Eficácia dos processos de linkagem na avaliação educacional em larga escala

WELLINGTON SILVA*

TUFI MACHADO SOARES**

RESUMO

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

Palavras-chave: Teoria da Resposta ao Item (TRI), Avaliação da educação, Escala de avaliação, Métodos de avaliação.

* Coordenador de Medidas Educacionais do Centro de Políticas Públicas e Avaliação da Educação – CAEd da Universidade Federal de Juiz de Fora (UFJF) (wellington@caed.ufjf.br).

** Professor do Programa de Mestrado e Doutorado em Educação e Coordenador de Pesquisa do Centro de Políticas Públicas e Avaliação da Educação – CAEd da Universidade Federal de Juiz de Fora (UFJF) (tufi@caed.ufjf.br).

RESUMEN

En 1997, por medio del Sistema Nacional de Evaluación de la Educación Básica (Saeb), se difundió la escala de proficiencia para Brasil. A partir de entonces, practicamente todas las evaluaciones a gran escala realizadas han intentado mantener una comparabilida de resultados con esa escala, por intermedio de la Metodología de la Teoría de Respuesta al Ítem (TRI). Sin embargo, se observa una diversidad de situaciones al analizar las diferentes evaluaciones realizadas por los diferentes Estados brasileños e incluso hasta en el propio Saeb. En este artículo presentaremos algunos aspectos técnicos necesarios para garantizar posibilidad de establecer comparaciones en los procedimientos de linkage de evaluaciones, así como las características de las evaluaciones del Saeb y de algunos Estados brasileños a lo largo del tiempo.

Palabras clave: Teoría de Respuesta al Ítem (TRI), Evaluación de la educación, Escala de evaluación, Métodos de evaluación.

ABSTRACT

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

Keywords: Item Response Theory (*IRT),* Education Assessment, Assessment scale, Assessment methods.

1 INTRODUÇÃO

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

Para estudar tal proposta, foi contratado o National Research Council (NRC), o qual produziu, em 1999, o relatório *Uncommon Measures: equivalence and linkage among educational tests*.

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

Este caso americano, com suas recomendações técnicas, será utilizado neste artigo para estudarmos e refletirmos sobre diversas situações de linkagem dos sistemas de avaliações aplicados no Brasil.

Um ponto forte e que diferencia a realidade brasileira da americana é o fato de, no Brasil, já existir uma cultura de escala única referenciada ao Sistema Nacional de Avaliação da Educação Básica – Saeb. Outro fator importante, que viabiliza essa escala única, é o fato de as avaliações realizadas, até então, pelos diversos Estados brasileiros, terem mantido a mesma matriz de referência com o Saeb, o que é bem diferente da realidade americana, em que cada Estado tem autonomia para elaborar sua própria matriz. Esses dois fatores, alinhados a diversos sistemas de avaliações adotados por diferentes Estados brasileiros, propiciam a adoção de uma escala única atrelada ao Saeb.

Neste estudo, focaremos os aspectos técnicos necessários para garantir qualidade e confiabilidade nas linkagens e, com isso, não perdermos os benefícios de termos uma escala única de proficiência em nosso país.

2 CONCEITOS BÁSICOS EM MEDIDAS EDUCACIONAIS

Como comparar o desempenho de grupos diferentes em diferentes períodos de tempo? Este é o desafio da psicometria. Podemos ter duas soluções: aplicar o mesmo teste ou comparar formas diferentes de testes. No primeiro caso, não temos o erro de medida, mas em compensação os resultados poderão ser infla-

cionados pelo fato de os grupos passarem informações entre si, sobre o conteúdo dos testes. No segundo caso, eliminamos este efeito, mas, em contrapartida, estaremos sujeitos aos erros inerentes aos processos de comparação entre diferentes testes.

