o teste de matemática foi aplicado juntamente com o de língua portuguesa para a série e apenas os resultados de matemática apresentaram diferenças significativas, considera-se que o número de casos utilizados para estimar as habilidades não tenha influenciado substancialmente na diferença entre os resultados.

Ambas as avaliações forneceram fatores motivadores semelhantes aos estudantes que os responderam. Não foram encontradas nos relatórios do SAEB informações que indicassem que os estudantes da Prova Brasil 2005 receberam estímulo motivador diferenciado que os fizessem responder com mais afinco a Prova Brasil, de forma que essas apresentassem resultados mais fidedignos e melhores resultados de desempenho. O mesmo se pode dizer com relação à ANEB 2005. Ações de gestão visando divulgar e incentivar a avaliação na escola têm bastante impacto no desempenho dos estudantes. Em 2005, no entanto, parecem ter ocorrido similarmente entre as avaliações.

Aspectos inerentes à administração dos testes impactam tanto na fidedignidade (Cohen & Swerdlik, 2002; Cronbach, 1996; Urbina, 2007), quanto na validade dos resultados da avaliação (Oshima, 1994; Bolt, Cohen e Wollack, 2002; Sireci, 2005; Sireci, Scarpati e Li, 2005; Lu e Sireci, 2007). Questionou-se até que ponto aspectos relacionados à aplicação, que inclui instruções, tempo disponível aos respondentes associado ao tamanho do teste, influência da velocidade em teste de potência no cansaço e na motivação (Oshima, 1994) influenciaram significativamente na validade e na fidedignidade dos resultados da ANEB e da Prova Brasil. Indagou-se até que ponto a administração de 48 itens (24 de matemática, 24 de língua portuguesa) para cada estudante da Prova Brasil, número superior ao teste da ANEB em que os estudantes responderam a 39 itens, pode ter gerado cansaço, desmotivação ou mesmo o tempo não tenha sido suficiente para os respondentes da Prova Brasil especificamente no caso de matemática 8ª série EF. Não se têm evidências para acreditar que os resultados de diferenças entre ANEB e Prova Brasil tenham ocorrido em função da falta de motivação e do cansaço associados ao número de

itens aplicados, já que para língua portuguesa 8ª série EF as avaliações apresentaram resultados idênticos.

Os procedimentos de treinamento dos aplicadores, as instruções de aplicação, de tempo disponibilizado para resposta a cada item (2,31 minutos por item para a ANEB e 2 minutos para a Prova Brasil) e a distribuição de testes para os estudantes não se diferiram substancialmente entre as avaliações. Assim, não devem ter impactado diferentemente entre as avaliações. Por se tratar avaliações em larga escala, sabe-se da dificuldade de se garantir a padronização da aplicação, de forma que os aplicadores transmitam as instruções uniformemente, distribuam os cadernos da forma programada, entre outros. Principalmente no caso da Prova Brasil, cujo campo foi significativamente maior ao da ANEB, problemas de padronização podem ter ocorrido. Utiliza-se aqui novamente o argumento da proximidade entre os resultados da ANEB e da Prova Brasil para a língua portuguesa 8ª série EF para se fazer a inferência que o trabalho de campo não foi fator significativo para gerar impacto na validade e na fidedignidade das estimativas de habilidade. O tema merece outros estudos.

Apresenta-se como hipótese que, como a Prova Brasil e o IDEB ganharam grande projeção nacional, impactando nos diversos setores educacionais, a partir de 2007, o grau de validade e de fidedignidade das estimativas de habilidade irá aumentar para os próximos ciclos. Assim, se houver relação de fidedignidade com magnitude do parâmetro de habilidade, o desempenho na Prova Brasil irá melhorar. A Teoria G (Brennan, 1983; Cronbach, Gleser, Rajaratnam & Nanda, 1972) pode contribuir com essa investigação, a partir da decomposição o erro em componentes para identificar sua fontes.

No que se refere aos procedimentos de análise, questionou-se até que ponto a calibração dos itens realizada para a Prova Brasil, a partir de respostas dos estudantes de escolas públicas urbanas, podem ter se diferido da calibração realizada para a ANEB, que tem por base as respostas dos estudantes de escolas particulares e públicas, rurais e urbanas. Embora os procedimentos utilizados para estimação dos parâmetros dos itens tenham sido praticamente os mesmos, é fundamental a realização de outros estudos para verificar o impacto da diferença das características dos respondentes para a calibração. Sugere-se a comparação dos parâmetros dos itens comuns da Prova Brasil 2005 com o SAEB 2003. Considerou-se, para o presente estudo, que o impacto não tenha sido substantivo.

Os procedimentos analíticos de consideração da não-resposta dos estudantes para os últimos itens dos blocos foram semelhantes entre as avaliações. Foram considerados não-

apresentados (Bock & Zimowski, 1995) para a estimação das habilidades da ANEB e da Prova Brasil.

