Proposta de Análise de Itens das Provas do Saeb sob a Perspectiva Pedagógica e a Psicométrica

MARGARIDA M. M. RODRIGUES
Consultora Externa da Diretoria de Avaliação da Educação Básica – DAEB/Inep/MEC
margaridarodrigues@uol.com.br

Resumo
O presente artigo é resultante de uma pesquisa que teve por objetivo analisar os itens das provas de Matemática – 8ª série –, do Sistema Nacional de Avaliação da Educação Básica (Saeb), dos anos de 1997 e 1999, considerando-se os seus aspectos qualitativos e quantitativos. Os aspectos qualitativos foram analisados por meio da análise da validade de conteúdo e dos procedimentos efetivos da elaboração dos itens. Os aspectos quantitativos foram analisados, considerando-se as propriedades estatísticas, as quais incluíram procedimentos de análise da Teoria Clássica dos Testes (TCT) e da Teoria de Resposta ao Item (TRI). As análises realizadas mostraram que as avaliações de desempenho, principalmente as de larga escala, devem levar em conta os aspectos psicométricos e pedagógicos dos itens de forma integrada, sem privilégio de nenhuma delas. Dessa forma, constata-se que há uma maior compreensão dos resultados dessas avaliações, os quais poderão reverter em políticas mais adequadas de intervenção na busca da melhoria da qualidade da educação, propósito final do Saeb.
Palavras-chave: Psicometria, Validade de conteúdo, Teoria Clássica dos Testes (TCT), Teoria de Resposta ao Item (TRI).

Resumen
El presente artículo resultó de una investigación que tuvo como objetivo analizar los ítems de las pruebas de Matemática de la *8ª série* (coresponde al alumno padrón de 14 años), del Sistema Nacional de Evaluación de la Educación Básica (Saeb), de 1997 y 1999. Fueron considerados tanto los aspectos cualitativos como los cuantitativos de la evaluación. Los aspectos cualitativos fueron analizados por medio del análisis de la validez de contenido y de los procedimientos efectivos de la elaboración de los ítems. Los aspectos cuantitativos fueron analizados en términos de las propiedades psicométricas, incluyendo procedimientos de análisis de la Teoría Clásica de los Tests (TCT) y de la Teoría de Respuesta al Ítem (TRI). Los análisis mostraron que las evaluaciones de desempeño, principalmente las de larga escala, deben llevar en cuenta los aspectos psicométricos y pedagógicos de los ítems. Cuando estos análisis son hechos en forma integral, sin privilegiar a ninguno de ellos, se observa que hay una mejor comprensión del proceso enseñanza-aprendizaje y los resultados pueden convertirse en políticas públicas de intervención más adecuadas y que buscan mejorar la calidad de la educación, propósito final del *Saeb*.
Palabras-clave: Psicometria, Validez de contenido, Teoría Clásica de los Testes (TCT), Teoría de Respuesta al Ítem (TRI).

Abstract
This article is the result of a study that aimed at analyzing the items of 8th grade mathematics tests from the National Basic Education Evaluation System (Saeb), between 1997 and 1999, considering both their qualitative and quantitative aspects. The qualitative aspects were analyzed using content validity analysis and effective item elaboration procedures. The quantitative aspects were analyzed considering statistical properties, including Classical Test Theory (CTT) and Item Response Theory (IRT) analysis procedures. The analyses revealed that performance evaluations, mainly wide-scale ones, should take both psychometric and pedagogical aspects of the items into consideration. When these analyses are carried out in an integrated way, without privileging either one of them, a greater comprehension of the teaching-learning process is achieved and the results can be converted into more adequate intervention policies to improve the quality of education, SAEB´s ultimate purpose.
Key words: psychometrics, content validity, Classical Test Theory (CTT), Item Response Theory (IRT).

1 INTRODUÇÃO

A concepção de avaliação educacional apresenta um caráter multifacetado com uma extensa bibliografia a respeito do tema. A avaliação pode ser contemplada de várias formas e por meio de diferentes métodos. Suas abordagens vinculam-se aos diversos paradigmas que vão se contextualizando através do tempo. Dessa forma, sua concepção reflete uma postura filosófica em face da educação. Observa-se, porém, que, independentemente do paradigma ou da postura filosófica, há um eixo comum entre as diversas concepções de avaliação educacional. Todas elas associam-se a um processo interpretativo de dados quantitativos e/ou qualitativos, supondo um juízo de valor, qualidade ou mérito que tem por meta diagnosticar e verificar o alcance dos objetivos propostos no processo ensino-aprendizagem.

Sabe-se que avaliar, se tais objetivos foram alcançados, não decorre de uma simples verificação da aprendizagem. Esse diagnóstico vai muito além, pois há toda uma conjuntura que propicia a aprendizagem do aluno ou não. No cotidiano, constata-se que o processo pedagógico ocorre por meio da relação que se estabelece entre professores, alunos, direção, administração, estrutura física da escola, comunidade, entre outros, e nessa relação estão envolvidas as múltiplas dimensões que formam cada ser humano. Portanto, uma avaliação, que pretenda avaliar a qualidade da educação oferecida por uma escola, por uma rede ou por um sistema, deve estar embasada em um modelo que contemple todas as relações possíveis de serem avaliadas.

O objetivo deste artigo é discutir o instrumento utilizado para avaliar o desempenho dos alunos pelo Sistema Nacional de Avaliação da Educação Básica (Saeb), apresentando um modelo de análise das provas e itens fundamentado em análises pedagógicas e psicométricas, as quais foram realizadas de forma integrada. A principal contribuição deste estudo é mostrar que, ao realizar as análises sugeridas, as provas podem se constituir em indicadores confiáveis e úteis para o sistema de informações da qualidade da educação brasileira.

2 O SAEB

A primeira discussão de um sistema de avaliação em larga escala surgiu durante o período de redemocratização do país, em 1985-1986. O objetivo principal da primeira proposta foi levantar informações úteis sobre o que estava sendo gerado no setor educacional, como, onde, quando e

quem eram os responsáveis pelo produto obtido. Dessa forma, surgiu o Sistema Nacional de Avaliação da Educação Básica (Saeb), como uma atribuição do Ministério da Educação que, em 1990, iniciou a coleta de informações sobre a qualidade da educação brasileira.

O Saeb, atualmente, avalia, de maneira sistemática e periódica, o desempenho dos alunos da educação básica em todo o território nacional. A finalidade primeira do Saeb é obter resultados sobre a qualidade do ensino ao longo do tempo e identificar os fatores que contribuem para a ocorrência desses resultados, visando a melhoria da qualidade da educação básica do Brasil. O segundo objetivo do Saeb é monitorar o avanço alcançado pelos programas e pelas políticas governamentais em relação às metas educacionais.

Para atingir esses objetivos, vários instrumentos são empregados, além das provas de avaliação do desempenho escolar. Adicionalmente, são utilizados questionários que permitem: 1) obter informações sobre as características da realidade socioeconômica e cultural e hábitos de estudo dos alunos; 2) avaliar o perfil e as práticas pedagógicas dos professores; 3) avaliar o perfil e as práticas de gestão escolar dos diretores; e 4) realizar o levantamento dos equipamentos disponíveis e das características físicas e de conservação das escolas. Os indicadores resultantes dessas avaliações permitem que se façam associações, correlações, análises hierárquicas e estudos relevantes sobre a realidade educacional brasileira.

A construção de instrumentos, que objetivam medir o nível de competência dos alunos, requer um conhecimento sistemático das habilidades específicas a serem alcançadas no processo ensino-aprendizagem. Assim, as provas para avaliar o desempenho dos alunos constituem um processo de coleta de dados de uma amostra representativa de comportamentos. Este processo envolve o conhecimento das diferentes habilidades que são requeridas para a construção de determinadas competências que usam como meio os conteúdos que servem de orientação para o processo ensino-aprendizagem.

3 CONSTRUÇÃO DAS PROVAS DO SAEB EM 1997 E 1999

Neste estudo, foi feita a análise das provas de Matemática da 8ª série, do Saeb, aplicadas em 1997 e 1999. Como os resultados dessas provas são indicadores úteis para o sistema de informação da qualidade da educação brasileira, é importante ressartar que eles devem demonstrar e comprovar a sua objetividade, confiabilidade e qualidade.

A elaboração das provas teve por base as Matrizes Curriculares de Referência (Pestana et al., 1997, 1999). É importante observar que essas provas têm alcance nacional; portanto, é fundamental que sejam orientadas pelo estabelecimento prévio dos conteúdos desejáveis e necessários às demandas e exigências implícitas no sistema educacional brasileiro, além de considerar todas as diferenças regionais.

As matrizes curriculares, tanto de 1997 quanto de 1999, foram desenvolvidas a partir de uma ampla consulta nacional e consensual sobre os conteúdos praticados nas escolas brasileiras de ensino fundamental e médio, bem como da reflexão de professores, pesquisadores e especialistas a respeito da produção científica em cada área que se torna objeto de conhecimento escolar. Estabelecidos os conteúdos, estes foram hierarquizados e distribuídos em três ciclos, com terminalidades na 4ª e 8ª séries do Ensino Fundamental (EF) e na 3ª série do Ensino Médio (EM), abrangendo as seguintes disciplinas: Língua Portuguesa, Matemática, Ciências, História, Geografia, Física, Química e Biologia.