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

Com relação à linkagem e equalização, seguiremos a taxonomia de Mislevy e Linn, tal como apresentada por Kolen e Brennan (2004). Segundo esses autores, a linkagem subdivide-se em quatro tipos, em função da precisão da comparação que se deseja obter, e a equalização seria um desses tipos. Quanto maior for o rigor na estruturação dos testes cujos resultados desejam ser comparados, maior será essa precisão. Dessa forma, as melhores comparações são obtidas quando, nas diferentes avaliações: os testes medem o mesmo constructo, possuem a mesma estrutura, mesmos descritores distribuídos em testes paralelos, mesmo método estatístico para cálculo das proficiências e populações equivalentes. Variações nessas características influenciarão na qualidade da precisão dos resultados, ou seja, no nível de robustez (força) da medida obtida nesses processos. Assim, serão apresentados os quatro tipos de linkagem em ordem decrescente de robustez:

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

Os escores obtidos por esse processo são intercambiáveis, ou seja, se um teste X é equalizado a um teste Y, as interpretações obtidas por meio do teste X são equivalentes às obtidas no teste Y. No entanto, conforme observado por Lord (1980), na prática, essa intercambialidade não é perfeita, pois podem existir pequenas variações estruturais entre as diferentes formas do teste.

Para garantir a qualidade da equalização, devemos verificar: 1) se a correspondência entre escores equalizados é simétrica, ou seja, uma única tabela de correspondência deve ser usada para obter os escores da forma X em Y e vice-versa; 2) a invariância de grupo: a função de equalização deve ser a mesma para qualquer subgrupo da população, por exemplo, sexo, raça, região ou política educacional adotada; 3) a

invariância no tempo: não faz diferença se a equalização é baseada, por exemplo, em dados obtidos em 2000 ou 2005.

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

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

Podemos distinguir dois tipos de calibração. O primeiro, quando temos diferentes formas de testes não necessariamente com a mesma estrutura e com diferentes conteúdos, aplicados a populações equivalentes. Isto ocorre, por exemplo, quando, no 5º ano de Matemática, se utiliza uma estrutura de Blocos Incompletos Balanceados (BIB) de 13 blocos com 13 itens, 26 cadernos com 3 blocos, sendo que cada caderno possui blocos comuns entre si.

O segundo, quando se deseja medir a *performance* dos alunos em diferentes níveis de escolaridade. Neste caso, a calibração é comumente denominada de equalização vertical ou escalonamento vertical (*vertical equating* ou *vertical scaling*), que ocorre ao colocarmos em uma mesma escala alunos do 5º e do 9º anos do EF e alunos do 3º ano EM. Nesta situação, se os testes possuírem a mesma estrutura, os resultados serão melhores do que se as estruturas forem diferentes. Este é o tipo de linkagem mais aplicado nas avaliações educacionais em larga escala, por exemplo: avaliações estaduais com avaliações nacionais (Saeb).

Como esse tipo de linkagem é o normalmente utilizado nas avaliações estaduais com o Saeb, devemos ressaltar que, para se atingir tanto o objetivo de estimar as proficiências individuais dos alunos quanto os percentuais de alunos em determinados níveis de proficiência, é essencial que as duas avaliações estejam ajustadas com relação à mesma abrangência de conteúdo, às mesmas demandas cognitivas exigidas dos alunos e às mesmas condições em que os testes são administrados. Variações no grau de similaridades dessas condições influenciarão na confiabilidade das medidas obtidas.

O termo calibração é também utilizado por alguns autores para descrever o processo do cálculo dos parâmetros dos itens em uma mesma escala nos processos de linkagem (quase sempre em modelos da Teoria da Resposta ao Item).

- Projeção: forma unidirecional de linkagem aplicada a um mesmo grupo de respondentes em que os escores de um teste são projetados, por exemplo, por meio da regressão (linear ou não-linear) para se obter os escores de outro teste, sem a expectativa de que os mesmos estejam medindo exatamente a mesma coisa. É importante mencionar que a projeção de A em B não é necessariamente a mesma que B em A.

