AVALIAÇÃO EDUCACIONAL EM LARGA ESCALA E *ACCOUNTABILITY*: UMA BREVE ANÁLISE DA EXPERIÊNCIA BRASILEIRA

Reynaldo Fernandes[1]

Amaury Patrick Gremaud[2]

[1] Professor Titular do Departamento de Economia – FEA/RP Universidade de São Paulo, Ribeirão Preto, São Paulo, Brasil.Contato: refernan@usp.br.

[2] Professor Doutor, Departamento de economia – FEA /RP Universidade de São Paulo, Ribeirão Preto, São Paulo, Brasil.Contato: agremaud@usp.br.

Resumo

O presente texto procura avaliar o movimento de avaliação educacional em larga escala e *accountability* que ocorreu no Brasil nas duas últimas décadas. No artigo, é feita uma breve revisão da literatura internacional sobre *accountability* educacional, destacando os argumentos em sua defesa, a posição dos críticos e a avaliação de seus resultados. Em seguida, é apresentado um histórico do movimento de avaliação em larga escala e *accountability* no Brasil e, ao final, é realizada uma avaliação desse movimento, abordando três questões: a) até onde a melhoria dos dados educacionais brasileiros no ensino fundamental refletem uma melhoria no aprendizado ou reflete uma "inflação de notas?; b) até que ponto os programas de avalição e *accountability* contribuíram para esta melhoria? e c) Por que a melhoria no ensino fundamental não atingiu o ensino médio da mesma forma?

Palavras-chaves: Avaliação educacional. Accountability. Educação básica. SAEB. IDEB.

Abstract

This paper seeks to evaluate the large-scale educational assessment and accountability movement that has occurred in Brazil over the past two decades. In the article a brief review of the international literature on educational accountability is made, highlighting the arguments in its defense, the position of the critics and the evaluation of its results. This is followed by a review of the large-scale assessment and accountability movement in Brazil. At the end, an evaluation of this movement is conducted, addressing three issues: a) to what extent does the improvement of Brazilian educational data in fundamental education reflect an improvement in learning or reflect a "score inflation"? b) to what extent have evaluation and accountability programs contributed to this improvement? and c) why did the improvement in primary education not reach high school in the same way?

Keywords: Educational evaluation. Accountability. Basic education. SAEB. IDEB.

INTRODUÇÃO

Enquanto as primeiras iniciativas voltadas à implantação e desenvolvimento do Sistema de Avaliação da Educação Básica (SAEB) datam do final dos anos 80 e que ocorreram duas aplicações anteriores (1990 e 1993), podemos considerar o ano de 1995 como o ano de institucionalização definitiva do SAEB. A partir dessa data, o SAEB vem divulgando resultados de leitura e matemática para uma amostra de estudantes do final de cada uma das etapas do ensino básico. Em conjunto com as informações de movimentação e fluxo escolar, extraídas do censo da educação básica, o SAEB permitiu a realização de um detalhado diagnóstico da qualidade da educação ofertada no Brasil e em cada uma de suas unidades federativas.

Nos anos 2000, a avaliação educacional em larga escala no Brasil passou a incorporar a noção de *accountability*. A criação da Prova Brasil em 2005, a divulgação do Exame Nacional do Ensino Médio (ENEM) por escolas em 2006 e o lançamento do Índice de Desenvolvimento da Educação Básica (IDEB) em 2007, são marcos dessa nova orientação. Os resultados passaram a ser divulgados não apenas para o país e unidades da federação, mas também por redes de ensino e escolas individuais. Esse movimento de avaliação e *accountability* não ficou restrito ao governo federal. Hoje, diversos estados e alguns municípios possuem sistemas próprios de avaliação e programas de *accountability*, incluindo bônus para professores com base no desempenho dos alunos nos exames.

Os programas de *accountability* educacional costumam gerar polêmica, seja no Brasil ou no exterior. Seus defensores destacam as dificuldades existentes para monitorar o trabalho da escola (por pais, autoridades e a sociedade em geral). Esse quadro facilitaria o surgimento de um problema típico de agente-principal, onde os interesses dos agentes (professores, diretores e gestores educacionais) podem não estar totalmente alinhados com os interesses dos principais (governantes, eleitores, estudantes e seus familiares). Assim, os exames externos providenciariam informações independentes às autoridades e ao público sobre o desempenho das escolas em disciplinas chaves. Isso, associado a outros incentivos para que os alunos desempenhem bem nos exames, encorajaria os educadores a se concentrarem em tarefas que elevem o desempenho dos estudantes nos exames, potencialmente alterando o método de ensino. Por sua vez, seus opositores sustentam que tais programas, além de não terem impactos comprovados na melhora do aprendizado, distorcem os incentivos das escolas. Duas das principais preocupações dizem

respeito à exclusão dos estudantes com mais dificuldades de aprendizado e ao estreitamento do currículo (concentrar o ensino naquilo que é pedido nos exames externos).