A esses conteúdos foram associadas as competências cognitivas exigidas para cada uma das disciplinas, assim como as habilidades instrumentais delas advindas. Citando Pestana et al., 1997:

> *Competências cognitivas são modalidades estruturais da inteligência, isto é, operações que o sujeito realiza para estabelecer relações com e entre os objetos, situações, fenômenos e pessoas (observar, representar, imaginar, reconstruir, comparar, classificar, ordenar, memorizar, interpretar, inferir, criticar, supor, levantar hipóteses, escolher, decidir etc.). Já as habilidades instrumentais referem-se especificamente ao plano do 'saber fazer' e decorrem diretamente do nível estrutural das competências adquiridas que se transformaram em habilidades.* (p.7)

As competências foram categorizadas em três níveis distintos de ações e de operações mentais, que se diferenciam pela qualidade das relações que se estabelecem entre o sujeito e o objeto do conhecimento: o nível básico, o operacional e o global.

No *nível básico* (presentativo) estão as ações que tornam presente o objeto do conhecimento para o sujeito. No *nível operacional* (procedural) estão as ações e operações que pressupõem o estabelecimento de relações com e entre os objetos. No *nível global* encontram-se as ações e operações mais complexas que envolvem a aplicação de conhecimentos e a resolução de problemas inéditos. Para cada nível de competências, são listadas as ações e as operações correspondentes esperadas para todos os conteúdos e séries avaliadas.

A construção das matrizes ocorreu pela constituição do universo possível de cruzamentos entre os conteúdos e as competências referentes aos diferentes níveis e ciclos de avaliação. Deste cruzamento, resultaram os descritores do desempenho desejável do aluno que, no seu conjunto, expressam a totalidade dos indicadores necessários para a orientação da construção dos itens constituintes das provas.

A matriz curricular de referência para cada disciplina ficou estruturada da seguinte forma: na dimensão conteúdos foram expostos os temas e tópicos e/ou assuntos relacionados a cada disciplina e série, e na dimensão competências foram colocados os três níveis de competências e habilidades ou descritores envolvidos. Os descritores referentes a cada tópico foram então associados a cada competência, e os itens foram construídos seguindo o critério de proporcionalidade.

As Matrizes Curriculares de Matemática de 1997 e 1999 se apoiaram em três premissas básicas: 1) os conceitos matemáticos não se constituem verdades absolutas e são formados de maneira inter-relacionada, contemplando diferentes procedimentos de solução; 2) a aquisição do conhecimento de Matemática dá-se por meio de aprendizagens significativas, as quais estão relacionadas com o mundo real do sujeito, interpretado e construído em diferentes linguagens; e 3) a avaliação deve aproximar-se o máximo possível da situação de aprendizagem do aluno. Essas três premissas, aliadas às limitações impostas ao tipo de avaliação a ser realizada, indicaram a proposição de uma matriz compreendida basicamente de situações-problema por meio da qual tem-se a possibilidade de avaliar satisfatoriamente as competências evidenciadas pela aprendizagem dos conteúdos matemáticos.

4 ANÁLISE DE ITENS

Os itens elaborados para cada prova podem ser analisados qualitativamente em termos pedagógicos, de conteúdo e forma, assim como quantitativamente em termos psicométricos, ou seja, das propriedades estatísticas. A análise qualitativa é realizada com base na validade de conteúdo e nos procedimentos efetivos da elaboração dos itens. A análise quantitativa inclui procedimentos de análise da Teoria Clássica dos Testes (TCT), da Análise Fatorial e da Teoria de Resposta ao Item (TRI). Ambas as análises (qualitativas e quantitativas) visam avaliar a validade, a fidedignidade e a objetividade dos testes.

Segundo Anastasi e Urbina (2000), os procedimentos de validação e descrição do conteúdo de uma prova envolvem, principalmente, o seu

exame sistemático, para determinar se ele abrange uma amostra representativa do domínio do comportamento a ser medido. O conteúdo precisa, portanto, ser amplamente definido para incluir todos os objetivos importantes desde a aplicação até o conhecimento factual da aprendizagem. Deve-se cuidar, ainda, para que o teste realmente meça o que propôs medir, de forma a incluir itens que cubram tão-somente o conteúdo a ser avaliado e que revele os processos usados pelo educando para fazer o teste.

Nunnaly e Bernstein (1994) afirmam que a validade de conteúdo também se refere a uma questão de generalização científica – a extensão segundo a qual, pode-se generalizar, de um conjunto particular de itens, todos os itens possíveis relacionados a um domínio maior.

Os procedimentos específicos para a validade de conteúdo incluem: 1) a escolha dos conteúdos apropriados; 2) a elaboração de uma tabela de especificações dos testes; 3) a distribuição proporcional por ordem de importância; e 4) a análise teórica dos itens, incluindo a análise semântica por sujeitos da própria população de interesse e a análise do conteúdo do teste por peritos nas áreas do conhecimento. Os itens que não alcançarem tais critérios deverão ser retirados do conjunto de itens.

A análise empírica dos itens é realizada por meio dos dados coletados de uma amostra representativa de sujeitos de uma população cujo sistema está sendo avaliado, utilizando-se análises estatísticas. A análise, embora utilize técnicas estatísticas diferentes, fornece informações que, na maioria das vezes, se confirmam.

4.1 ANÁLISE DE ITENS PELA TCT

O modelo clássico da psicometria tradicional (Pasquali, 1997) está fundamentado na Teoria Clássica dos Testes (TCT). Esta considera os testes como um conjunto de estímulos comportamentais (itens) cuja qualidade é definida em termos de um critério; este, por sua vez, é representado por comportamentos presentes ou futuros. A TCT está apoiada no seguinte paradigma: o escore empírico ou bruto do sujeito é constituído de dois componentes: 1) o escore real ou verdadeiro (V) do sujeito no comportamento avaliado; e 2) o erro de medida (E). O erro, sempre presente em qualquer medida empírica, resulta no modelo fundamental da psicometria, o qual confirma a tese de que o escore bruto de um examinando é a soma do escore verdadeiro e do erro (T= V + E). Este modelo implica alguns postulados básicos: a) o escore esperado é o escore verdadeiro. Isto decorre do conceito de esperança matemática do escore

empírico, ou seja, se o sujeito responde infinitas vezes ao mesmo teste, ele terá infinitos diferentes escores empíricos, e a média destes infinitos escores será o escore verdadeiro, porque ela eliminaria os erros; b) não há correlação entre o escore verdadeiro e o erro, pois a correlação entre o escore verdadeiro e o erro é zero; portanto, não há nenhuma razão para supor que escores verdadeiros maiores terão erros positivos e escores verdadeiros menores terão erros negativos; e c) os erros em testes paralelos não são correlacionados.

O modelo da TCT é baseado em dados empíricos coletados de um conjunto de itens agrupados inicialmente de maneira intuitiva. O teste é construído por meio da seleção de uma amostra de itens coletados de um universo que parece medir um dado construto. Essa maneira de construir instrumentos psicométricos está fundamentada na idéia de que existe, para cada construto, um conjunto indefinido de itens, a partir do qual uma amostra é extraída para construir o teste. A definição dos itens, que comporão o teste, é feita por meio da validade aparente, ou seja, escolhem-se aqueles itens que parecem estar medindo a mesma coisa. Na TCT, os parâmetros do item e da habilidade são dependentes da amostra e do teste.

A validade na TCT consiste na verificação da hipótese de que o teste é capaz de predizer um critério externo, o qual é representado por comportamentos. Assim, a demonstração da validade é uma questão de legitimação do instrumento em relação ao erro de estimação, ou seja, é a verificação da magnitude do escore verdadeiro que é concebido como representante legítimo do traço latente.

Um parâmetro importante a ser analisado, utilizando-se a TCT, é a dificuldade dos itens que compõem um teste. Esta pode ser definida como a porcentagem de examinandos que respondem corretamente aos itens. O cálculo da dificuldade de cada item, ou o valor *p*, é feito dividindo-se o número de pessoas que acertaram o item pelo número total de pessoas que o responderam. Geralmente, testes que alcancem um índice médio de dificuldade em torno de 0,5 produzem distribuições de escores no teste com maior variação (Bloom, 1971; Vianna, 1982; Pasquali, 1997; McIntire, Miller, 2000; Anastasi, Urbina, 2001). Para fins de avaliação de larga escala, os testes devem ser compostos de itens que alcancem todo o *continuum* da escala, ou seja, devem ter uma amplitude que inclua itens fáceis, medianos e difíceis (Vianna, 1989).

Outro parâmetro importante é a discriminação dos itens, que se refere ao poder que um item possui para distinguir sujeitos com magnitudes de traços diferentes, do qual o item constitui a representação comportamental (Pasquali, 1997). Quanto mais próximas forem as magnitudes do traço que o item puder diferenciar, mais discriminativo ele

será. Estatisticamente, esse conceito, na TCT, representa a correlação dos escores dos sujeitos no item com seus escores no teste total. De acordo com Marshall e Hales (1972), em Wilson, Wood e Gibbons (1991), existem mais de 60 índices propostos para medir o poder de discriminação de um item.