A precisão da projeção depende do quão forte é a relação entre os testes e necessita ser reavaliada frequentemente, pois a projeção é muito sensível e dependente do contexto, grupo utilizado para estabelecer a relação e tempo.

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

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

No Quadro 1 apresentamos os tipos de linkagem e suas principais características.

Quadro 1 – Características dos diferentes tipos de linkagem

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

Fonte: Kolen e Brennan (2004).

Nos processos de avaliação em larga escala, após a aplicação do método de linkagem, o passo seguinte é a construção de uma escala, ou seja, o escalonamento, que é o processo de transformação dos escores brutos em escores de escala, cujo objetivo é fornecer um significado para a medida pela incorporação de informações normativas ou de conteúdo, facilitando a sua interpretação.

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

3 FATORES QUE AFETAM A VALIDADE DA LINKAGEM

Podemos destacar quatro fatores que influenciam a confiabilidade dos resultados obtidos nos processos de linkagem.

3.1 Conteúdo do teste

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

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

3.2 Formato do teste

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

1. Testes de tamanhos variados. Por exemplo, na 4ª série do ensino fundamental da Prova Brasil, nos anos de 2005 e 2007, o número de itens por caderno aumentou de 40 para 44.
2. Testes que avaliam diferentes disciplinas. Por exemplo, a linkagem de Língua Portuguesa da avaliação do Estado do Rio de Janeiro de 2004 com o Saeb 2003. Nesse Estado, em um mesmo caderno de teste, foram avaliadas as disciplinas Língua Portuguesa, Matemática, Ciências Humanas e Ciências da Natureza, e, no Saeb 2003, apenas uma disciplina. Outro exemplo dessa situação seria a linkagem de Língua Portuguesa e Matemática do Saeb 2005 com o Saeb 2007, pois no de 2005 havia um teste para cada disciplina, e no de 2007 as duas disciplinas estavam no mesmo teste.
3. Testes com itens fechados e abertos. Linn, Sheoard e Hartka (1992) demonstraram diferenças percentuais de alunos nos níveis do NAEP, ao considerar seus desempenhos em itens abertos e fechados.
4. Testes com itens apenas lidos pelos alunos linkados com testes com itens lidos pelo aluno, lidos parcialmente e/ou totalmente lidos pelo aplicador. Por exemplo, a linkagem entre a 2ª e 3ª séries do ensino fundamental do Programa de Avaliação da Alfabetização de Minas Gerais em que na 2ª série os itens são lidos para os alunos e na 3ª série os alunos leem os itens.
5. Testes aplicados por um aplicador externo, não pertencente à escola, e testes aplicados por um professor da escola. Por exemplo, a linkagem das avaliações do Programa de Avaliação do Ensino Básico de Minas Gerais (Proeb) com o Saeb. No primeiro caso, o aplicador de testes é o professor da própria escola (de uma turma diferente da qual ele leciona), enquanto no Saeb há um aplicador externo.
6. Ordem das disciplinas nos testes: ao se avaliar duas disciplinas, que é a situação mais comum nas avaliações realizadas no Brasil, em que geralmente se avalia Língua Portuguesa e Matemática, podemos observar quatro diferentes tipos de montagem de cadernos: a) testes com apenas uma disciplina aplicados em dias diferentes; b) testes em que todos os cadernos têm a primeira metade dos itens de uma disciplina e a segunda metade de outra

disciplina, c) testes em que a primeira metade dos cadernos começa com uma disciplina e a segunda metade com outra disciplina; d) testes em que há uma mistura de blocos das duas disciplina de forma alternada, por exemplo: metade dos cadernos contendo o primeiro bloco com a disciplina A, o segundo com a disciplina B, o terceiro com a disciplina A e o último com a disciplina B; enquanto a outra metade dos cadernos começa com a disciplina B e termina com a disciplina A. Mostraremos, na seção 5.2, um exemplo de como a comparabilidade de resultados é afetada pela utilização de diferentes *designs* de testes.