Ao completarmos 25 anos da institucionalização do SAEB, parece ser uma boa oportunidade para avaliar esse movimento de avaliação educacional em larga escala e *accountability*. O presente texto pretende caminhar nessa direção. O restante do artigo está organizado em três seções, além da conclusão. Na próxima seção, realizamos uma breve revisão da literatura internacional sobre *accountability* educacional, destacando os argumentos em sua defesa, a posição dos críticos e, principalmente, a avaliação de seus resultados. A seção III traz um histórico do movimento de avaliação em larga escala e *accountability* no Brasil. Por fim, na seção IV, faz-se uma avaliação desse movimento no Brasil.

ACCOUNTABILITY EDUCACIONAL: SUA RACIONALIDADE, A POSIÇÃO DOS CRÍTICOS E AS EVIDÊNCIAS DISPONÍVEIS DE SEUS RESULTADOS.

O uso de avaliações educacionais em larga escala com objetivo de proporcionar estudos sobre aprendizagem e para monitorar sistemas educacionais não é algo novo. Um marco no uso de avaliações em larga escala no estudo da aprendizagem é dado pelo relatório Coleman (COLEMAN et al., 1966), que tinha como objetivo estudar a segregação racial no sistema educacional dos EUA. Esse trabalho é importante por uma série de razões, entre as quais a mudança de foco na forma de identificar a qualidade das escolas: de insumos e processos para resultados. Para efeitos de monitoração do sistema educacional, o National Assessment of Educational Progress (NAEP) dos Estados Unidos surge em 1969. Para essas duas finalidades, as avaliações tinham base amostral (cujos resultados não eram apresentados por escolas, professores ou estudantes) e não costumavam dar origem a grandes polêmicas, ainda que alguns questionamentos já estivessem presentes e incidiam, principalmente, sobre a limitação dos instrumentos (questionários, testes e matrizes) em captar o trabalho que se realizava nas escolas e sobre os processos infra escolares.

É só a partir do final dos anos 80 que as avaliações com fins de *accountability* ganham corpo, tendo como marco a reforma educacional inglesa de 1988. Ao incorporar a função de *accountability*, as avaliações em larga escala passam a ser universais e dão origem a uma grande polêmica sobre sua validade. Apesar da polêmica, as avaliações educacionais universais para fins de *accountability* se espalharam ao redor do mundo. Hoje, a maioria dos principais

países desenvolvidos e muitos países em desenvolvimento possuem sistemas universais de avaliação que impõem algum grau de *accountability* sobre seus sistemas educacionais, escolas e mesmo professores individuais.

Por *accountability* educacional, entende-se o processo de avaliar o desempenho de sistemas educacionais, escolas e professores individuais com base em medidas de desempenho dos alunos. Seguindo Hanushek e Raymond (2005), vamos dividir os programas de *accountability* em dois tipos: i) aqueles que se limitam a divulgar os resultados dos estudantes por escolas e sistema educacional, a "*accountability* fraca" e ii) aqueles que, além disso, atrelam prêmios, sanções e assistência a tais resultados, a "*accountability* forte". Enquanto nos programas do segundo tipo, as recompensas e sanções são explícitas – como bônus para professores em escolas consideradas excelentes ou ameaças de reestruturação ou fechamento de escolas de baixo desempenho – , nos programas do primeiro tipo, elas estão implícitas – operando menos por ação direta dos gestores do programa e mais por pressão da comunidade.