O Saeb utiliza a correlação bisserial. Esta é uma medida de associação entre o desempenho no item e o desempenho no teste. A correlação bisserial é menos influenciada pela dificuldade do item e tende a apresentar menos variação de uma situação de testagem para outra (Wilson, Wood, Gibbons, 1991). Sua fórmula é:

$$r_b = \frac{\overline{M_p} - \overline{M}}{S} x \frac{p}{h(p)}, \text{ onde}$$

M_p = média no teste dos sujeitos que acertam o item (p)

M = média total do teste

S = desvio padrão do teste

p = proporção de sujeitos que acertam o item

$h(p)$ = é a ordenada na curva normal no ponto de divisão dos segmentos que contêm as proporções **p** dos casos.

4.2 ANÁLISE GRÁFICA DOS ITENS

Esta nova técnica foi desenvolvida por T. A. van Batenburg e J. A. Laros (2001) e é baseada no pressuposto de que os construtores de itens devem conhecer muito bem o conteúdo ensinado e seus objetivos. Eles precisam de habilidades específicas para construir um bom item de múltipla escolha. Construir um item de múltipla escolha bom é uma tarefa complexa; o item deve ter uma – e somente uma – resposta correta, sem haver nenhuma discussão a esse respeito. As alternativas devem ser atrativas, mas não confusas. É importante não somente avaliar a dimensão de reconhecimento do que foi ensinado, mas também a dimensão de aplicação e de compreensão.

Os dois pressupostos válidos para essa análise são: a) um aluno que dá uma resposta certa em um item de múltipla escolha sabe mais que um aluno que dá a resposta errada; e b) um aluno que tem mais itens certos sabe mais que um aluno com menos itens certos.

Isso implica que aqueles que acertam todos os itens (o escore máximo) têm probabilidade 1 de terem marcado a alternativa correta; e aqueles que têm todos os itens errados, têm probabilidade 0 de terem marcado a alternativa correta. Entretanto, num caderno com 39 itens de múltipla escolha, com quatro alternativas, como é o caso das provas avaliadas nesse estudo, poucos alunos terão o escore 0, em razão da possibilidade de acerto ao acaso. Um aluno que somente "chuta" as questões terá uma chance de acertar, aproximadamente, dez questões (39 x 0,25). Assim, pode ser esperado que a proporção de acerto ao item aumente de 0 para 1 conforme vai aumentando o escore total. Acredita-se, também, que as alternativas falsas decresçam com o aumento do escore total. Até um certo escore, pode-se esperar que as alternativas certas e as falsas fiquem nos valores da chance de acerto ao acaso (0,25, neste caso). Depois deste escore total específico, a proporção de marcação da alternativa correta aumenta, e a proporção de marcação das alternativas falsas decresce. A análise da dificuldade do item pela AGI é realizada considerando-se a inclinação (*slope*). Na TRI, a dificuldade de um item é definida no ponto onde a linha de proporção 0,5 corta a "linha do item". Em uma abordagem visual isso é definido da mesma forma: a linha do item discrimina entre pessoas no intervalo acima das alternativas no máximo de 1 (um). Isto é chamado de intervalo de informação. Se a proporção de respostas para a alternativa correta aumenta rapidamente com o escore total, o item terá um alto poder discriminativo; caso contrário, será baixo.

No método gráfico usado, as proporções das alternativas dos itens estão sendo plotadas em contraposição ao escore total. Nas figuras a seguir, são apresentados exemplos de gráficos para a análise de itens.

Figura 1 – Item bom

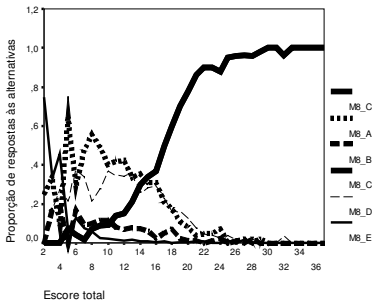


Figura 2 – Item ruim

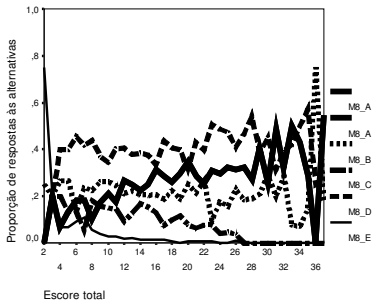


Figura 3 – Item muito difícil

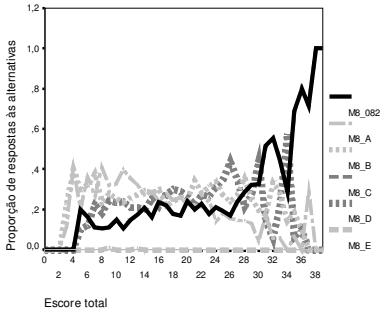
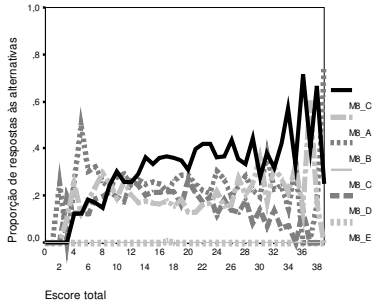


Figura 4 – Item com discriminação ruim



4.3 ANÁLISE FATORIAL

O modelo da análise fatorial está embasado no pressuposto de que uma série de variáveis observáveis pode ser explicada por um número menor de variáveis hipotéticas, não-observáveis, chamadas de fatores. Essas variáveis seriam a causa do fato de as variáveis observáveis se relacionarem entre si. Dessa forma, supõe-se que se as variáveis se relacionam entre si é porque elas têm uma causa comum que produz essa correlação. Tal causa chama-se fator e é do que a análise fatorial trata.

A relação entre cada item e o fator é expressa por meio da covariância ou correlação e é denominada carga fatorial. Esta mostra o grau com que cada item contribui para a mensuração do fator único. Itens que têm cargas mais altas no fator são considerados unidimensionais, pois estão medindo o mesmo fator. O critério mínimo da carga fatorial, citado na literatura, para que o item componha um mesmo fator, é 0,32 (Tabachnick, Fidell, 1996). Esse critério indica que a contribuição do item na composição do fator seria de aproximadamente 10%.

4.4 ANÁLISE DE ITENS PELA TRI

O modelo da psicometria moderna está fundamentado na Teoria de Resposta ao Item (TRI) que se relaciona ao modelo do traço latente ou da habilidade possuída. A idéia básica da TRI apóia-se em dois postulados fundamentais: a) o desempenho de um examinando em um teste pode ser predito ou explicado por fatores chamados traços latentes ou habilidades; e b) o relacionamento entre o desempenho de um examinando no item e os traços subjacentes ao desempenho no item pode ser descrito como uma

função monotonicamente crescente, chamada *função característica do item* ou *curva característica do item* (CCI). Esta função especifica que, à medida que o nível do traço ou da habilidade aumenta, a probabilidade de uma resposta correta ao item aumenta. Portanto, examinandos com valores mais altos no traço examinado têm probabilidades mais altas de responderem corretamente ao item do que estudantes com valores mais baixos no traço, independentemente do grupo a que pertencem (Hambleton, Swaminathan, Rogers, 1991).

Existem muitos modelos possíveis de resposta ao item que se diferem na forma matemática da função característica do item e/ou no número de parâmetros especificados no modelo. Todos os modelos de TRI contêm um ou mais parâmetros descrevendo o item e também um ou mais parâmetros descrevendo o examinando. Um dado modelo de TRI pode ou não ser apropriado para um conjunto particular de dados de um teste, isto é, o modelo pode não predizer ou explicar adequadamente os dados. Em qualquer aplicação da TRI, é essencial avaliar a adequação do modelo aos dados.

Quando um modelo de TRI é adequado aos dados do teste de interesse, várias características desejáveis são obtidas. As estimativas da habilidade dos examinandos não são dependentes do teste, e os índices não são dependentes do grupo. Estimativas de habilidade obtidas de diferentes conjuntos de itens serão as mesmas (exceto por erros de medida) e as estimativas dos parâmetros do item em diferentes grupos de examinandos serão também as mesmas (exceto por erros amostrais). Resumindo, tem-se que os parâmetros do item e da habilidade são invariantes, considerando-se uma escala única, e esta propriedade é obtida pela iteração da informação acerca do processo de estimação das habilidades dentro do processo de estimação dos parâmetros do item.

Os modelos matemáticos empregados na TRI pressupõem que a probabilidade de um examinando responder a um dado item corretamente depende de sua habilidade e das características do item. A TRI inclui um conjunto de pressupostos acerca dos dados para os quais o modelo será aplicado. Os dois principais pressupostos são o da *unidimensionalidade* e o da *independência local*. A unidimensionalidade supõe que somente uma habilidade esteja sendo medida pelos itens que compõem o teste. A independência local está relacionada ao conceito da unidimensionalidade e pressupõe que as respostas dadas aos itens dependem somente da habilidade que está sendo medida e não de outras habilidades. Assim, as respostas dos examinandos para qualquer par de itens deverão ser estatisticamente independentes. Para todos os modelos da TRI, a função característica do item deve refletir o relacionamento verdadeiro entre

variáveis não-observáveis (habilidades) e variáveis observáveis (respostas aos itens).