3.3 Usos e consequências

A estabilidade do processo de linkagem é afetada quando sanções e premiações são adotadas em um grupo e não no outro. Isso ocorre ao linkarmos determinados Estados, que adotam políticas como premiações para escolas e alunos, bonificação de professores e ranqueamento de escolas, com o Saeb, que não possui esse tipo de política.

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

Conforme mencionado por Feuer et al. (1999, p. 89), quando testes relacionados com sanções e premiações são linkados com testes que não possuem essa característica, a dificuldade relativa entre eles é alterada, isto é, o teste parecerá mais fácil para avaliações inseridas no primeiro contexto, e isso pode afetar a estabilidade da linkagem ao longo do tempo.

3.4 Erro do método estatístico de equalização

Os diferentes procedimentos de equalização exigem uma série de pressupostos que podem não ser verificados na prática, por exemplo, a quantidade e qualidade de informação trazidas pelos itens comuns com o grupo ao qual se deseja realizar a linkagem (isso em *design* de itens comuns). A literatura tem estudos extensivos para avaliar tais erros, sejam analíticos, sejam obtidos por simulação.

Os diversos fatores apresentados, se considerados isoladamente, poderão não ter grandes efeitos na linkagem, entretanto o problema maior ocorre quando temos diferentes conjugações dos mesmos, o que acarretará grandes influências na comparação de resultados. Como exemplo, podemos citar a queda de proficiência em Leitura detectada pelo Naep, em 1984 e 1986, para os alunos de 17 e 9 anos. Constatou-se que essa queda foi em razão da mudança de *design* dos testes, o que afetou a estrutura dos itens comuns utilizados na linkagem, conforme relatado por Kolen e Brennan (2004, p. 22). Em 1984, os testes avaliaram apenas Leitura e Escrita e, em 1986, Leitura, Matemática e/ou Ciências para as idades de 9 e 13 anos e Leitura, Ciências da Computação, História e/ou Literatura para os alunos de 17 anos. Dessa forma, os itens de leitura utilizados na linkagem apareciam em diferentes ordens, e o tempo disponível para responder aos itens foi modificado nesses anos. Tal acontecimento, então denominado anomalia de Leitura Naep, ilustra a importância de administrar os itens comuns no mesmo contexto nas formas a serem linkadas. Caso isso não seja levado em consideração, os efeitos podem produzir resultados enganosos.

4 CARACTERÍSTICAS DAS AVALIAÇÕES EM LARGA ESCALA NO BRASIL

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

- Avaliações amostrais com representatividade de agregação de resultados para todos os Estados brasileiros.
- Participação das redes de ensino estaduais, municipais, federais e particulares
- Avaliações em Língua Portuguesa e em Matemática na 4ª e 8ª séries do ensino fundamental e no 3° ano do ensino médio.
- Criação de escalas de habilidades para Língua Portuguesa e para Matemática, por meio da técnica estatística da TRI, tendo a 8ª série do ensino fundamental de 1997 média de 250 pontos e desvio padrão de 50 pontos, garantindo, portanto, a comparabilidade de resultados entre os anos avaliados. Essa média e esse desvio padrão são a referência de escala de habilidades
- Avaliações realizadas a cada dois anos: 1997, 1999, 2001, 2003, 2005 e 2007.

Em 2005, procurando um mapeamento maior da educação básica, foi instituída a Prova Brasil, com característica censitária, avaliando todos os alunos, apenas

da rede pública, nas disciplinas de Matemática e de Língua Portuguesa, na 4ª e 8ª séries do ensino fundamental de oito anos. O Ministério da Educação (MEC), por intermédio do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), elabora, aplica e entrega os resultados da Prova Brasil, cabendo às escolas a participação na aplicação dos testes e o devido uso de seus resultados, tornando a avaliação um importante instrumento para gestão dentro de cada unidade escolar.