Os programas de *accountability* educacional têm por objetivo mudar a estrutura de incentivos para que professores, escolas ou sistemas educacionais proporcionem um melhor aprendizado aos seus estudantes. A racionalidade desses programas tem como base o problema do agente-principal, em que os interesses dos agentes (professores, diretores e gestores educacionais) não estariam totalmente alinhados com os interesses dos principais (pais, autoridades e a sociedade como um todo).[3] Nesse quadro e na presença de informação imperfeita (assimétrica e/ou incompleta), os educadores poderiam se comportar de uma maneira que não esteja totalmente de acordo com os interesses dos estudantes e/ou de seus responsáveis.[4] Admite-se que os responsáveis pela oferta de educação (professores, diretores e gestores) podem alterar suas condutas e, assim, proporcionar aos estudantes um melhor ensino. Entretanto, tais mudanças são consideradas custosas e, por esse motivo, os educadores não as efetivam. Então, ao providenciar informações independentes às autoridades e ao público em geral sobre o desempenho das escolas em disciplinas chaves, os programas de *accountability* atuariam no sentido de promover incentivos para que os educadores se concentrarem em tarefas

[3] Por exemplo, os educadores podem preferir gastar recursos em ações não diretamente associadas ao ensino (tornar o ambiente mais aprazível para eles ou usar da estrutura da escola para execução de serviços pessoais), podem conceder "benefícios" a eles mesmos, que acabam por comprometer a aprendizagem dos estudantes (faltar sem ter que repor as aulas ou focar o ensino na parte do currículo que mais gostam, ao invés daquelas mais importantes para os estudantes) etc.

[4] Para uma discussão mais detalhada sobre *accountability* e incentivos baseados em avaliações, ver, por exemplo, Fernandes e Gremaud (2009), Figlio e Loeb (2011) e Hout e Elliott (2011).

que elevam o desempenho dos estudantes nos exames. Se a simples divulgação dos resultados de desempenho não for considerada suficiente, pode-se atrelar recompensas e sanções às escolas com base no desempenho de seus estudantes (*accountability* forte).

De modo geral, os programas de *accountability* assumem que é do interesse dos estudantes e/ou de seus responsáveis que as escolas concentrem seus esforços no ensino de algumas disciplinas chaves, cujo aprendizado pode ser aferido por avaliações em larga escala. Evidentemente, quanto mais alinhado com esse objetivo estiver o interesse dos educadores, menos efetivo tenderá a ser o programa de *accountability*. Assim, alguém que considere que os professores já fazem o máximo para proporcionar aos estudantes o melhor aprendizado, tenderia a considerar que qualquer política de incentivos para eles seria, no mínimo, inócua. Por outro lado, aqueles que acreditam que a qualidade da educação pode melhorar implicitamente admitem que alguns dos responsáveis pela educação (professores, diretores, gestores de rede ou governantes) podem fazer algo diferente do que vêm fazendo. Nessa perspectiva, em algum nível (professores, escolas ou sistemas de ensino), a a*ccountability* pode ser necessária.

É importante destacar que diferentes desenhos de programas fornecem diferentes estruturas de incentivos. Por exemplo, se a medida *accountability* utilizada for a proporção de estudantes considerados proficientes (aqueles com pontuação acima de determinado nível), o incentivo fornecido para as escolas é que elas se concentrem naqueles estudantes logo abaixo do nível de proficiência, retirando atenção daqueles que já ultrapassam esse nível e daqueles com desempenho muito baixo (considerados com baixa probabilidade de de ultrapassar o nível fixado). Por sua vez, se a medida de *accountability* for a pontuação média da escola, o incentivo é para as escolas se preocuparem com todos os estudantes.

Os críticos dos programas de *accountability* levantam uma série de pontos, entre os quais destacam-se: 1) os programas são incompletos, pois não consideram todos os resultados importantes das escolas; 2) suas medidas de aprendizagem são imprecisas; 3) são injustos, ao responsabilizar os educadores por aspectos sobre os quais eles não possuem total controle; e 4) podem gerar distorções como o estreitamento curricular e a exclusão de alunos com maiores dificuldades de aprendizado.

Em relação ao primeiro ponto, é difícil discordar da alegação de que os objetivos dos programas existentes são limitados, frente aos múltiplos objetivos que podemos atribuir às escolas. A questão fundamental, no entanto, é sa-

ber se podemos considerar correto sinalizar para as escolas que priorizem suas ações no aprendizado de determinadas disciplinas consideradas chaves. Evidentemente, não há uma resposta óbvia para essa questão, a qual pode variar entre diferentes programas. Programas com objetivos mais estreitos podem incentivar as escolas a tirar o foco de aspectos importantes do ensino, enquanto programas com objetivos muito amplos podem ter dificuldade de obter medidas confiáveis de todos eles, além de proporcionar uma sinalização confusa sobre quais deveriam ser os principais objetivos das escolas. A escolha das medidas de desempenho é uma das decisões mais importantes do desenho do programa.