A *função característica do item* ou a *curva característica do item* é uma expressão matemática que relaciona a probabilidade de sucesso (dar uma resposta correta) em um determinado item, segundo a habilidade medida pelo teste e segundo as características do item. A escolha do número de parâmetros a serem usados no modelo envolve pressupostos acerca dos dados, e tais suposições podem ser verificadas mais tarde pelo exame de quão bem o modelo explica os resultados observados pelo teste. Os três modelos de TRI mais populares são os modelos logísticos de um, dois e três parâmetros.

No Saeb, é usado o modelo logístico de três parâmetros, que é dado pela expressão matemática:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad i = 1, 2, ..., n;$$

onde

$P_i(\theta)$ = probabilidade de um examinando com habilidade θ responder corretamente um item *i;*

a_i = parâmetro de discriminação do item*;*

b_i = parâmetro de localização do item;

c_i = parâmetro de pseudo-chance*;*

n = número de itens do teste;

e = é um número transcendental cujo valor aproximado é 2,718;

D = 1,7, que é um fator introduzido para tornar a função logística tão próxima quanto possível da função ogiva normal.

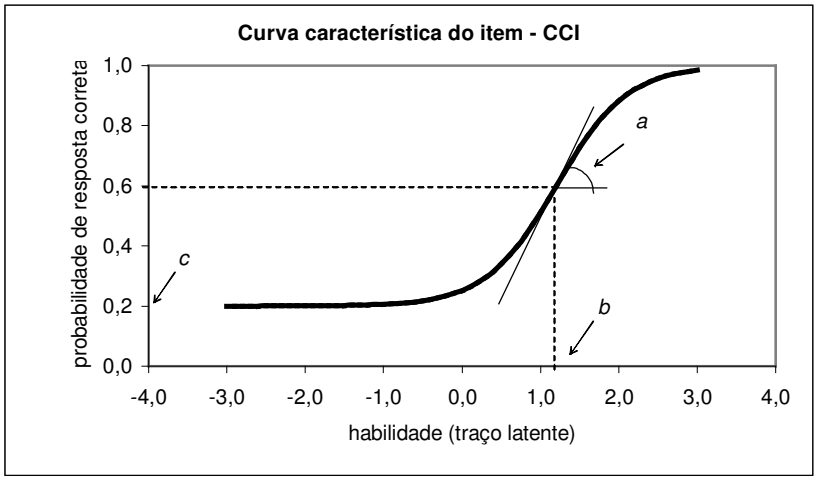
Os modelos da TRI permitem que, com base em informações indiretas sobre uma determinada característica não-observável do sujeito (traço latente, ou habilidade, ou *theta* – θ), se possa obter uma estimação da **localização** para cada indivíduo da amostra na escala. O **parâmetro *b*** para um item é o ponto correspondente na escala da habilidade em que a probabilidade de uma resposta correta é 0,5. Este é, portanto, um parâmetro de localização, indicado pela posição da CCI em relação à escala de habilidade. Quando os valores de habilidade de um grupo são transformados para a escala de média 0 e desvio padrão 1, os valores de *b*

normalmente variam de –3 a +3. Valores de *b* próximos de -3 correspondem aos itens que são muito fáceis e, ao contrário, valores de *b* próximos de +3 correspondem aos itens que são muito difíceis para o grupo de examinandos. Entretanto, o parâmetro *b* é influenciado pelo parâmetro *c* (acerto ao acaso). Se o modelo de três parâmetros for o escolhido, deve-se somar ao ponto 0,5, pois é nele que ocorre a probabilidade de 50% de uma resposta correta, ou seja, o valor do parâmetro *c* dividido por dois.

O parâmetro de **discriminação** do item é chamado **parâmetro** a_i. Este é proporcional à inclinação (*slope)* da CCI no ponto b_i na escala da habilidade. Itens com inclinações mais altas são mais úteis para separar examinandos com diferentes níveis de habilidades. Teoricamente, o parâmetro de discriminação do item pode assumir valores na escala de -∞ a +∞. Itens que apresentam valores negativos do índice de discriminação devem ser descartados. Os valores mais comuns do índice de discriminação dos itens variam entre 0 e +2.

O parâmetro c_i é o parâmetro da assíntota inferior do item e representa a probabilidade de examinandos com baixa habilidade responderem corretamente ao item. A seguir, a figura 5 mostra um exemplo da curva característica de um item.

Figura 5 – Exemplo de curva característica do item



Em um teste, cada item contribui com uma parcela significativa para o resultado final da avaliação. A análise de itens possibilita identificar aqueles que desempenham bem e aqueles que são problemáticos em relação à produção da informação desejada.

5 CONSIDERAÇÕES SOBRE A METODOLOGIA UTILIZADA PARA O ESTUDO

A proposta do presente estudo é apresentar um modelo de análise dos instrumentos construídos para avaliar o desempenho do aluno com base em análises pedagógicas e psicométricas dos itens de forma integrada, bem como, por meio deste modelo, verificar, ainda, a qualidade desses instrumentos que o Saeb utiliza para avaliar a educação básica brasileira.

Construir provas, apoiadas nas diretrizes curriculares da educação brasileira e nos propósitos norteadores da prática educacional, de forma que seja possível avaliá-las, constitui-se em um grande desafio. Portanto, este estudo está apoiado na premissa de que uma avaliação desse porte deve estar fundada nas mais modernas técnicas de avaliação e pautada por um extremo rigor científico.

A metodologia proposta para este estudo fundamentou-se nos procedimentos de análise que revelam esse nível de qualidade das provas e dos itens. Foram realizadas análises de cunho pedagógico, paralelamente às análises de cunho psicométrico, procurando-se, ao longo das interpretações, demonstrar que os dois tipos propostos não se bastam isoladamente. Além disso, se realizadas de forma integrada, podem revelar informações importantes que impactam o sistema educacional brasileiro. Para esses fins específicos, foram utilizados os bancos de dados coletados pelo Saeb, em 1997 e em 1999, para a disciplina de Matemática no nível da 8ª série.

As provas aplicadas pelo Saeb, a partir de 1995, adotaram o delineamento usado pelo sistema de avaliação norte-americano – *National Assessment of Educational Progress – NAEP*, chamado *Balanced Incomplete Blocks* (Blocos Balanceados Incompletos – BIB) – em espiral (Beaton, Johnson, Ferris, 1987). Os Blocos Balanceados Incompletos são uma variante de matriz amostral. Nem sempre é viável ou desejável que todos os itens do teste sejam administrados a todos os respondentes. Entretanto, muitas vezes, é necessário assegurar uma ampla e representativa cobertura do conteúdo da avaliação. Uma maneira pela qual tal representação é realizada é por meio do BIB. Em essência, significa que um conjunto completo de itens é dividido em um número menor de blocos. Os blocos

são, então, designados para os cadernos, de modo que cada bloco seja emparelhado com outro bloco para formar um caderno.

Características desejáveis da abordagem em espiral do BIB são aquelas em que: a) cada bloco apareça na mesma freqüência; b) efeitos da posição sejam controlados, pois cada bloco aparece uma vez em cada uma das três posições; e c) cada combinação de dois blocos apareça apenas uma vez em um caderno (Kirsch, Jungeblut, 1986).

Por meio desse delineamento, o total de itens ficou disposto em 13 blocos que, combinados de forma espiralada, compuseram 26 diferentes cadernos. Cada caderno da prova de 1997 teve, em sua composição, de 35 a 39 itens e os cadernos da prova de 1999 foram compostos por 39 itens, todos dispostos em três blocos. Nesse sistema, cada bloco apareceu seis vezes, e cada combinação de blocos apareceu somente uma vez. O total de itens aplicados em 1997, que foram objeto deste estudo, foi de 161; já, em 1999, o número de itens totalizou 169. Em ambas as edições, os itens foram distribuídos em 26 cadernos.

Na Tabela 1, a seguir, é apresentada a distribuição aproximada do número de respondentes por item, bloco e caderno para as edições de 1997 e 1999 na disciplina Matemática – 8ª série.

Tabela 1 – Distribuição aproximada do número de respondentes por item, bloco e caderno – 1997 e 1999

Matemática – 8ª série	1997	1999
Item	4.300	4.100
Bloco	1.480	1.380
Caderno	720	680
Total	18.806	17.890

Os procedimentos para a análise dos dados deste estudo seguiram os passos adiante especificados. Todas as análises propostas foram de cunho exploratório, buscando-se sempre apresentar os resultados psicométricos obtidos paralelamente à análise pedagógica de cada item. As análises pedagógicas das provas como um todo e as pedagógicas dos itens abrangeram apenas os itens construídos para cada prova, não considerando

os itens comuns[1]. Em 1997, com a exclusão dos itens comuns, o total de itens, construídos exclusivamente para as provas de Matemática – 8ª série, foi de 104, e, em 1999, de 117; tem-se que 57 itens foram comuns entre séries e anos na prova de 1997, e 52 tiveram esta mesma característica em 1999. A análise do nível de dificuldade das provas, bem como as análises psicométricas dos itens, individualmente, incluíram todos os itens apresentados nas provas (161 itens em 1997 e 169 em 1999).

1. Analisou-se **pedagogicamente a prova** como um todo, observando-se:
   1.1 a distribuição e a proporção de conteúdos abrangidos do total esperado;
   1.2 o nível de dificuldade dos itens que compuseram cada tema abrangido;
   1.3 a distribuição das competências exigidas para a resolução do item.