4.1 Diferenças entre o SAEB e a Prova Brasil

Verificamos, no Quadro 2, as principais características dessas duas avaliações que vêm ocorrendo de forma simultânea, no Brasil, a partir de 2005:

Quadro 2 – Características da Prova Brasil e do Saeb

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

Fonte: www.inep.gov.br

4.2 Avaliações em larga escala nos Estados brasileiros

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

Uma característica importante nessas avaliações estaduais, e não observada nos EUA, como mencionado anteriormente, é a preocupação de comparabilidade dos resultados dessas avaliações estaduais com os resultados do país. Para tanto, a parceria com o Inep, por meio de disponibilização de itens e bases de dados do Saeb foram imprescindíveis.

A partir de 2005, observa-se, com a criação da Prova Brasil, a tendência de o Inep avaliar, também de forma censitária, os alunos da rede pública de ensino, com o objetivo de fornecer maiores subsídios para os Estados, municípios e escolas. No entanto, o que se nota é que alguns Estados continuam com seus sistemas de avaliação, pois estes possuem algumas características diferentes da Prova Brasil, e, a princípio, não querem abandonar a metodologia empregada ao longo dos anos. Por exemplo:

- Minas Gerais, que possui uma série histórica de avaliação nos anos de 2000, 2002, 2003, 2006, 2007 e 2008, aplica testes, em dias distintos, de Língua Portuguesa e Matemática, em todas as escolas do Estado, independentemente do número de alunos. Diferencia-se, portanto, da Prova Brasil que aplica Língua Portuguesa e Matemática no mesmo caderno e não aplica testes em escolas com menos de 15 alunos.
- Rio Grande do Sul, que avalia a 5ª série/6º ano do ensino fundamental e 1º ano do ensino médio, o que difere das séries avaliadas pela Prova Brasil.
- Ceará, que, além do 5º e 9º anos do ensino fundamental, avalia todo o ensino médio, fornecendo os resultados por aluno avaliado, contrapondo-se à Prova Brasil cuja menor unidade de avaliação é a escola.

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

5 A INFLUÊNCIA DO DESIGN DOS TESTES NA COMPARABILIDADE DE RESULTADOS

Conforme relatado anteriormente, uma das principais diferenças entre o Saeb e a Prova Brasil é o fato de o primeiro ser amostral e, a segunda, censitária. Fa-

remos, entretanto, uma descrição mais detalhada das características particulares dos *designs* utilizados nesses dois sistemas de avaliação e suas consequências nas comparabilidades de resultados.

5.1 Histórico do Saeb e da Prova Brasil

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

Até 2005, o BIB utilizado pelo Saeb era composto por 26 modelos diferentes de cadernos por disciplina, e cada caderno era composto por 3 blocos de 8 itens. A posição dos diferentes blocos nos cadernos é apresentada no quadro abaixo.

Quadro 3 – BIB de 26 cadernos

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

Fonte: Arquivos Saeb.

As diversas versões da Prova Brasil e do Saeb, a partir de 2007, utilizaram um BIB diferente, composto por 21 modelos diferentes de cadernos, e cada caderno era composto por 4 blocos de itens e 2 disciplinas, conforme quadro abaixo.

Quadro 4 – BIB de 21 cadernos

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

Fonte: Arquivos Saeb.

Será feita a seguir uma descrição mais detalhada desses *designs* nas aplicações do Saeb e da Prova Brasil, em que poderemos observar as metodologias empregadas na construção dos testes, por esses dois sistemas de avaliação, em três momentos distintos.

1º momento: *Design* Saeb até 2005

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

Quadro 5 – *Design* de montagem dos blocos de itens nas versões do Saeb até 2005

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

Fonte: Arquivos Saeb.

2º momento: *Design* Prova Brasil 2005

Em um mesmo teste, o aluno foi avaliado em Língua Portuguesa e em Matemática, conforme *design* apresentado no quadro abaixo:

Quadro 6 – *Design* de montagem dos blocos de itens na Prova Brasil 2005

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

Fonte: Arquivos Saeb.