A segunda crítica é a de que os resultados das avaliações são uma medida imprecisa do aprendizado dos estudantes na disciplina considerada. De fato, o resultado dos estudantes nas provas não depende apenas da aprendizagem. Depende também da motivação e preparação específica para realizar o exame; das condições da aplicação; da sorte etc. Kane e Staiger (2002) mostram que os resultados de exames padronizados são medidas sujeitas a muito ruído, particularmente entre as pequenas escolas. A volatilidade das escolas no ranking de desempenho pode desacreditar o indicador utilizado. Mas, é possível adotar procedimentos para minorar esse problema como, por exemplo, usar a média das últimas edições do exame ao invés dos resultados de uma única edição, ou adotar um índice composto que agrega mais de uma medida de resultado. Uma preocupação na avaliação dos programas de *accountability* é que sua implantação pode levar as escolas a adotar medidas que elevam a pontuação nos exames sem que a aprendizagem seja afetada como, por exemplo, motivar e treinar os estudantes para o teste. Um fenômeno conhecido como inflação de notas (*score inflation*).

Quanto ao ponto três, é verdade que os resultados dos exames padronizados incorporam, além do esforço da escola e de seus professores, influências advindas da família, dos amigos e das habilidades inatas dos estudantes, bem como do erro aleatório de medida. Entretanto, isso não é necessariamente um problema em um programa de "*accountability* fraca", limitado à ampla divulgação dos resultados. O público interessado pode "extrair o sinal de qualidade" de uma escola, por exemplo, por comparar os resultados de mais de uma edição dos exames com o de escolas próximas e/ou que possuem público similar. Esse não é o caso, no entanto, para os programas de "*accountability* forte". Para esses, a questão da medida de desempenho é um elemento sensível, já que as premiações e/ou punições são automaticamente atreladas a ela. Nesse caso, seria necessário adotar alguma medida de valor adicionado.[5]

[5] Para uma discussão sobre Modelos de Valor Adicionado ver, entre outros, Reardon e Raudenbush (2009).

Por fim, os programas podem promover distorção de incentivos, mas existem formas de, se não as eliminar, reduzi-las. Se os objetivos dos programas são adequados (concentram-se no que é prioritário), o estreitamento do currículo não vem a ser um problema. Por outro lado, pode haver incentivo para a exclusão de alunos de baixa proficiência. Assim, os programas devem procurar incluir formas de penalizar a exclusão de alunos com baixa proficiência.

Existe hoje uma importante literatura avaliando o impacto de diversos programas de *accountability* implementados nos Estados Unidos e em outros países. Duas revisões dessa extensa literatura são Figlio e Loeb (2011) e Hout e Elliott (2011). As duas revisões apontam que a resposta das escolas aos programas de *accountability* podem ir tanto na direção intencionada pelo programa, quanto em adotar ações que visam a elevar o desempenho nos exames sem a correspondente melhora no aprendizado. Por exemplo, Hout e Elliott (2011, p. 62-63) apontam:

> Além das mudanças no ensino de determinado assunto, há evidências de tentativas de aumentar as pontuações de maneiras completamente não relacionadas à melhoria do aprendizado. As tentativas incluíram o ensino de habilidades para a realização dos testes, a exclusão de alunos de baixo desempenho dos testes, a alimentação dos alunos com refeições de alto teor calórico nos dias de teste, o fornecimento de ajuda aos alunos durante um teste e até a alteração das respostas dos alunos após a conclusão do teste.

Os dois estudos também encontram evidências de inflação de notas (*score inflation*), no sentido que o impacto positivo dos programas nos exames utilizados pelo programa (*high-stakes test*) tendem a ser maior do que nos exames não utilizados pelo programa (*low-stakes test*). Quanto a avaliação dos programas em elevar a aprendizagem nas disciplinas consideradas, as revisões chegam a conclusões diversas. Figlio e Loeb (2011) chegam a uma conclusão mais positiva dos programas, enquanto Hout e Elliott (2011) têm uma posição mais negativa. Isso é interessante, uma vez uma parcela importante dos artigos revisados é comum às duas revisões.[6] Figlio e Loeb (2011, p. 383) concluem que "A preponderância de evidências sugere efeitos positivos, do movimento de accountability nos Estados Unidos durante os anos 90 e início dos anos 2000, sobre o desempenho dos alunos, especialmente em matemática".