O delineamento da ANEB e da Prova Brasil apresentam semelhanças entre si, que possibilita o estudo sobre a relação dos testes com as estimativas de habilidade. Permite-nos realizar uma série de análises comparativas, já que utilizam a mesma escala para parâmetros de itens e de habilidade, mesmo modelo TRI, mesma estrutura de questão de teste construídos sobre uma mesma matriz de referência.

Urbina (2007) considerou que os procedimentos de validação de testes são defendidos com argumentos lógicos e relações demonstráveis entre o conteúdo do teste e o construto que esse pretende representar. Para ambos os testes, encontraram-se evidências que contribuíram para a validade dos resultados, já que os itens foram elaborados e revisados por especialistas da área de matemática, capacitados em técnicas de construção de itens, preocupados de garantir a convergência entre dos itens com a matriz de referência. Não bastasse a análise realizada pelo INEP para as etapas de elaboração e de revisão de itens, que ocorreram anteriormente à composição do teste definitivo, após a administração dos testes, promoveu-se nova análise pedagógica. Inclusive, verificou-se que alguns itens foram excluídos por motivos pedagógicos e não foram incluídos na análise psicométrica.

Também com relação à validade baseada no conteúdo do teste, os testes de matemática 8ª série EF da ANEB e da Prova Brasil não se diferiram em termos das expressões utilizadas nas tarefas ou mesmo em termos do formato dos itens. As orientações e os procedimentos de elaboração e de revisão técnico-pedagógicas foram os mesmos. Assim, foram utilizados itens com idêntica estrutura teórica e formato de enunciado e com quatro opções de respostas para uma possibilidade de resposta correta. Não se pode afirmar que as expressões utilizadas para a elaboração dos itens, ou seu formato, tenham prejudicado individualmente a validade dos resultados da ANEB ou da Prova Brasil e propiciado diferenças entre os resultados de estimativas de habilidade.

Cabe questionar se a validade dos resultados foi influenciada em função da falta de familiaridade dos estudantes com relação ao formato de múltipla escolha das questões. Os achados do presente estudo não são capazes de permitir inferências sobre o tema e isso remeteria à discussão da validade do SAEB como um todo, desde sua primeira aplicação, em 1995, sob a atual estrutura.

No que tange ao tipo de evidência de validade baseada no conteúdo do teste (AERA, APA & NCME, 1999), em seus aspectos associados aos temas avaliados, à

cobertura ou ao alinhamento (Herman, Webb e Zuniga, 2002; Bhola, Impara e Buchendahl, 2003), analisou-se o percentual de itens por tema e por descritor para os testes. Cada um deles respeitou, na medida do possível, o esquema de prioridades previsto na matriz de referência, considerando-se o percentual de itens por tema (espaço e forma, grandezas e medidas, números e operações/ álgebra e funções e tratamento da informação). Essa é uma evidência que contribuiu para o grau de validade dos resultados de ambos os testes, já que o planejamento estrutural previsto para o teste no *framework*, também fruto de discussões entre especialistas em matemática e pedagogia, foi cumprido. Para a Prova Brasil, alguns descritores não foram cobertos, mas de modo geral, foram utilizados outros itens de mesmo tema para suprir sua ausência.

A dimensionalidade do teste tem relação com a validade dos resultados da testagem, pois se refere à estrutura interna do teste, ao grau de relação entre os itens e os componentes do teste em conformidade ao construto que o teste propôs medir (AERA, APA & NCME, 1999). Os quatro efeitos negativos gerados pela violação do pressuposto de unidimensionalidade são (Laros, Pasquali & Rodrigues, 2000): (a) diminuição da validade de construto do teste, dificultando a interpretação dos escores; (b) aumento da função diferencial do item; (c) dificuldade de realização da equalização dos resultados de várias formas de uma prova; e (d) as estimativas de habilidade apresentam baixo grau de validade, com um impacto especial para os desvios-padrão do parâmetro de habilidade que podem ser errôneos.

O INEP não realizou estudos de verificação da dimensionalidade do teste de 8ª série EF da Prova Brasil. Para a ANEB, o instituto de pesquisa realizou um estudo de verificação da unidimensionalidade (CESPE, 2007c) e identificou um conjunto de itens que não contribuíam significativamente com o fator principal do teste. A estimação das habilidades dos estudantes da ANEB 2005, por sua vez, foi realizada sem a exclusão desses itens.

A não-realização de estudos de verificação da dimensionalidade e a não utilização de estudos realizados podem estar associadas ao prazo que as empresas responsáveis pelas análises possuíam para entregar os resultados finais ao INEP. O tempo político e a necessidade de divulgação dos resultados das habilidades dos estudantes acabam não permitindo o cumprimento do cronograma estipulado nos projetos básicos (MEC/INEP/DAEB, 2005a; 2005b) e no contrato, embora o estudo de unidimensionalidade tenha sido previsto.