2. Analisou-se **pedagogicamente cada item**, por meio das seguintes observações:
   2.1 construção do enunciado, sua linguagem, ilustrações e nível de complexidade;
   2.2 plausibilidade dos distratores;
   2.3 coerência do gabarito;
   2.4 adequação entre o propósito do descritor e o item apresentado;
   2.5 adequação ou não para a série avaliada.

3. Examinou-se **psicometricamente os itens** por meio das seguintes análises:
   3.1 análise gráfica dos itens (AGI);
   3.2 análise da unidimensionalidade dos itens pela análise fatorial *full information*. As cargas fatoriais foram extraídas desta análise;
   3.3 análise da dificuldade e discriminação dos itens através da (TCT);
   3.4 análise dos três parâmetros da TRI (discriminação, localização e acerto ao acaso).

As análises pedagógicas e psicométricas realizadas permitiram a construção de uma tabela sumário com os índices gerados de todas as análises de cada item das provas de 1997 e 1999.

[1] O motivo da não-inclusão dos itens comuns deve-se à falta de informações do descritor e da competência correspondentes a cada item o que impossibilitaria a análise da distribuição dos conteúdos e das competências na prova.

Para a realização desses estudos propostos, as análises psicométricas foram feitas por meio dos seguintes softwares: *Statistical Package for the Social Sciences* (SPSS), Bilog - W, versão 3.0 e TESTFACT 2.0.

A seguir, há o modelo da tabela utilizada para análise da qualidade dos itens e da prova (Figura 6). O exemplo é apenas para ilustrar como foram reunidas as análises. Constam, nessa ilustração, três itens relacionados ao primeiro tópico de Matemática 1997.

Figura 6 – Modelo de tabela, contendo especificações pedagógicas e psicométricas para a análise de alguns itens da prova de Matemática – 8ª série – 1997

Especificações pedagógicas dos itens – 1997

Tema	Tópico	Descritor	Item	Competência	Construção do item*
Geometria	Formas – bidimensionais e tridimensionais (elementos e propriedades)	Classificar representação de figuras tridimensionais simples, de acordo com alguns critérios, como, por exemplo, número de faces, número de pontas, medida dos lados, formas arredondadas e não arredondadas.	1	Operacional	SP
		Comparar figuras bidimensionais e descrever propriedades a partir de suas representações.	2	Operacional	SP
			12		PG/PE/PA

Análise psicométrica dos itens – 1997

Item	AGI	Dificuldade	Bisserial	Análise Fatorial	Par. *a*	Par. *b*	Par. *c*
1	Bom	0,510	0,443	0,330	0,806	0,439	0,123
2	Bom	0,549	0,471	0,340	1,027	0,319	0,158
12	Ruim	Item anulado *a priori*					

* Legenda: SP- sem problemas; PG - problema no gabarito; PE - problema no enunciado; PA - problema nas alternativas; NAS - não adequado à série.

6 RESULTADOS E DISCUSSÃO

6.1 ANÁLISE PEDAGÓGICA E PSICOMÉTRICA DAS PROVAS COMO UM TODO

6.1.1 Distribuição dos Conteúdos nas Provas

A análise da distribuição das competências e de sua abrangência são muito importantes para as avaliações educacionais. Uma prova que contenha amostras insuficientes de competências relacionadas aos conteúdos ou temas para avaliar o conhecimento do aluno numa determinada disciplina estará com sua validade comprometida.

As Tabelas 2 e 3 apresentam a proporção de itens, por temas e tópicos, avaliados em Matemática – 8ª série, em 1997 e 1999.

Tabela 2 – Proporção do número de itens por temas

Temas	1997	1999
Números	0,39	0,37
Geometria e Medidas	0,38	0,48
Estatística	0,23	0,15

Verifica-se, pela Tabela 2, que a distribuição dos conteúdos foi abordada de modo diferenciado nas provas dos dois anos, tendo havido um privilégio do tema Geometria e Medidas no ano de 1999.

A Tabela 3 apresenta a proporção dos conteúdos por tópico. Observa-se que há uma maior concentração de itens no tópico "Figuras planas" para as provas dos dois ciclos. Esta ocorrência também indica uma falta de atenção na distribuição de conteúdos das provas avaliadas. Um fato curioso relatado por especialistas é que esse conteúdo, muitas vezes, não é abordado plenamente até o final do ano letivo, embora o devesse ser.

Tabela 3 – Proporção do número de itens por tópico

Temas	Tópicos	1997	1999
Números	Nos naturais e operações	0,05	0,04
	Nos inteiros e operações	0,13	0,07
	Nos racionais e operações	0,13	0,18
	Operações algébricas	0,09	0,08
Geometria e Medidas	Retas	0,04	0,04
	Ângulos	0,07	0,08
	Figuras planas	0,18	0,27
	Figuras tridimensionais	0,09	0,09
Estatística	Noções de proporcionalidade, porcentagem e juros	0,23	0,15

Considerando que as provas são construídas e aplicadas tendo por meta traçar uma radiografia do sistema educacional como um todo, o fato de ter havido uma distribuição não-proporcional de conteúdos e, além disso, privilegiando alguns deles, indica um comprometimento da validade da prova de Matemática de 1999, como instrumento avaliativo do processo ensino-aprendizagem do sistema educacional brasileiro.

A Tabela 4, a seguir, foi retirada do Relatório Saeb 1999 e mostra a relação entre o desenvolvimento do conteúdo curricular (informação obtida dos questionários aplicados aos professores) e o desempenho do aluno segundo a Escala de Desempenho do Saeb.

Tabela 4 – Desempenho médio dos alunos por disciplina e série, segundo o desenvolvimento dos conteúdos curriculares em sala de aula

Disciplinas	Série	Menos da metade (<50%)	Um pouco mais da metade (50% a 79%)	Quase todo (80% a 99%)	Todo o conteúdo (100%)
Língua Portuguesa	4ª EF	159,82	163,05	175,54	189,73
	8ª EF	222,87	227,87	237,16	247,28
	3º EM	259,72	259,20	270,04	284,14
Matemática	4ª EF	170,38	173,99	186,32	213,36
	8ª EF	236,55	239,64	252,41	261,43
	3º EM	271,61	271,76	284,62	303,33

Fonte: MEC/INEP/DAEB, 1999.

As escalas de desempenho são utilizadas desde 1995 pelo Saeb para descrever e comparar o desempenho dos alunos brasileiros nas disciplinas e séries avaliadas. Essas escalas variam de 0 a 500 pontos e o desempenho é apresentado em ordem crescente e cumulativa. A estimativa de desempenho obtida para os alunos ou grupos de alunos em cada uma das três séries avaliadas indica, portanto, o lugar que eles ocupam na escala. Em 1997 e 1999, os alunos da 8ª série encontravam-se, em média, no *"Nível 225 a ≤ 275"*, o qual se caracteriza pelo domínio das seguintes habilidades: resolver as quatro operações com números naturais, identificar elementos das figuras geométricas, interpretar gráficos e tabelas, ler medidas de temperatura, estabelecer relações entre diversas unidades de tempo e manipular o sistema monetário.

O principal objetivo da escala de desempenho do Saeb é fornecer uma forma de interpretação do desempenho do aluno que descreva o que os alunos sabem e são capazes de fazer em determinados intervalos da escala, denominados níveis de desempenho. Esse tipo de interpretação favorece a análise da evolução do desempenho entre os diferentes ciclos de avaliação, uma vez que torna possível a interpretação pedagógica de todos os pontos da escala (Klein, 2003).

A Tabela 5, indica que há uma relação direta e significativa, a um nível de confiança de 95%, entre o desenvolvimento dos conteúdos e o desempenho dos alunos. Dessa forma, constata-se que, à medida que o percentual de conteúdo previsto para ser desenvolvido durante o ano letivo aumenta, o desempenho dos alunos também aumenta.

6.1.2 O Nível de Dificuldade das Provas

O nível de dificuldade dos itens que compõem uma prova de avaliação de sistema deve obedecer ao critério de equilíbrio: os itens de um mesmo *continuum* devem cobrir todos os seus segmentos em termos de dificuldade (fáceis, médios ou difíceis), e essa distribuição deve ter como base a curva normal (Pasquali, 1997). A seguir, nas figuras 7 e 8, é apresentada a distribuição dos índices de dificuldade nas duas provas completas.

Figura 7

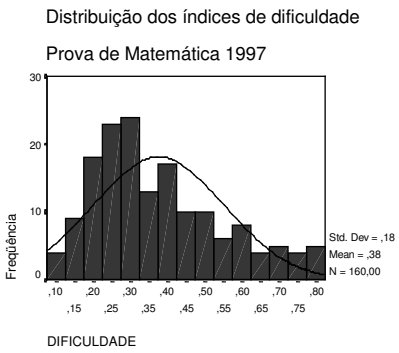
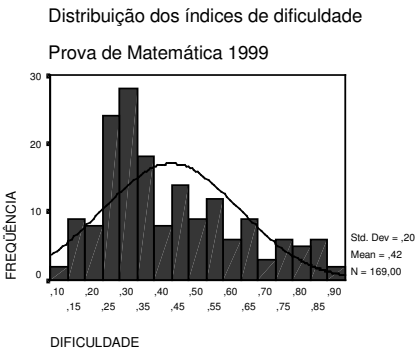


Figura 8



A dificuldade média de todos os itens de 1997 foi de 0,38, enquanto a de 1999, foi de 0,42. Observa-se que as provas se apresentaram bastante difíceis. Pelo fato de serem provas para avaliação de um sistema, os altos índices de dificuldade passam a representar um fator negativo para a prova do Saeb. A literatura tem indicado que um nível de dificuldade médio de 0,50 é o ideal para esse tipo de prova, pois se a variância for pequena a fidedignidade da prova será reduzida e, conseqüentemente, os resultados também serão menos fidedignos.