[6] Enquanto Hout e Elliott (2011) consideram apenas programas com consequências explicitas, Figlio e Loeb (2011) consideram todos os tipos de programas (*accountability* "fraca" e "forte"). Hout e Elliott (2011) excluem também programas que se utilizam de regressões descontínuas (por avaliar o impacto de apenas uma parcela dos alunos da escola) e programas para os quais não foi possível obter uma medida de um teste não considerado para efeitos de *accountability* (*low-stakes test*).

Por sua vez, Hout e Elliott (2011, p. 92) concluem que "Apesar de usá-los por várias décadas, os formuladores de políticas e os educadores ainda não sabem como usar os incentivos baseados em testes para gerar consistentemente efeitos positivos no desempenho e melhorar a educação".

Apesar da diferença nas conclusões, ao analisar objetivamente as duas revisões, vemos que elas apresentam um quadro muito parecido. Por exemplo, na tabela 4-1 da revisão de Hout e Elliott (2011), que resume os resultados dos artigos considerados, 17 impactos de programas com base em *low-stakes tests* são apresentados e os resultados são: 8 positivos, 7 não estatisticamente significativos e 2 negativos. Portanto, em sintonia com Figlio e Loeb (2011), os resultados são majoritariamente positivos. Entretanto, Hout e Elliott (2011) preferem destacar o pequeno valor das estimativas. A valor médio das estimativas por eles considerada é de 0,08 desvios padrão da distribuição de notas dos alunos do estado norte-americano (ou país) considerado. Eles também destacam que os impactos positivos se concentram em matemática e nas séries iniciais. Ainda que o uso do valor de 0,08 não esteja isento de críticas, ele dá uma dimensão dos impactos obtidos.[7]

Ao rever esses trabalhos, uma conclusão mais apropriada seria: as evidências disponíveis sugerem que, em média, os programas de *accoubtability* educacional apresentam efeitos positivos, mas modestos. Além disso, se concentram em matemática e nas séries iniciais. Para finalizar essa seção, seria importante destacar dois aspectos relacionados à conclusão acima. Primeiro, encontrar impactos modestos e concentrados nas séries iniciais em matemática não é uma exclusividade dos programas de *accountability*. Análises de impacto de reformas educacionais mais dispendiosas que *accountability* (aumento de salários dos professores, redução do tamanho da turma etc.) têm, na melhor das hipóteses, encontrado resultados modestos e, da mesma forma, concentrado em matemática e nas séries iniciais.[8] Por fim, a utilização de uma valor médio das estimativas esconde variações não apenas entre programas, mas também entre escolas de um mesmo programa. Parece que, quando o programa é posto em funcionamento, algumas escolas respondem de modo

[7] Hanushek (2012), em texto bastante crítico a Hout e Elliott (2011), contesta o uso do valor de 0,08. Ele incluiria estudos que não encontram os critérios definidos para inclusão na revisão, além de tirar o foco de programas que apresentam um excelente desempenho.

[8] Por exemplo, Aos e Pennucci (2003) revisam 53 estudos que avaliam o impacto da redução no tamanho da turma. Em todos os casos, a elasticidade do desempenho em relação ao tamanho da classe, em termos absolutos, é menor que 0,15. A maioria dos estudos na revisão mediu os resultados dos alunos (mudança de um ano) com pontuações padronizadas em testes; alguns examinaram as taxas de conclusão do ensino médio. Isso significa que reduzir o tamanho de sala de 30 para 27 alunos eleva o aprendizado anual médio dos estudantes em, no máximo, 1,5%. O efeito positivo de diminuir o tamanho das turmas é mais forte nas séries iniciais.

significativo, enquanto outras não dão qualquer resposta. Se for esse o caso, seria interessante investigar o que essas escolas têm de diferente. O que faz com que umas respondam aos incentivos e outras não.

UM BREVE HISTÓRICO DA AVALIAÇÃO EDUCACIONAL EM LARGA ESCALA E ACCOUNTABILITY NO BRASIL[9]

Como visto, é praxe considerar como marco inicial para os processos de avaliação educacional externa, o Relatório Coleman, com testes de desempenho aplicados a mais de 650 mil estudantes e de "*surveys*" aplicados aos próprios estudantes, pais, professores e diretores de escola, buscando levantar características do contexto e do processo educativo. Junto com este, o relatório Plowden na Inglaterra e a criação do NAEP fazem do final dos anos 60 um marco internacional das avaliações diagnósticas externas das redes educativas.