Neste *design*, os cadernos ímpares começaram com 2 blocos de Língua Portuguesa e terminaram com 2 blocos de Matemática, já nos cadernos pares a montagem das disciplinas foi invertida, ou seja, começaram com 2 blocos de Matemática e terminaram com 2 blocos de Língua Portuguesa.

3º momento: *Design* Prova Brasil 2007 e Saeb 2007

A partir de 2007, o Saeb mudou o *design* de seus testes, passando a utilizar, assim como na Prova Brasil, cadernos com as duas disciplinas juntas. O *design* do Saeb foi o mesmo da Prova Brasil, apenas com a diferença que no Saeb foi avaliado o 3º ano do ensino médio e na Prova Brasil essa série não foi avaliada. Também houve uma mudança nesse *design* com relação ao utilizado em 2005; conforme podemos

observar, o número de itens aumentou no 5º ano (4ª série) e a disposição dos blocos de Língua Portuguesa e de Matemática nos cadernos foi alterada com relação à versão anterior:

Quadro 7 – *Design* de montagem dos blocos de itens na Prova Brasil 2007

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

Fonte: Arquivos Saeb.

Quadro 8 – *Design* de montagem dos blocos de itens no Saeb 2007

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

Fonte: Arquivos Saeb.

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

Tais diferenças de *design* merecem estudos mais profundos, no que diz respeito à comparabilidade de resultados entre o Saeb, a Prova Brasil e os Estados, pois, em diferentes anos e projetos, os modelos adotados nessas instâncias não são os mesmos.

5.2 Evidência do efeito do *design* dos testes na proficiência

O fato de se ter duas disciplinas em um mesmo caderno de teste provoca cansaço e, consequentemente, a queda na proficiência na disciplina que está no final do caderno.

Faremos duas análises, por meio de diferentes projetos, sobre o efeito da ordem das disciplinas nos cadernos de testes.

5.2.1 Nova Escola em 2005 e 2006

No projeto Nova Escola, no Estado do Rio de Janeiro, em 2005 e 2006, os cadernos de testes eram compostos por Língua Portuguesa e Matemática (cadernos ímpares, iniciados com Língua Portuguesa, e cadernos pares, com Matemática). Segundo o *design* do Saeb 2005, observou-se uma diferença significativa nos cálculos de proficiência, ao se comparar os resultados dos grupos formados pela ordem das disciplinas nos cadernos. Alunos que fizeram os testes que começavam com Língua Portuguesa tiveram um valor de proficiência maior que aqueles que fizeram essa disciplina ao final do caderno. Os alunos que realizaram os testes que iniciavam com Matemática também obtiveram um resultado maior que os alunos que fizeram essa disciplina ao final do caderno. Os valores relativos a essas diferenças são apresentados na tabela 1.

Tabela 1 – Influência da ordem das disciplinas, Língua Portuguesa e Matemática, na proficiência dos alunos no Projeto Nova Escola

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

Fonte: Arquivos CAEd.