Na prática, as empresas responsáveis pelas análises priorizaram para 2005 um rigor metodológico para as etapas de análise clássica, análise pedagógica, da calibração, da análise DIF e da estimação das habilidades. A empresa responsável pelas análises da ANEB 2005 (CESPE, 2007b, 2007c) replicou todo o processo de calibração e de estimação das habilidades do SAEB 2003 (CESGRANRIO, 2004), antes de dedicarem-se à análise dos dados da ANEB 2005, para garantir que as próximas análises não fossem influenciadas por fatores inerentes aos procedimentos de análise. Sugere-se ao INEP um redimensionamento do tempo disponível para as análises, tendo em vista a relevância do pressuposto de unidimensionalidade dos testes para a validade dos resultados das avaliações.

A Função Diferencial do Item (DIF), também relacionada à validade (AERA, APA & NCME, 1999, p.13), foi verificada para a ANEB (CESPE, 2007b, 2007c) e a Prova Brasil (CESGRANRIO, 2006). Itens foram excluídos ou deixaram de funcionar como comuns entre séries ou entre anos da avaliação, para que as estimativas de habilidade não fossem significativamente influenciadas pelas características dos itens em seu funcionamento para grupos de estimativas de habilidade semelhantes. Essa evidência contribuiu para a validade dos resultados das duas avaliações.

Citam-se alguns fatores relacionados ao teste e que podem influenciar na fidedignidade: poder de informação do teste (Hambleton, Jones & Rogers, 1993); amostragem de conteúdo e consistência entre itens (Urbina, 2007); e tamanho do teste (Cronbach, 1996).

Os testes A, B e C, com 104 itens, buscaram a redução do número de itens total, mas sem redução do número de itens que cada estudante respondeu na Prova Brasil. Para o Teste A, observaram-se discriminação e dificuldade médias iguais ao Teste ANEB. Os resultados para o Teste A indicaram um aumento da habilidade estimada média associado à redução do número de itens no teste, já que os parâmetros psicométricos médios mantiveram-se os mesmos. O Teste B foi o mais discriminativo possível para 104 itens e apresentou resultados de estimativas de habilidade maiores, comparado à ANEB e ao Teste A, mas ainda distantes dos da Prova Brasil. Quando o teste D foi aplicado, menos itens (81) estavam em jogo. O total de itens foi similar ao da Prova Brasil, mas o número de itens respondidos por bloco e por caderno por aluno foi inferior. Como o critério para redução foi a exclusão dos itens menos discriminativos, atingiu-se o teste mais discriminativos de todos possíveis para a simulação proposta. Os resultados de estimativas

de habilidade foram superiores para o Teste D (0,27), mas aquém dos resultados da Prova Brasil.

A relação direta encontrada entre discriminação e estimativas de habilidade médias deve ser analisada com cautela. Quando se excluem itens menos discriminativos, por vezes, retiram-se dos testes itens com parâmetro *b* baixo. Os itens dos testes passam a estar concentrados nas faixas central e superiores da escala, como a Prova Brasil. Os procedimentos adotados para simulação dos testes, em termos de distribuição de itens, tornaram o Teste ANEB mais próximo das características da Prova Brasil. Isso pode ter gerado aumento das estimativas médias de habilidade.

Há evidências que o número de itens total e em cada caderno influencia na fidedignidade dos resultados obtidos pelo teste. O EPM médio ponderado pelo número de estudantes foi maior para os testes com menor número de itens. O Teste ANEB com 155 itens apresentou os melhores resultados de índice de fidedignidade e o Teste D, os piores. Evidência que quanto maior o número de itens do teste, maior a fidedignidade.

No entanto, como o EPM é estimado para cada estudante que responde a um determinado número de itens (Hambleton, Swaminathan & Rogers, 1991), houve razões para acreditar que o tamanho do caderno de teste respondido por estudante tinha sido determinante para a fidedignidade. Os resultados encontrados para o presente estudo indicaram que os examinandos que responderam a cadernos de testes com um maior número de itens apresentaram resultados de parâmetro de habilidade com índice de fidedignidade maior (menores EPM). Os resultados da ANEB, teste em que os estudantes respondiam a um número maior de itens (39), foram os mais fidedignos. O Teste D, simulado de forma que os estudantes respondessem em torno de 18 itens, apresentou resultados de habilidade estimada menos fidedignos.

No SAEB, a aplicação de um teste avaliando duas disciplinas é vantajosa. Avaliam-se duas áreas do conhecimento e de competência diferentes de cada estudante. O SAEB mudou substancialmente seu delineamento com a Prova Brasil 2005, o mesmo estudante passou a responder testes de duas disciplinas. O formato está associado a uma diminuição da fidedignidade que os resultados apresentam em cada disciplina, já que o número de itens que cada estudante responde para a Prova Brasil de uma mesma disciplina (24) é menor que o para a ANEB (39).

Observaram-se que todos os testes foram mais discriminativos (maiores parâmetro *a*) para a faixa de habilidade +1 a +2, seguido das faixas +2 a +3 e 0 a +1. Constatou-se também para essas faixas os menores EPM e conseqüentemente as maiores informações.