Nas décadas de 60 e 70, não se observam, no Brasil, levantamentos como estes. Boa parte das pesquisas e dos debates aqui travados diziam respeito a questões acerca dos fluxos educacionais: entrada, abandono e reprovações dos estudantes nas escolas e sistemas educativos. Nesse período, as questões de acesso à escola eram graves e preocupações sobre quantos e que tipo de alunos avançavam dentro do sistema estavam na base das pesquisas. Na década de 70, as pesquisas também foram muito influenciadas pela hipótese atrelada à teoria do capital humano de que a concentração de renda brasileira era consequência, pelo menos em parte, do baixo nível educacional da população brasileira. Essas pesquisas se baseavam nas informações acerca do (não) acesso ao sistema educacional e na (baixa) quantidade de séries que os diferentes grupos da população brasileira alcançavam.[10]

Nessa mesma década, os problemas da evasão e/ou abandono e da retenção e/ou reprovação também eram pontos importantes do debate dando sustentação a iniciativas, ainda na década de 70, de políticas de progressão continuada. Essas discussões avançaram pelos anos 80 junto com o próprio crescimento do número de jovens que adentravam no sistema educacional brasileiro. Se, por um lado, se observa, ao longo da década de 80, fortes

[9] Para uma discussão das avaliações externas e do processo de institucionalização do SAEB, ver Bonamino e Franco (1999), Castro (2016), Horta Neto (2007) e Pestana (2016).

[10] Segundo Horta Neto (2007), as primeiras medições da educação brasileira fizeram parte *do Anuário Estatístico Brasileiro*, produzido a partir de 1906, concentrando informações principalmente do Distrito Federal sobre número de escolas, de pessoal docente, matrículas e repetências. Estes dados foram interrompidos em 1918 e retomados, agora nacionalmente, em 1936.

melhorias nos dados de acessibilidade para as crianças nos anos iniciais do ensino fundamental, o abandono e a repetência continuavam a estar no centro das controvérsias e acabaram por trazer diversos questionamentos aos levantamentos empíricos realizados sobre estes fluxos educacionais e suas interpretações. Por exemplo, os alunos que deixavam de frequentar a escola e voltavam a se matricular na mesma série no ano subsequente eram considerados evadidos pelas estatísticas oficiais em vez de repetentes. Isso inflava as taxas de evasão e subestimava as taxas de repetência, o que ensejou a introdução de novos modelos para a interpretação dos dados como, por exemplo, o modelo Pro Fluxo.[11]

Nos anos 80 surgem, junto com algumas pesquisas de cunho etnográficos, as primeiras pesquisas de rendimento escolares (testes cognitivos) atrelados a fatores associados (questionários e observações contextuais). Tais avaliações foram realizadas dentro do Programa de Expansão e Melhoria do Ensino no Meio Rural do Nordeste Brasileiro (EDURURAL), que tinha por objetivos expandir o acesso à escola primária, diminuir as taxas de repetência e evasão e melhorar o rendimento escolar dos alunos. Fazia parte do programa – que contou com apoio do Banco Mundial, da Fundação Carlos Chagas, da Fundação Cearense de Pesquisa e da Universidade Federal do Ceará, além de pesquisadores estrangeiros[12] – avaliar os impactos dos investimentos realizados sobre os rendimentos dos alunos e sobre as taxas de reprovação. Esta investigação ocorreu em 1981, 1983 e 1985 e envolveu a aplicação de testes de Português e Matemática em aproximadamente 6.000 alunos da 2ª e 4ª séries do ensino fundamental de 600 escolas em 60 municípios dos estados do Ceará, Piauí e Pernambuco (HORTA NETO 2007; BONAMINO; FRANCO, 1999).

Interessante notar que, nessas pesquisas, juntam-se de modo explícito a questão dos fluxos, notadamente do abandono e da repetência, e dos problemas relativos à aprendizagem. As conclusões desses estudos, segundo Bonamino e Franco (1999, p. 105) foram:

> Com relação ao aspecto da melhoria da performance dos alunos a partir de investimentos realizados, a pesquisa detectou resultado positivo apenas no Estado do Piauí. Mas quanto à melhoria da taxa de promoção, os resultados foram negativos nos três estados, o que sugere que a melhoria nas taxas de promoção é um objetivo ainda mais difícil de ser alcançado do que a melhoria do desempenho dos alunos. O resultado sugere que, no contexto das novas condições viabilizadas pelos investimentos,

[11] Ver Fletcher e Ribeiro (1989) e Klein e Ribeiro (1991).