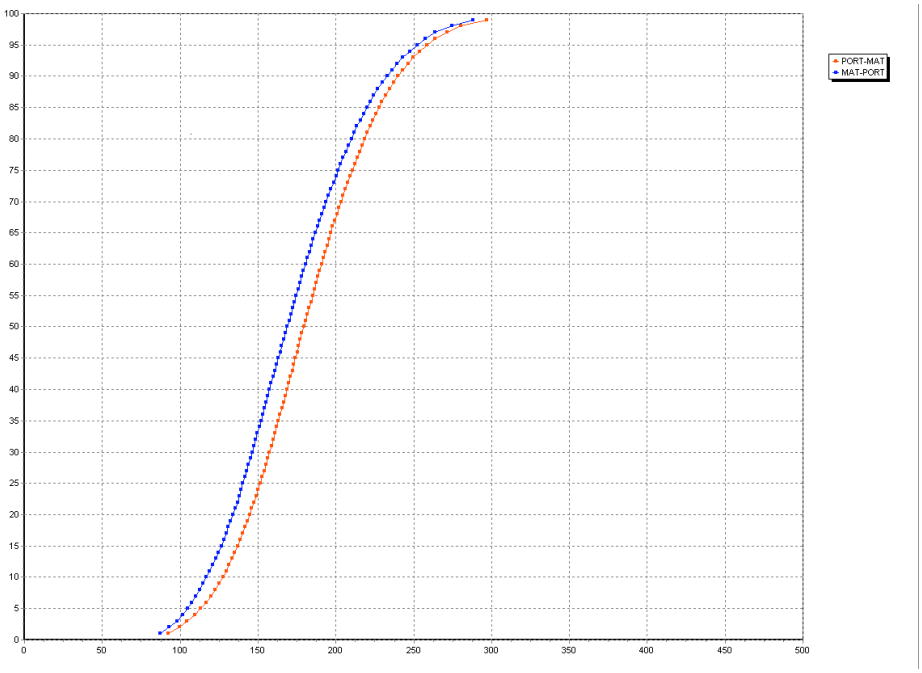
Observamos, nessa tabela, que as diferenças de proficiências em Língua Portuguesa foram mais significativas do que em Matemática, ao se considerar a ordem das disciplinas nos cadernos. Também podemos verificar que, em Língua Portu-

guesa, essa diferença, na 8ª série do EF e no 3º ano do EM, é mais significativa do que na 4ª série do EF e que, em Matemática, as diferenças entre esses três anos de escolaridade é praticamente a mesma, ou seja, o efeito na proficiência dos alunos, provocado por esse *design* é mais forte em Língua Portuguesa na 8ª série do EF e no 3º ano do EM.

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

Ao analisarmos a distribuição dos percentis de Língua Portuguesa, dos alunos da 4ª série, em 2006 (Gráfico 1), levando em consideração a posição dessa disciplina no teste, ou seja, no início ou no final, observamos que as maiores diferenças ocorrem no meio da curva e que, nas extremidades, essas diferenças são menores, evidenciando que para alunos com proficiência muito baixa ou muito alta o fato de mudar a posição da disciplina no caderno de teste produz pouco efeito na proficiência. Esse efeito é maior para os alunos medianos. Essa característica foi a mesma observada nas demais séries para as duas disciplinas, nos dois anos analisados.

Gráfico 1 – Percentis dos alunos na avaliação de Língua Portuguesa da 4ª série do EF, em 2006, no programa Nova Escola para cadernos na ordem Língua Portuguesa/Matemática e Matemática/Língua Portuguesa



5.2.2 Saeb nos anos 2005 e 2007

Observamos na tabela 2 a seguir que no Saeb de 2005 a diferença de proficiências entre alunos que responderam cadernos pares e alunos que responderam cadernos ímpares é insignificante. Já no Saeb de 2007 em que o teste era formado por disciplinas de Língua Portuguesa e Matemática, podemos verificar diferenças significativas em função das disciplinas nos cadernos.

Tabela 2 – Influência da ordem das disciplinas, Língua Portuguesa e Matemática, no caderno de teste e a proficiência dos alunos no Saeb

Série	Saeb 2005			Saeb 2006		
	Ordem das disciplinas	Língua Portuguesa	Matemática	Ordem das disciplinas	Língua Portuguesa	Matemática
4 EF	LP/LP/LP/LP	171.4	182.3	LP/MAT/LP/MAT	175.9	191.0
	MAT/MAT/MAT/MAT	172.6	182.7	MAT/LP/MAT/LP	173.2	194.4
	DIFERENÇA	-1.2	-0.4	DIFERENÇA	2.7	-3.4
8 EF	LP/LP/LP/LP	230.7	239.1	LP/MAT/LP/MAT	235.5	242.7
	MAT/MAT/MAT/MAT	232.4	240.1	MAT/LP/MAT/LP	229.3	248.6
	DIFERENÇA	-1.6	-1.0	DIFERENÇA	6.1	-5.9
3 EM	LP/LP/LP/LP	256.7	270.7	LP/MAT/LP/MAT	264.6	269.7
	MAT/MAT/MAT/MAT	257.7	271.8	MAT/LP/MAT/LP	257.8	276.3
	DIFERENÇA	-1.0	-1.2	DIFERENÇA	6.8	-6.6

Fonte: Arquivos Saeb.