Isso confirmou a dependência da magnitude do EPM não só à quantidade, mas também à qualidade dos itens, de forma que EPM baixos são associados à alta discriminação, como previam Hambleton, Swaminathan & Rogers (1991).

Embora as faixas +1 a +2 e +2 a +3 tenham apresentados os itens mais discriminativos, menores EPM e maiores informações, o percentual de estudantes com parâmetro de habilidade localizado nessas faixas para todos os testes envolvidos neste estudo foi pequeno. Observa-se que o parâmetro de habilidade dos estudantes de 8ª série EF em matemática foi estimado, para a Prova Brasil e para a ANEB, com base em testes mais informativos para as faixas de habilidades maiores, o que não representa o perfil da maioria dos estudantes. A magnitude do EPM tem relação com a associação entre o parâmetro *b* e o parâmetro de habilidade, de forma que EPM pequenos são associados com testes compostos de itens com parâmetro *b* aproximadamente igual ao parâmetro de habilidade dos examinandos (Hambleton, Swaminathan & Rogers, 1991). Tanto para a Prova Brasil, quanto para a ANEB, observou-se maior concentração de itens para as faixas 0 a +1 e +1 a +2. Considera-se, no entanto, que o número de itens com boa qualidade discriminativa foi insuficiente para os testes exatamente para as faixas em que se localiza o maior percentual de estudantes (de -2 a -1 a 0 a +1).

Essas evidências têm uma relevância prática. É fundamental o planejamento do teste a partir da seleção de itens com parâmetro *a* alto para cada uma das faixas, principalmente para as faixas em que se localizam a maior parte dos estudantes. O rigor com relação ao critério de discriminação deve ser adotado para a composição dos testes do SAEB. Sugere-se, para as próximas edições do SAEB, especialmente, a seleção de itens mais discriminativos para as faixas de -2 a -1 a 0 a +1, de forma a melhorar a fidedignidade dos resultados para essas faixas, detectadas no presente estudo como razoavelmente fracas.

O Teste ANEB apresentou o maior índice de fidedignidade para todas as faixas de habilidades estimadas. É importante notar que o Teste ANEB não é mais discriminativo e que os resultados de informação da Prova Brasil foram sistematicamente inferiores que o da ANEB para todas as faixas de parâmetro de habilidade. Teoricamente, com resultados de discriminação maiores para a Prova Brasil, esperavam-se índices de informação maiores para esse teste. Não foi o observado. Novamente, os resultados sugerem que o número de itens que cada estudante responde tem relação com a fidedignidade do teste.

O Teste ANEB se mostrou mais adequado quanto à distribuição dos itens pelas faixas de habilidade. Ressalta-se que a Prova Brasil concentrou demais seus itens nas faixas 0 a +1 e +1 a +2 e não incluiu nenhum item para a faixa -3 a -2. Essa concentração

pode ter gerado um menor desvio-padrão para as estimativas de habilidade da Prova Brasil. Os itens de 4ª série EF incluídos na ANEB tiveram uma relevância para cobrir a faixa inferior da escala.

Os resultados de parâmetro de habilidade estimados a partir da simulação dos Testes A a D não atingiram os obtidos para a Prova Brasil, que apresentou estimativa de habilidade média maior que todos os testes. Os testes B e D alcançaram as estimativas de habilidade mais próximas da Prova Brasil, mas inferiores. O aumento da discriminação dos itens, associado à diminuição do número de itens gerou resultados de habilidades estimadas superiores aos da ANEB, mas fidedignidade inferior. O Teste D, inclusive, apresentou o menor índice de informação de todos os testes para todas as faixas, com exceção da faixa +2 a +3, em função do número de itens considerados por estudante ser inferior aos demais testes. O Teste B apresentou índice de informação superior ao da Prova Brasil para as faixas superiores da escala, em função do número de itens do teste como um todo (104) ter sido superior ao da Prova Brasil (81). Para o Teste B e a Prova Brasil, foram considerados cerca de 24 itens por estudante.

Para os testes B e D, observou-se, em comparação com a ANEB, um aumento do parâmetro *b* médio, superior à dificuldade da Prova Brasil. Associado a isso, verificou-se um aumento do parâmetro de habilidade média para os testes, o que indica que a propriedade de invariância do parâmetro de habilidade em função do parâmetro *b* parece proceder. Aumentando-se a dificuldade dos testes, não houve uma diminuição das estimativas de proficiências, como na TCT.

## 8. Conclusões

O estudo para verificação da associação entre características dos testes e estimativas de habilidade é pertinente, pois seus resultados podem orientar algumas decisões referentes ao planejamento metodológico do SAEB e, inclusive, podem servir de base para a construção de testes fora do âmbito da referida avaliação. Permite, ainda, uma reavaliação dos resultados encontrados para o SAEB 2005 (ANEB e Prova Brasil).