[12] Ver Harbison e Hanushek (1992) e Gatti (1994).

> os professores optam, ao menos em um primeiro momento, por aumentar seus padrões de exigência, o que redunda na tendência de manutenção, ou mesmo de elevação, das taxas de repetência. (...) o baixo rendimento curricular das crianças e o grande número de repetência decorrente de um conjunto de fatores, como baixos salários, influência política na designação de professores e infraestrutura curricular insuficiente, além de precariedades associadas às condições de vida dos alunos e suas famílias, em especial no que se refere às condições de saúde.

As bases dessas primeiras investigações com testes e *surveys* foram aproveitadas nos passos iniciais de institucionalização de um programa de avaliação educacional nacional. De modo a ampliar as informações disponíveis sobre a educação, incorporando, então, mais elementos em torno do processo de aprendizagem do alunado, a Fundação Carlos Chagas foi contratada, em 1987, pelo INEP para a realização do estudo "Avaliação do Rendimento dos Alunos de Escola do 1º Grau da Rede Pública: Um Estudo em 15 Capitais e 24 Cidades". A própria fundação estendeu este estudo no ano seguinte, com o Governo do Estado do Paraná, e acrescentou 29 cidades ao estudo.

Pestana (2016, p. 74) destaca que o próprio Ministério da Educação nesse período sente:

> necessidade de informações sobre as diversas realidades educacionais. Somente com base em um sistema de dados abrangente e robusto seria possível conhecer amplamente essas realidades e melhorar a capacidade de proposição e execução de políticas educacionais e de auditoria social. Os processos de avaliação educacional destacam-se então, como meio privilegiado de geração de informações do tipo requerido por essa forma de fazer política educacional.

Também em 1988, o MEC anunciou a criação do Sistema de Avaliação do Ensino Público de 1º Grau – SAEP. A pretensão era desenvolver mecanismos de avaliação de abrangência nacional, estendendo para todo o país os entendimentos que então se debatiam com o Banco Mundial, no interior do Projeto Nordeste, de promover avaliações semelhantes às realizadas no EDURURAL.[13] Uma questão importante, na implementação desses testes, dizia respeito à falta de um currículo nacional a ser utilizado como base para a construção das matrizes de referência. Essas foram construídas com base em consultas aos professores e na verificação sobre o que efetivamente era ministrado nas sa-

[13] A literatura, por exemplo Bonamino e Franco (1999), destaca nesse momento inicial o apoio de organismos internacionais como o Banco Mundial para a promoção destes mecanismos de avaliação. No caso do SAEP, Horta Neto (2007) aponta o financiamento do Instituto Interamericano de Cooperação para a Agricultura – IICA.

las de aula e contou com as duas experiencias anteriores da Fundação Carlos Chagas. Ainda em 1988, pré-testes foram realizados, mas a primeira aplicação só ocorreu em 1990 envolvendo alunos das então 1ª, 3ª, 5ª e 7ª séries. Os dados foram publicados em 1992 já com uma mudança no nome do estudo que passou a adotar o nome de SAEB, sendo esta aplicação de 1990 considerado 1º ciclo do SAEB. Pretendia-se institucionalizar ciclos a cada dois anos, o que foi efetivamente realizado a partir do segundo ciclo que, no entanto, ocorreu apenas em 1993. Atualmente completamos o 15º ciclo avaliativo.

Esses ciclos avaliativos do SAEB podem ser divididos em três fases, os dois primeiros (1990 e 1993) compondo a fase de implementação do SAEB Depois, a partir do terceiro ciclo de 1995, temos a segunda fase, de consolidação do SAEB como instrumento de diagnóstico e monitoramento da educação básica brasileira e, a partir de 2005, uma terceira fase marcada por acoplar ao SAEB elementos de *accountability* que, como visto, já tinham começado a se destacar nas avaliações internacionais no final dos anos 80.