Outra análise realizada foi a da dificuldade média dos itens nos tópicos. Os grupos de itens por tópico referem-se apenas àqueles que continham a informação do descritor. A Tabela 5 apresenta esses dados.

Tabela 5 – Índice de dificuldade média, segundo o tópico

Temas	Tópicos	1997	1999
Números	Nos naturais e operações	0,35	0,44
	Nos inteiros e operações	0,39	0,39
	Nos racionais e operações	0,32	0,32
	Operações algébricas	0,36	0,35
Geometria e Medidas	Retas	0,30	0,36
	Ângulos	0,41	0,39
	Figuras planas	0,27	0,31
	Figuras tridimensionais	0,27	0,35
Estatística	Noções de proporcionalidade, porcentagem e juros	0,31	0,45

Os dados apresentados na Tabela 5 indicam que há um nível médio de dificuldade pouco variável entre os tópicos. Nota-se que os alunos submetidos ao Saeb de 1999 apresentaram igual ou maior dificuldade em relação aos de 1997 nos tópicos: "Números inteiros e operações", "Números racionais e operações", "Operações algébricas" e "Ângulos". Essa informação pode estar revelando problemas no processo ensino-aprendizagem em relação a esses tópicos em específico, pois a proporção de acertos nos tópicos citados é baixa.

A Tabela 6 apresenta a porcentagem de itens, segundo o tema, com índice de dificuldade menor que 0,30, ou seja, menos de 30% dos alunos respondem ao item corretamente. Em números absolutos, apurou-se que a prova de 1997 apresentou 58 itens com índices menores que 0,30, enquanto a prova de 1999 apresentou 55 itens com tal característica.

Tabela 6 – Porcentagem do número de itens com índice de dificuldade menor que 30,0, segundo o tema

Temas	1997	1999
Números	41,0	44,0
Geometria e Medidas	59,0	51,0
Estatística	54,0	41,0
Prova total	51,0	47,0

6.1.3 Distribuição das Competências Exigidas para a Resolução do Item

Foram considerados, para análise, os três níveis de competências cognitivas: o nível básico, o nível operacional e o nível global. Os itens relacionados às competências do nível básico requerem habilidades como lembrar e reconhecer noções e operações básicas; os itens das competências do nível operacional exigem que o aluno compreenda, explique e relacione os conceitos matemáticos aprendidos para aplicá-los a situações cotidianas e práticas da vida; os itens das competências do nível global são aqueles que requerem habilidades de maior complexidade na busca da solução dos problemas.

A Tabela 7 apresenta a distribuição das competências, segundo os temas e tópicos, das provas de Matemática dos dois anos avaliados.

Tabela 7 – Distribuição proporcional do número de itens, por competências, segundo os temas e tópicos – Matemática – 1997/1999

Temas/Tópicos	Ano	Competências		
		Básica	Operacional	Global
Número e operações	1997	0,15	0,68	0,17
	1999	0,28	0,70	0,02
Geometria e medidas	1997	0,23	0,62	0,15
	1999	0,23	0,67	0,11
Estatística	1997	--	0,88	0,12
	1999	--	0,88	0,12
Total	1997	0,15	0,70	0,15
	1999	0,21	0,71	0,08

Observa-se que a distribuição por competências aparece um pouco mais equilibrada em 1997 do que em 1999. O nível de competência operacional foi o mais privilegiado nas provas, o que é uma prática comum.

6.2 ANÁLISES PEDAGÓGICA E PSICOMÉTRICA DOS ITENS DAS PROVAS

Os itens devem ser analisados com rigor em seus dois campos de análise possíveis: o pedagógico e o psicométrico. Um não deve ser mais privilegiado do que o outro. Ao contrário, eles devem complementar-se. Essas análises têm por objetivo avaliar a validade dos itens.

6.2.1 Aspectos Pedagógicos dos Itens

A respeito da construção do item, foram analisados aspectos de conteúdo e da forma. Considerando a natureza do conteúdo, avaliou-se se o item conseguiu atingir o objetivo proposto. Nesse aspecto específico, constatou-se que, tanto para a prova do ano de 1997 quanto para a de 1999, em sua grande maioria, os itens alcançaram os objetivos propostos. Percebeu-se que alguns, embora tivessem cumprindo o seu papel, não se adequaram à série em questão ou ao desenvolvimento cognitivo esperado para o aluno nesse nível. Levanta-se a hipótese de que elaboradores de itens, ao construí-los, não consideraram a amplitude do sistema educacional brasileiro e o desenvolvimento mental do aluno em cada nível

de escolaridade. O aluno, no nível pesquisado (8ª série), tem uma idade média de 14 anos, o que pressupõe que ele esteja num processo de maturação cognitiva, não tendo atingido, ainda, a plenitude do pensamento formal (Piaget, 1967). Além do problema da maturação, tem-se também o problema da influência do conhecimento prévio na resolução de problemas. Muitas vezes, no entanto, esse conhecimento ainda não está consolidado, de forma que o aluno não consegue alcançar o sucesso esperado na solução da questão. Na análise realizada, itens com problemas dessa natureza foram considerados não adequados à série.

Outro ponto levado em consideração no que se refere à análise da construção do item foi o aspecto formal. Aqui, o item foi analisado considerando-se o enunciado, as alternativas (gabarito e distratores) e as ilustrações. Na Tabela 8, é exibida a porcentagem do número de itens que apresentaram problemas pedagógicos.

Tabela 8 – Porcentagem do número de itens que apresentam problemas pedagógicos

Problemas pedagógicos dos itens	1997	1999
No enunciado	2,0	2,0
Nas alternativas	2,0	1,0
Nas ilustrações	1,0	2,0
Não adequados à série	12,0	5,0

Em relação ao enunciado, observou-se se cada item abordou apenas um problema. Em princípio, este deveria ser bem formulado de modo que o aluno, apenas lendo-o, fosse capaz de raciocinar sobre a resposta sem depender da leitura de todas as alternativas. Além disso, verificou-se se o item expressava um comportamento, e não uma abstração, permitindo ao sujeito uma ação clara e precisa do que ele deveria fazer.

Outro aspecto importante relacionado ao enunciado é o que diz respeito à linguagem e à simplicidade em sua formulação. Foi observada a clareza de linguagem, a objetividade e a simplicidade na forma de composição do problema. O item deve apresentar um equilíbrio formal-estrutural para atingir todos os estratos da população-alvo, sem prejuízo ou privilégio para qualquer parte deles.

De um modo geral, houve poucos problemas relacionados à construção do enunciado. O problema mais comum foi a falta de clareza, pois é difícil para o elaborador de itens colocar-se no lugar do respondente. Ele costuma seguir a sua própria lógica. Outro problema que surgiu foi a falta de objetividade na exposição da questão, levando à interpretação dúbia.

Em relação às alternativas, observou-se: 1) a coerência da estrutura e do tamanho; 2) a plausibilidade dos distratores, ou seja, se elas mantiveram um grau de racionalidade com o enunciado; e 3) a clareza do gabarito, não dando chance ao aluno que sabe ficar em dúvida com a resposta.

Houve também poucos problemas relativos às alternativas. Os problemas mais freqüentes foram em relação à estrutura e ao tamanho. Algumas alternativas tornavam-se atrativas em função do tamanho; outras exigiam mais operações mentais em razão de sua estrutura. Por exemplo, questões que em suas alternativas apresentam, ao mesmo tempo, operações mentais de naturezas diferentes, tornam-se mais difíceis para o aluno. Ao contrário, essas devem se apresentar simples e diretas.

Constatou-se que houve poucos problemas relacionados à poluição visual causada pelas ilustrações que, às vezes, em vez de ajudar, atrapalham.

A maior proporção de problemas foi com relação à adequação do item à série avaliada (12%). Este é um problema sério numa avaliação de larga escala, pois pode comprometer os resultados. Alguns especialistas acreditam que isso ocorre em razão da falta de prática, em sala de aula, dos elaboradores dos itens e do mau planejamento na montagem da prova. Entretanto, quando esse planejamento é realizado de maneira cuidadosa a não-adequação é detectada pelos próprios resultados psicométricos. Outra questão a ser notada relaciona-se ao conhecimento da fase de desenvolvimento cognitivo em que o aluno se encontra. É comum a elaboração de itens que exigem um nível de abstração para o qual eles ainda não têm amadurecimento suficiente para compreendê-los e respondê-los.

6.2.2 Aspectos Psicométricos dos Itens

6.2.2.1 Análise gráfica dos itens

A Análise Gráfica dos Itens (AGI) dispõe de recursos visuais, em que é apresentada a relação entre o escore total e as porcentagens de

respostas às alternativas verdadeiras e falsas dos itens. Essa análise permitiu identificar: bons itens; itens extremamente difíceis; itens que apresentam uma ou mais alternativas falsas e mantêm um aumento da porcentagem de respostas com o aumento do escore total (problema de discriminação); e itens cujas alternativas verdadeiras apresentam um decréscimo na porcentagem de respostas em relação ao aumento do escore total (itens ruins). O principal pressuposto dessa análise, segundo van Batenburg e Laros (2001), é: "a proporção da alternativa correta deve aumentar com um aumento do escore total, e a proporção de alternativas falsas deve decrescer com um aumento do escore total".