Os objetivos do presente estudo foram atingidos, pois foi possível identificar em que medida características dos testes, tais como cobertura da matriz, qualidade pedagógica, configuração psicométrica e tamanho, estão associadas à validade e à fidedignidade das estimativas de habilidade de examinandos. Adicionalmente, forneceu informações sobre fatores associados às diferenças dos resultados de estimativas de habilidade entre ANEB e Prova Brasil.

Citam-se as principais evidências identificadas e que contribuem para a validade dos resultados Prova Brasil e da ANEB: (a) utilizaram uma matriz de referência elaborada em consulta nacional com os especialistas e pautada nas Diretrizes Curriculares Nacionais; (b) envolveram itens de teste elaborados e revisados por especialistas nas disciplinas avaliadas e em técnicas de construção; (c) utilizaram pré-teste para cálculos das informações psicométricas dos itens; (d) envolveram novamente especialistas para seleção de itens e composição dos testes a partir dos resultados do pré-teste, de um esquema de prioridades pedagógicas e visando uma cobertura equilibrada da matriz; (e) envolveram novamente especialistas para uma última análise pedagógica; e (f) utilizaram estudos de verificação do DIF para tomada de decisão.

Os resultados do estudo de verificação da dimensionalidade, também relacionado à validade, realizado para a ANEB não foram utilizados para exclusão de itens antes da estimação das habilidades dos estudantes. Estudos de verificação da unidimensionalidade para a Prova Brasil não foram encontrados na literatura. A inclusão de possíveis itens que não contribuíram significativamente para o fator principal pode ter impactado negativamente na validade das estimativas de habilidade. Sugere-se a realização de estudos que busquem verificar esse impacto. Ainda, recomenda-se identificar as causas da não-realização ou da não-utilização dos estudos de verificação da dimensionalidade antes da realização das estimações das habilidades, já que foram previstos. Caso o motivo esteja relacionado à pressão do INEP e do MEC para divulgação dos resultados, o que é possível, sugere-se revisão de calendário, já que estudos anteriores mostraram que o distanciamento da unidimensionalidade apresenta efeito na invariância do parâmetro de habilidade da TRI.

Um maior número de itens no teste e no caderno fornece estimativas de habilidade mais fidedignas. Se há outras vantagens da diminuição do número de itens total e por caderno na Prova Brasil como, por exemplo, permitir a avaliação de duas disciplinas por estudante, deve-se ter ciência de certo prejuízo para a fidedignidade das estimativas de habilidade.

É importante compor testes com itens de alta discriminação, já que o parâmetro *a* está diretamente associado à fidedignidade dos resultados. Evidenciou-se certa dependência entre o parâmetro de habilidade e o parâmetro *a*, diferentemente do que pressupõe a TRI (Baker, 2001; Hambleton, Swaminathan & Rogers, 1991).

No caso do teste de matemática 8ª série EF do SAEB, sugere-se uma maior atenção quanto à discriminação dos itens localizados na faixa -1 a 0, onde estão localizadas as habilidades estimadas da maioria dos estudantes brasileiros. Observou-se que o poder informativo dos testes para essa faixa foi baixo em 2005, especialmente para a Prova Brasil. Essa constatação está associada claramente ao número de itens discriminativos utilizados na faixa, inferior às faixas superiores, onde a informação dos testes foi máxima.

Sugere-se que, na composição dos testes do próximo SAEB, as informações, por faixa de estimativas de habilidade, de parâmetro *a* médio, de número de itens (relacionadas ao parâmetro *b*) e de EPM, obtidos com base no presente estudo ou recalculados para o SAEB 2007, sejam apresentados aos especialistas. Sugere-se um planejamento do teste de matemática 8ª série EF de forma que (a) um maior número de itens seja incluído para a faixa -1 a 0; (b) um número não tão grande de itens seja alocado para a faixa de +1 a +2; e (c) itens com parâmetros *a* altos sejam selecionados para o teste, em especial para a faixa -1 a 0. Estas sugestões têm por base não só os resultados deste estudo, mas os fatores considerados por Hambleton, Swaminathan e Rogers (1991) como determinantes para a magnitude do EPM, ou seja, da fidedignidade das estimativas: (a) o número de itens do teste; (b) a qualidade discriminativa dos itens; e (c) a associação do parâmetro *b* com o de habilidade. O poder discriminativo e de informação do teste do SAEB tenderá a aumentar, acompanhado de um aumento da fidedignidade das estimativas de habilidade dos estudantes.

Após todos os procedimentos analíticos realizados, o que se pôde concluir com relação à qualidade dos testes ANEB e Prova Brasil? Quais resultados apresentam maior grau de validade e de precisão? Em nenhum momento, o presente estudo afirmou que os resultados da Prova Brasil ou da ANEB seriam os mais corretos, os mais válidos e fidedignos. No entanto, o conjunto de evidências sugere que os resultados da ANEB

apresentam um grau de fidedignidade maior. Se a Prova Brasil apresentou a maior discriminação média, a ANEB apresentou (a) maior número de itens total e por caderno; (b) maior número de itens para as faixas de habilidades com maior percentual de estudantes; (c) melhor distribuição dos itens pelas faixas de habilidades. Ainda, o poder informativo da ANEB foi superior ao da Prova Brasil para todas as faixas. Se a diferença entre as estimativas de habilidade entre ANEB e Prova Brasil foi gerada pela diferença de fidedignidade, supõe-se que os resultados da ANEB sejam mais confiáveis.