5.2.3 Comparação entre os designs Nova Escola 2005 e Saeb 2007

Podemos verificar que ao alternar as disciplinas no caderno de teste, como foi o caso do Saeb 2007, e não concentrando as mesmas no início ou final do caderno, como no caso do Nova Escola 2005, a diferença de proficiência entre as disciplinas ficou praticamente a mesma nas três séries avaliadas, diferentemente do que ocorreu no Nova Escola 2005.

6 CONSIDERAÇÕES FINAIS

Nesse momento de transição do *design* do Saeb e tendo em vista o grande número de avaliações realizadas pelos Estados brasileiros, cada um com suas particularidades, as comparações entre resultados são muitas vezes conduzidas sem as devidas considerações.

Uma reflexão, no sentido de levantar sugestões para contornar a situação descrita acima, e até mesmo uma proposta para a reestruturação das práticas avaliativas no país, parte da constatação de que as avaliações conduzidas pelo Inep e pelos Estados são de características transversais, ou seja, caracterizam-se pela coleta periódica de dados em algumas séries da educação básica e que praticamente inexistem estudos longitudinais.

Conforme salientado por Franco e Alves (2008), os sistemas de avaliação deveriam incluir três componentes principais:

- estudos transversais;
- estudos longitudinais;
- indicadores escolares por meio de censos escolares anuais.

Tais componentes deveriam ser manipulados, utilizando-se métodos quantitativos e qualitativos, a fim de se obter uma análise ampla e profunda de seus dados.

O que se pretende, portanto, é uma análise conjunta dos componentes e técnicas, levando-se em consideração os recursos de cada um:

- Estudos transversais: indicadores do desempenho do sistema.
- Estudos longitudinais: possibilidade de realizações de estudos de valores agregados, ou seja, o que a escola está acrescentando de aprendizado ao aluno, levando-se também em consideração análises contextuais, mediante estudos de análises multiníveis.
- Censos escolares: informações sobre a infraestrutura das escolas, formação de professores e diretores.

A integração de métodos quantitativos e qualitativos é fundamental para a análise dos sistemas educacionais, pois, em linhas gerais, os métodos quantitativos localizam, por exemplo, o que acontece no interior das escolas, porém, para explicar o porquê, é necessária a utilização de métodos qualitativos.

Logo, o grande desafio para a avaliação passa pela utilização, harmonização e padronização de métodos e técnicas conduzidas de forma otimizada por órgãos

do governo, secretarias de educação, universidades e escolas, tendo como objetivo único a melhoria da qualidade do ensino.

Fica evidente, neste momento de transição de metodologias de modelos de testes, a necessidade de estudos com o objetivo de elucidar dúvidas e definir diretrizes na condução das avaliações, de forma a não perdermos a cultura de uma escala única em nosso país.

REFERÊNCIAS BIBLIOGRÁFICAS

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

FRANCO, C.; ALVES, M. T. G. A Pesquisa em eficácia escolar no Brasil: evidências sobre o efeito das escolas e fatores associados à eficácia escolar. In: BROOKE, N.; SOARES, J. F. (Org.). *Pesquisa em eficácia escolar:* origem e trajetória. Belo Horizonte: UFMG, 2008. p. 482-500.

KOLEN, M. J.; BRENNAN,= R. L. *Test equating, scaling, and linking:* methods and practices. 2. ed. New York: Springer, 2004.

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

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

Recebido em: outubro 2009
Aprovado para publicação em: dezembro 2009