Os resultados dessa análise, que têm por base o escore total, retratam a tendência real dos alunos quando respondem ao item. Esses resultados radiografam a realidade. Quando a análise de um item revela que houve uma dispersão nas respostas às alternativas, não significa que o problema seja da construção do item, mas pode estar indicando uma falta coletiva de conhecimento de determinado assunto abordado neste item. Assim, é importante que, em conjunto com a AGI, seja realizada uma análise pedagógica desses itens. Essas análises poderão dar indicativos do processo mental utilizado para a solução da questão, associando-se o escore total e as respostas aos distratores. A Tabela 9 apresenta a porcentagem do número de itens distribuídos em cada categoria, considerando-se essa análise.

Tabela 9 – Porcentagem do número de itens por categoria de qualidade psicométrica, com base na AGI

Qualidade Psicométrica do Item	1997	1999
Bom (sem problema)	69,0	85,0
Ruim	4,0	7,0
Difícil	2,0	4,0
Com baixa discriminação	7,0	4,0

Observa-se que, na prova de 1997, 69% dos itens apresentaram um bom comportamento, enquanto em 1999 a porcentagem foi de 85%. Os demais itens, nos dois anos avaliados, apresentaram algum tipo de problema.

6.2.2.2 Análise da dificuldade dos itens

A dificuldade dos itens, baseada na TCT, é calculada com base na percentagem de examinandos que respondem corretamente a um dado item. Associando-se os índices gerados por essa análise às informações pedagógicas do item, podem-se obter dados que mostram onde os alunos estão mais defasados, em termos das competências que deveriam ter construído.

O estudo realizado aponta que, na prova de Matemática de 1997, os descritores que apresentaram maior número de itens com índices de dificuldade inferiores a 30% foram os seguintes: "Utilizar as relações métricas no triângulo retângulo (Teorema de Pitágoras), para solucionar problemas" (Descritor 38, do tema "Geometria e Medidas", do tópico "Figuras planas") e "Solucionar situações-problema analisando informações apresentadas em tabelas e gráficos mais usuais" (Descritor 64, do tema "Estatística", do tópico "Noções de proporcionalidade, porcentagem e juros"). O descritor 38 apresentou três itens com dificuldades que variaram de 12,0 a 29,0, e o descritor 64 três itens com dificuldades entre 15,0 e 23,0.

Em relação aos itens da prova de Matemática, de 1999, o descritor que teve maior número de itens com índices menores que 30,0 foi "Aplicar a noção de área de figuras planas como triângulo, paralelogramo e trapézio" (Descritor 19, do tema "Geometria e Medidas", do tópico "Figuras planas"). Este descritor apresentou seis itens com dificuldades que variaram de 10,0 a 29,0. A Tabela 10 mostra o número de itens mais difíceis (índices menores que 30,0) por tópico.

Tabela 10 – Distribuição do número de itens com índice de dificuldade menor que 30,0, por tópico

Temas	Matemática 1997		Matemática 1999	
	Tópicos	N de itens		N de itens
Geometria	Retas	03	Retas no plano	03
	Ângulos	01	Ângulos	01
	Figuras planas	14	Figuras planas	18
	Figuras espaciais	06	Figuras tridimensionais	07
Números	Números naturais e operações – inteiros, racionais e reais	01	Números naturais e operações – inteiros, racionais e reais	01
	Números inteiros e operações	01	Números inteiros e operações	02
	Números racionais e operações	09	Números racionais e irracionais e operações	12
	Operações algébricas	06	Operações algébricas	04
Estatística	Noções de proporcionalidade, porcentagem e juros	13	Noções de proporcionalidade, probabilidade, porcentagem e juros	03

A comparação da dificuldade de tais itens entre esses anos (1997 e 1999) mostra que, de modo geral, ela se repete nos mesmos conteúdos. Considera-se esse dado de extrema relevância, pois os itens construídos para medir essas habilidades são considerados bons pedagogicamente; no entanto, se os alunos não conseguem resolver essas questões, podem ser levantadas hipóteses que vão desde a falta de informações básicas dos alunos para processarem elementos mais elaborados, passando pela imaturidade dos mesmos para o desenvolvimento dessas habilidades, até a falta de domínio desses conteúdos por parte dos professores.

Uma vez que os itens analisados compõem as provas de avaliação de um sistema nacional, e são constatadas recorrências de dificuldades em determinadas áreas, alguma intervenção pode ser feita. Por exemplo, com a adoção de políticas públicas que envolvam maior investimento na formação de professores, enfocando determinados conteúdos e, até, políticas educacionais que implementem uma extensão do tempo destinado às aulas de Matemática.

6.2.2.3 Análise da discriminação dos itens

A análise da discriminação dos itens foi realizada considerando-se os índices da correlação bisserial, ou seja, a correlação item-total, para cada item. Os resultados mostraram que esses coeficientes apresentaram-se ruins quando havia algum problema com a construção do item, ou quando o conhecimento exigido para solucionar a questão não era de domínio de quem, supostamente, o sabia. Na prova de 1997, 10 itens apresentaram correlação bisserial menor que 0,20 e, na prova de 1999, 16 itens apresentaram esse intervalo.

6.2.2.4 Análise fatorial dos itens

A análise fatorial mostra o grau com que cada item contribui para a mensuração do fator único. Itens com cargas fatoriais menores que 0,32 não contribuem para a unidimensionalidade da prova (Tabachnick, Fidel, 1996). Tal fato implica na exclusão desses itens do conjunto. As cargas apresentadas foram geradas por meio da análise fatorial *full information*.

Na prova de 1997, 25 itens apresentaram cargas fatoriais menores que 0,32, enquanto a de 1999 apresentou 24 itens com tal característica. O item que apresentou a carga fatorial mais alta, na prova de 1997, foi 0,76, enquanto na prova de 1999 foi 0,77.

6.2.2.5 Análise dos parâmetros da TRI

Conforme já foi dito anteriormente, a análise da TRI está baseada no pressuposto de que o desempenho de um examinando em um teste pode ser predito ou explicado por um conjunto de fatores chamados traços latentes ou habilidades. Estes devem refletir o relacionamento verdadeiro entre variáveis não observáveis (habilidades) e variáveis observáveis (respostas aos itens). A análise da TRI é importante porque a unidade de análise é o item e não o teste, como na TCT. Na TRI, os parâmetros do item e da habilidade são considerados invariantes.

Observou-se que os resultados apresentados pela TRI coadunam-se melhor com a análise pedagógica dos itens. Todas as análises realizadas fornecem indicadores importantes da qualidade dos itens dentro de suas especificidades, mas é a análise dos parâmetros da TRI que reflete melhor as especificações pedagógicas do item.

Os critérios adotados para o julgamento dos parâmetros da TRI foram os seguintes: para o parâmetro "a", itens com índices abaixo de 0,60 foram considerados com discriminação ruim; para o parâmetro "b", considerando-se a população pesquisada (8ª série), itens com índices acima de 2,00 foram considerados mais difíceis e abaixo de –2,00, mais fáceis, podendo, no entanto, ocorrer; para o parâmetro "c", itens com índices maiores que 0,30 foram considerados como aqueles que possuem alta probabilidade de acerto ao acaso. A Tabela 11 mostra a porcentagem do número de itens que apresentaram problemas associados a cada parâmetro e prova.

Tabela 11 – Porcentagem do número de itens que apresentam problemas em parâmetros da TRI

Edição	Parâmetro *a*	Parâmetro *b*	Parâmetro *c*
1997	4,0	17,0	4,0
1999	7,0	14,0	5,0

Nota-se que, de um modo geral, os maiores problemas surgiram em torno do parâmetro "b".

Ele é um parâmetro de localização ou de dificuldade do item, que indica a posição da CCI em relação à escala de habilidade. Quanto maior o valor do parâmetro "b", maior a habilidade requerida para que um

examinando dê uma resposta correta e, ao contrário, quanto menor o valor do parâmetro "b", menor a habilidade requerida para o examinando acertar o item. Os resultados da análise desse parâmetro ratificam os anteriores. Tanto a análise pedagógica de adequação dos itens à série quanto a dos índices de dificuldade gerados pela TCT já haviam detectado esse problema: os itens, de maneira geral, se apresentaram difíceis para a população avaliada. Os resultados apresentados na Tabela 11 referem-se aos parâmetros maiores que 2,00, pois não foram apresentados itens com índices menores que -2,00. Dessa forma, constata-se que um número significativo de itens exigia um elevado grau de proficiência para a sua resolução.

O parâmetro "a" é chamado parâmetro de discriminação e é proporcional à inclinação da CCI no ponto b_i da escala da habilidade. Itens com inclinações mais altas são mais úteis para discriminar os diferentes níveis de habilidade dos examinandos. Na prova de 1997, apenas 4% dos itens não apresentaram um alto poder de discriminação. Na prova de 1999, esse percentual foi de 7%.