Com os testes simulados, Testes A a D, foi possível identificar que a exclusão dos itens com menores parâmetro *a* gera exclusão dos itens com menores parâmetros *b,* já que esses discriminam menos, pois localizam-se em uma posição inferior da escala. Isso forçou uma concentração de itens para as faixas média e alta e foi verificada uma diminuição da variabilidade das estimativas de habilidade dos estudantes, como a observada na Prova Brasil, e uma aproximação com relação à media da Prova Brasil. Associada a essa configuração há uma perda da fidedignidade gerada pela queda no número de itens dos testes.

Todas essas evidências relacionam-se ao quanto os testes de matemática 8ª série EF do SAEB, compostos em 2005, foram apropriados para o seu público alvo. Pode-se dizer que os testes foram mais discriminativos e informativos para os estudantes com maiores estimativas de habilidades. Por sua vez, os resultados careceram de fidedignidade para os estudantes com menores habilidades estimadas, ou seja, a maioria.

A testagem adaptativa pode oferecer vantagens ao SAEB, quando itens discriminativos e apropriados à sua magnitude de habilidade estimada são apresentados aos estudantes. Assim, estudantes com baixas estimativas de habilidade seriam avaliados com itens mais apropriados ao seu nível de competência. Essa sugestão, no entanto, requer condições logísticas, de infra-estrutura e analíticas apropriadas para administração de testes por meio de computadores.

Identificaram-se algumas limitações do presente estudo: (a) não foi possível simular um teste com as mesmas características da Prova Brasil; (b) não foram simulados testes a partir da exclusão dos itens que não contribuíram significativamente para o fator principal, a partir de estudo de verificação da dimensionalidade; (c) não foi verificado o grau de ajuste do modelo aos dados, fundamental para avaliação da validade.

## 9. Referências


Alchieri, J. C. & Cruz, R. M. (2004). *Avaliação psicológica: conceitos, métodos e instrumentos*. São Paulo: Casa do Psicólogo.

American Educational Research Association – AERA, American Psychological Association – APA & National Council on Measurement in Education – NCME (1999). *Standards for educacional and psychological testing*. New York: AERA.

Araújo, C. H., Condé, F. N. & Luzio, N. (2004). Índice de Qualidade da Educação Básica, IQE: proposta para discussão. *Revista Brasileira de Estudos Pedagógicos, INEP, 85* (209/210/211), 126-136.

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

Barreto, E. S. S. & Pinto, R. P. (2001). Avaliação na Educação Básica (1990-1998). *Série Estado do conhecimento, n.4*. Brasília: MEC/INEP/Comped.

Bekman, R. M. (2001). Aplicação dos blocos incompletos balanceados na teoria de resposta ao item. *Estudos em Avaliação Educacional, 24*, 119-135.

Bhola, D. S., Impara, J. C. & Buckendahl, C. W. (2003). Aligning tests with states' content standards: methods and issues. *Educational measurements: issues and practice, 22* (3), 21-29.

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

Bock, R. D., Gibbons, R. & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Bolt, D. M., Cohen, A. S. & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: applications of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurements, 39*, 331-348.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City: ACT Publications.

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

Camilli, G. (2006). Test fairness. Em R. L. Brennan (Org.), *Educational Measurement* (pp. 221-256). Westport, CT: American Council on Education/Praeger.

CESGRANRIO (2004). *SAEB 2003: relatório técnico da análise da teoria clássica dos testes e da teoria de resposta ao item.* Rio de Janeiro: Fundação CESGRANRIO.

CESGRANRIO (2006). *Prova Brasil 2005: relatório técnico da análise da teoria de resposta ao item e da teoria clássica dos testes*. Rio de Janeiro: Fundação CESGRANRIO.

CESPE (2007a). ANEB 2005: relatório da teoria clássica dos testes. Brasília: CESPE/UnB.

CESPE (2007b). ANEB 2005: relatório técnico da análise da teoria de resposta ao item. Brasília: CESPE/UnB.

CESPE (2007c). ANEB 2005: relatório técnico da análise da teoria de resposta ao item (versão 2). Brasília: CESPE/UnB.

Cohen, R. J. & Swerdlik, M. E. (2002). *Psychological testing and assessment: an introduction to tests and measuremente*. USA: McGraw Hill.

Condé, F. N. & Laros, J. A. (2007). Unidimensionalidade e a propriedade de invariância das estimativas da habilidade pela TRI. *Revista Avaliação Psicológica, 2007, 6*(2), 205-215.

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: UFBA/ISP/FAPEX.

Condé, F. N. (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.* Dissertação de Mestrado, Universidade de Brasília.

Condé, F. N. (2007). O efeito dos modelos de testes na estimativa da habilidade dos estudantes: comparação entre Prova Brasil e SAEB 2005. Em *1ª Primeira Jornada de Avaliação Formativa do Programa de Pós-graduação em Psicologia Social do Trabalho e das Organizações – PSTO da Universidade de Brasília*. Brasília: UnB.

Condé, F. N. (2007b). O Índice de Qualidade da Educação Básica: estrutura e comparação com o IDEB. *Resumo de apresentação em mesa redonda do III Congresso Brasileiro de Avaliação Psicológica e XII Conferência Internacional de Avaliação Psicológica: Formas e Contextos*. Brasília: IESB, UnB.

Cronbach, L. J. (1996). *Fundamentos da testagem psicológica*. Porto Alegre: Artes Médicas.

Cronbach, L. J., Gleser, G. C, Rajaratnam, N. & Nanda, H. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349.

Fan, X. & Ping, Y. (1999). Assessing the effect of model-data misfit on the invariance property of IRT parameter estimates. Artigo apresentado na *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.

Fernandes, R. (2007). Índice de Desenvolvimento da Educação Básica, IDEB. *Série Documental, Texto para Discussão, INEP*. Retirado em 11/09/2007 no World Wide Web: http://www.publicacoes.inep.gov.br/.

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

Ferrara, S. & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12. Em R. L. Brennan (Org.), *Educational measurement* (4ª ed., pp. 579-621). Westport, CT: American Council on Education/Praeger.

Ferrara, S. (2006). Toward a Psychology of large-scale educational achievement testing: some features and capabilities. *Educational measurements: issues and practice, 25*(4), *2-5*.

Fundação Carlos Chagas – FCC (2001). Avaliação na educação básica (1990-1998). Em E. S. S. Barreto & R. P. Pinto (Orgs), *Série estado do conhecimento, 4*. Brasília: MEC/INEP/ COMPED.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R. K, Jones, R. W. & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of educational measurement, 30*(2), 143-155.

Hambleton, R. K. & Jones, R. W. (1993). An NCME Instructional Module on Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R. K., Swaminathan, H. & 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.

Herman, J. L., Webb, N. & Zuniga, S. (2002). *Alignment and college admissions: The match of expectations, assessments, and educator perspectives*. Artigo apresentado em the annual meeting of the American Educational Research Association, New Orleans, LA.

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

Instituto Nacional de Estudos e Pesquisas Educacionais – INEP (1999). *Matrizes Curriculares de Referência*. Brasília: INEP.

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

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

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP (2005a). *ANRESC 2005: manual do aplicador*. Brasília: Fundação CESGRANRIO.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP (2005b). *ANEB 2005: manual do aplicador*. Brasília: CESPE.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP (2007a). *Saeb 2005 primeiros resultados: médias de desempenho do SAEB 2005 em perspectiva comparada.* Brasília: INEP.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP (2007b). *Relatório Psicométrico de Montagem das Provas do Pré-teste*. Brasília: INEP.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29*(2), 95-110.

Kirisci, L., Hsu, T. & Yu, L (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. Applied Psychological Measurement, 25(2), 146-162.

Klein, R. & Klein, T. S. (1998). *Programa para Teoria Clássica dos Testes*. Rio de Janeiro: Fundação CESGRANRIO.

Kvanli, A. H., Guynes, C. S. & Pavur, R. J. (1991). *Introduction to business statistics*. 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.

Lu, Y. & Sireci, S. G. (2007). Validity issues in test speededness. *Educational measurements: issues and practice, 26* (4), 29-37.

Luckesi, C. C. (2003). *Avaliação da Aprendizagem escolar: estudos e preposições*. São Paulo: Cortez.

Maloney, M. P. & Ward, M. P. (1976). *Psychological assessment: a conceptual approach*. New York: Oxford University Press.

Manfredi, S. M. (1998). Trabalho, qualificação e competência profissional das dimensões conceituais e políticas. *Educ. Soc. [online], 19* (64), 13-49. Retirado em 13/09/2007 no World Wide Web: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-73301998000300002&lng=pt&nrm=iso

McIntire, S. A. & Miller, L. A. (2000). *Foundations of psychological testing*. USA: McGraw-Hill.

Ministério da Educação – MEC, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP, Diretoria de Avaliação da Educação Básica – DAEB (2005a). Projeto Básico, ANEB. Brasília: MEC/INEP.

Ministério da Educação – MEC, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP, Diretoria de Avaliação da Educação Básica – DAEB (2005b). Projeto Básico, ANRESC. Brasília: MEC/INEP.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. Em R. L. Brennan (Org.), *Educational measurement* (4th ed., pp. 257-305). Westport, CT: American Council on Education/Praeger.