O parâmetro "c" representa a probabilidade de examinandos com baixa habilidade responderem corretamente ao item. A porcentagem de itens com essa característica, na prova de 1997, foi de 4%. Na prova de 1999, o percentual foi de 5%.

CONSIDERAÇÕES FINAIS

A cultura da avaliação educacional no Brasil está, ainda, em fase de desenvolvimento, contudo tem alcançado importantes conquistas. O Ministério da Educação, responsável pela condução da política educacional do país, vem implementando sistemas de avaliação nos diversos níveis de ensino, como o Sistema Nacional de Avaliação do Ensino Superior (Sinaes), o Exame Nacional de Estudantes do Ensino Médio (Enem), e o Exame Nacional de Certificação de Competências para o Ensino de Jovens e Adultos (Encceja). Em conseqüência, vários Estados e alguns municípios, consoante orientação técnica dos órgãos centrais, vêm adotando seus próprios sistemas de avaliação. Esta iniciativa traz, em seu bojo, a mudança do foco da investigação sobre as políticas educacionais praticadas em todos os níveis da educação brasileira. Isto alavanca as mudanças de rumo da educação e alinha o Brasil com os países que já desenvolvem tecnologias educacionais de ponta.

O Saeb, como sistema responsável pela avaliação da educação básica brasileira, tem se esforçado para disseminar essa prática da maneira

mais competente possível. Como já foi exposto anteriormente, a sua função é obter dados sobre a qualidade do ensino ao longo do tempo, identificar os fatores que contribuem para a ocorrência dos resultados e intervir no sistema educativo, visando à melhoria da qualidade da educação básica do Brasil. Para alcançar esse intento, vários instrumentos são desenvolvidos. Dentre todos, o que avalia o desempenho dos alunos – a prova – constitui-se no mais importante, pois é ela que fornece informações sobre o estágio de desenvolvimento dos estudantes.

Partindo-se da premissa de que a construção e a análise desses instrumentos garantirão a fidelidade dos dados informados a respeito da realidade educacional brasileira, as provas têm, portanto, o dever de comprovar a sua objetividade, confiabilidade e qualidade. O presente estudo procurou demonstrar que um modelo de análise fundamentado em aspectos psicométricos e pedagógicos integrados poderá ser uma importante e confiável referência de informações da qualidade da educação brasileira.

O Saeb dispõe de dados que subsidiam uma escolha mais adequada de itens; no entanto, as análises apontam que deve-se tomar mais cuidado no que se refere à distribuição de conteúdos. Problemas como os detectados neste estudo, que revelam que cerca de 49% dos itens da prova de Matemática de 1999 – 8ª EF – contemplaram o tema "Geometria", devem ser evitados, tendo em vista a concentração de um só tema, problema que se torna ainda mais grave, pois as habilidades relacionadas a tal tema são pouco desenvolvidas em sala de aula.

Outro problema a ser evitado, e que pode ser constatado antes da montagem das provas, é com relação ao nível médio de dificuldade. A prova de 1997 apresentou um nível médio de 38%, e a prova de 1999 de 42%. Os altos níveis de dificuldade constituem-se em fator negativo para as provas de avaliação de sistemas, pois esses testes mostraram-se muito difíceis para a população amostrada. A literatura tem indicado que os níveis médios ideais de dificuldade devem estar em torno de 50,0 a 60,0, garantindo uma maior variabilidade.

Uma análise que dá bastante informação, e que não é de uso corrente na avaliação do Saeb, é a Análise Gráfica de Itens. Essa análise, juntamente com a análise bisserial das alternativas de cada item e a análise pedagógica dos distratores, dá pistas sobre os processos cognitivos utilizados pelos alunos para responderem ao item, e pode fornecer subsídios para discussões pedagógicas.

Ao final da pesquisa, constatou-se que estudos a respeito do desenvolvimento cognitivo do aluno para a compreensão do comportamento de respostas aos itens das provas devem ser levados em

conta. Esses aspectos precisam ser observados desde a construção do item até a análise de seus resultados. Não basta que ele apresente todas as características estruturais de um bom item. É imperativo ter em mente para quem este item está sendo construído.

Retomando o objetivo do Saeb, espera-se que o modelo de análise de provas apresentado neste estudo – Modelo Integrado das Análises Pedagógicas e Psicométricas – contribua para a fidelidade dos dados que procuram retratar a realidade educacional observada e as informações disseminadas.

REFERÊNCIAS BIBLIOGRÁFICAS

ANASTASI, A.; URBINA, S. *Testagem Psicológica*. Porto Alegre: Artmed, 2000

BEATON, A. E.; JOHNSON, E. G.; FERRIS, J. J. The assignment of exercises to students. In: BEATON, A. E. *Implementing the new design*: the NAEP 1983-1984 technical report. Princeton, NJ: Educational Testing Service, 1987. p.97-118.

BLOOM, B. S.; HASTINGS, J. T.; MADAUS, G. F. *Handbook on Formative and Summative Evaluation of Student Learning*. New York: McGraw-Hill,1971.

BOCK, R. D.; GIBBONS, R.; MURAKI, E. Full-information item factor analysis. *Applied Psychological Measument*, n.12, p. 261-280, 1988.

HAMBLETON, R. K.; SWAMINATHAN, H. *Item Response Theory*: Principles and Applications. Boston: Kluwer. Nijhoff Publishing, 1995.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H. J. *Fundamentals of Item Response Theory*. Newbury Park, California: Sage Publications, 1991.

KIRSCH, I. S.; JUNGEBLUT, A. *Literacy*: Profiles of American's young adults. Princeton, NJ: Educational Testing Service, 1986.

KLEIN, R. Utilização da Teoria de Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica (Saeb). *Ensaio*: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 11, n. 40, p. 283-296, jul./set. 2003.

LAROS, J. A.; PASQUALI, L.; RODRIGUES, M. M. M. *Análise da unidimensionalidade das provas do Saeb*. Brasília: Centro de Pesquisa em Avaliação Educacional. Universidade de Brasília, 2000. (Relatório Técnico)

MARSHALL, J. C.; HALES, L. W. *Essentials of testing*. Reading, M. A.: Addinson-Wesley, 1972.

McINTIRE, S. A.; MILLER, L. A. *Foundations of Psychological Testing*. New York: McGraw-Hill, 2000.

MURAKI, E.; ENGELHARD, G. Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement*, n. 9, p. 417-430, 1985.

NUNNALY, J. C.; BERNSTEIN, I. H. *Psychometric Theory*. 3.ed. New York: McGraw-Hill, 1994.

PASQUALI, L. *Psicometria*: teoria e aplicações. Brasília: Editora da Universidade de Brasília, 1997.

_____. *Psicometria*: teoria dos testes na psicologia e na educação. Petrópolis: Vozes, 2003.

_____. *Instrumentos psicológicos*: manual prático de elaboração. Brasília: LabPAM/IBAPP, 1999.

PESTANA, M. I. G. S. et al. *Matrizes Curriculares de Referência para o Saeb*. Brasília: MEC/Inep, 1997.

_____. *Matrizes Curriculares de Referência para o Saeb*. 2.ed. Rev. Ampl. Brasília: MEC/Inep, 1999.

PIAGET, J. *Biologia e Conhecimento*: ensaio sobre as relações entre as regulações orgânicas e os processos cognoscitivos. Petrópolis: Vozes, 1973

RODRIGUES, M. M. M. *Instrumentos de avaliação educacional*: uma visão pedagógica e psicométrica integradas – estudo das provas do Saeb. Brasília, 2001. Dissertação (mestr.) em Psicometria. Instituto de Psicologia/ Universidade de Brasília.

BRASIL. Ministério da Educação e Cultura. *Sistema Nacional de Avaliação da Educação Básica*: Saeb 1995; relatório técnico. São Paulo: Fundação Carlos Chagas; Rio de Janeiro: Fundação Cesgranrio, 1996.

BRASIL. *Sistema Nacional de Avaliação da Educação Básica*: Saeb 1997; relatório técnico. São Paulo: Fundação Carlos Chagas; Rio de Janeiro: Fundação Cesgranrio, 1998.

_____. *Sistema Nacional de Avaliação da Educação Básica*: Saeb 1999; relatório técnico. São Paulo: Fundação Carlos Chagas; Rio de Janeiro: Fundação Cesgranrio, 2000.

_____. Sistema Nacional de Avaliação da Educação Básica: Saeb 2001. *Novas Perspectivas*. Brasília: MEC/Inep/DAEB, 2002.

TABACHNICK, B. G.; FIDEL, L. S. *Using multivariate statistics*. New York: Harper Collins, 1996.

VAN BATENBURG, T. A.; LAROS, J. A. Graphical Analysis of Test Items. In: *Educational Research and Evaluation (An International Journal on Theory and Practice)*. Lisse: Swets e Zeitlinger, 2001.

VAN DER LINDEN, W. J.; HAMBLETON, R. K. *Handbook of Modern Item Response Theory*. New York: Spring-Verlag, 1997.

VIANNA, H. M. *Testes em Educação*. São Paulo: Ibrasa, 1982.

_____. *Introdução à Avaliação Educacional*. São Paulo: Ibrasa, 1989.

WILSON, D. T.; WOOD, R.; GIBBONS, R. *TESTFACT*: test scoring, item statistics, and item factor analysis. Chicago: Scientific Software, 1991.

Recebido em: março 2006.

Aprovado para publicação em: maio 2006