Nacional Center for Education Statistics – NCES (1992a). *Geografy Framework for the 1994 and 2001 National Assessment of Educational Progress*. Washington, DC: U.S. Department of Education.

Nacional Center for Education Statistics – NCES (1992b). *U.S. History Framework for the 1994 and 2001 National Assessment of Educational Progress*. Washington, DC: U.S. Department of Education.

Nacional Center for Education Statistics – NCES (1995a). *Science Framework for the 1996 and 2000 National Assessment of Educational Progress*. Washington, DC: U.S. Department of Education.

Nacional Center for Education Statistics – NCES (1995b). *Writing Framework for the 1998 National Assessment of Educational Progress*. Washington, DC: U.S. Department of Education.

Nacional Center for Education Statistics – NCES (1996). *The NAEP Guide*, by Ballator, N., editors. Washington, DC: U.S. Department of Education.

Nacional Center for Education Statistics – NCES (1997). *The NAEP Guide*, NCES 97-990, by Calderone, J., King, L.M., & Horkay, N., editors. Washington, DC: U.S. Department of Education.

Nacional Center for Education Statistics – NCES (1999). The NAEP Guide, NCES 2000-456, by Horkay, N., editor. Washington, DC: U.S. Department of Education.

Nacional Center for Education Statistics – NCES (2002a). *Mathematics Framework for the 2003 National Assessment of Educational Progress*. Washington, DC: U.S. Department of Education.

Nacional Center for Education Statistics – NCES (2002b). *Reading Framework for the 2003 National Assessment of Educational Progress*. Washington, DC: U.S. Department of Education.

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

Organisation for Economic Co-operation and Development – OECD (2000). *Measuring student knowledge and skills: the PISA assessment of reading, mathematical and scientific literacy*. França: OECD.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31* (3), 200-219.

Pasquali, L. & Alves, A. R. (1999). Testes referentes a conteúdos: medidas educacionais. Em L. Pasquali (Org), *Instrumentos psicológicos: manual prático de elaboração* (pp. 141-182). Brasília: LabPAM/IBAP.

Pasquali, L. (1996). *Teoria e métodos de medida em ciências do comportamento*. Brasília: LabPAM/UnB/INEP.

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

Pasquali, L. (1998). Princípios de elaboração de escalas psicológicas. *Revista de Psiquiatria Clínica, 25* (5). Retirado em 04/02/2008 no World Wide Web: http://www.hcnet.usp.br/ ipq/revista/r255/conc255a.htm.

Pasquali, L. (2003). *Psicometria: teoria dos testes na psicologia e na educação.* Petrópolis: Vozes.

Perrenoud, P. (1993). *Práticas pedagógicas, profissão docente e formação*. Lisboa: Don Quixote.

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

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

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

Pestana, M. I. G. S. (2006). *A polissemia da noção de competência: uma análise do conteúdo do discurso do setor público sobre gestão, avaliação e certificação de competências*. Dissertação de mestrado, PUC/SP.

Programa das Nações Unidas para o Desenvolvimento – PNUD (2006). *Termos de referência para contratação de empresa especializada para elaborar itens de Língua Portuguesa e de Matemática para o Banco Nacional de Itens da Diretoria de Avaliação da Educação Básica, Anexo I*. Brasília: INEP.

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

Rabello, G. C. (2007). *Relatório Técnico das análises estatísticas a partir dos dados da Prova Brasil, para subsidiar a elaboração de documentos de divulgação*. Brasília: PNUD/ INEP.

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, Universidade de Brasília.

Sant'anna, F. M., Enricone, D., André, L. C. & Turra, C. M. G. (1996). *Planejamento de ensino e avaliação*. Porto Alegre: Sagra Luzzatto.

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

Sireci, S. G. (2005). Unlabeling the disabled: a perspective on flagging scores from accommodated test administrations. *Educational Researcher, 34*(1), 3-12.

Sireci, S. G., Scarpati, S. & Li, S. (2005). Test accommodations for students with disabilities: an analysis of the interaction hypothesis. *Review of Educational Research, 75*, 457-490.

Snow, R. E. & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. Em R. L. Linn (Org.), *Educational measurement* (pp. 263-331). New York: American Council on Education/Macmillan.

Tyler, R. W. (1950). *Basic principles of curriculum and instruction*. Chicago, JL: University of Chicago Press.

Universidade Federal de Juiz de Fora – UFJF (2001). *Minas Gerais: avaliação da educação*. Juiz de Fora: UFJF.

Urbina, S. (2007). *Fundamentos da testagem psicológica*. Porto Alegre: Artmed.

Vianna, H. M. (1982). *Testes em educação*. São Paulo: Ibrasa.

Wilson, D. T., Wood, R. & Gibbons, R. (1991). *Testfact: test scoring, item statistics and item factor analysis*. Chicago: Scientific Software International (SSI).

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E. & Bock, D. (2003). Testfact 4. Em M. Du Toit (Org.), *IRT from SSI*. Chicago: Scientific Software International (SSI).

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).