Quanto aos dois ciclos iniciais destaca Pestana (2016, p. 75):

> O arranjo institucional inicial do sistema de avaliação, em período de grande luta por espaços de atuação em todos os níveis, contava com a participação e a contribuição efetiva das administrações estaduais de educação em termos técnicos, operacionais e financeiros. Em menor grau, algumas administrações municipais, especialmente capitais que possuíam grandes redes de ensino, também participaram nessa fase inicial. Nessa interação, que previa divisões de responsabilidades, foi possível aprofundar o debate sobre o significado, os meios utilizados e as consequências do uso da avaliação como uma política de melhoria da qualidade da educação, além de aspectos técnicos do processo de avaliação. Como resultado, essa dinâmica auxiliou na formação de uma primeira leva de técnicos e especialistas aptos a operar o sistema que se desenhava. Em relação aos custos, os estados assumiram diversas atribuições, entre elas o levantamento de dados, aspecto decisivo e o que mais contribuiu para a institucionalização do sistema e, principalmente, para introdução e disseminação de uma cultura de avaliação no setor educacional brasileiro.

Ainda segundo Pestana (2016, p. 78):

> Embora sejam inegáveis os ganhos técnicos obtidos com as modificações realizadas em 1995 (no método de análise dos testes e na metodologia de amostragem adotada), o mesmo não se pode afirmar em relação ao arranjo institucional do Saeb, que centralizou atribuições no MEC e diminuiu as atividades realizadas em parceria com estados e municípios.

Depois dos dois primeiros ciclos, em 1994, institucionalizou-se o SAEB como um sistema nacional de avaliação para diagnóstico e monitoramento. Algumas características foram introduzidas e se mantiveram pelos ciclos seguintes com poucas mudanças até 2005. Além da centralização, as principais modificações implementadas a partir do 3º ciclo de 1995 foram: a) a definição da aplicação dos testes nos alunos nas chamadas "series conclusivas", ou seja, na 4ª e 8ª série do ensino fundamental (atualmente 5º e 9º anos do ensino fundamental), além da inclusão da 3ª série do ensino médio; b) a inclusão de escolas particulares nas amostras; c) a manutenção dos testes versando sobre habilidades e competências nas áreas de Português (leitura) e Matemática (resolução de problemas) – ainda que, em alguns ciclos, terem sido aventados e, por vezes, acrescidos testes em outras disciplinas; d) as avaliações com base em amostras complexas, representativas em diferentes estratos, especialmente com a possibilidade de resultados em termos estaduais e por dependência administrativa; e) a consolidação de instrumentos contextuais, fornecendo informações das características socioeconômicas, culturais e dos hábitos e práticas dos alunos – além dos aspectos infra escolares e das práticas dos docentes e gestores educacionais, que já eram colhidos nos primeiros ciclos; f) a consolidação do uso da teoria da resposta ao item (TRI) como metodologia de apuração dos resultados e definição das escalas (uma única escala por disciplina, envolvendo todas as séries).[14]

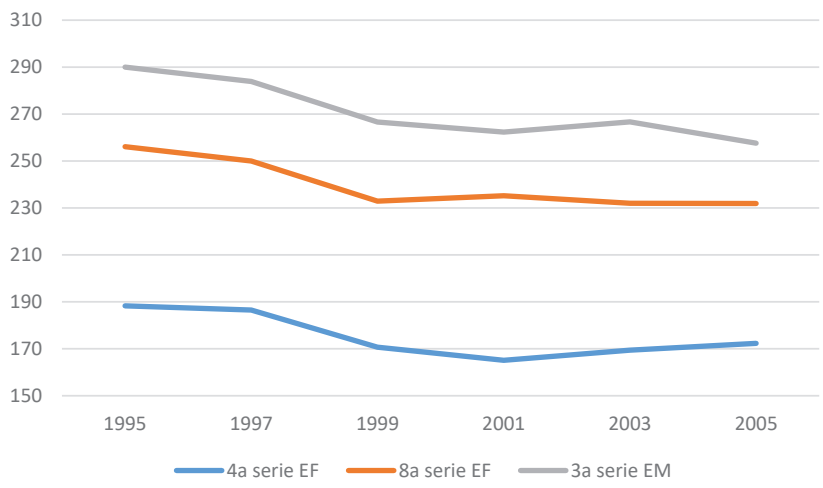
As duas últimas alterações acima permitiram importantes mudanças nas análises até então realizadas. Por um lado, a interpretação dos resultados da avaliação brasileira passou a dialogar com investigações internacionais, em que a preocupação sobre a influência das condições socioeconômicas e culturais sobre os resultados das proficiências dos alunos sempre foi objeto de atenção. No caso brasileiro, se confirma a importância dos aspectos socioeconômicos na explicação das diferenças de resultados entre os alunos, mesmo que o chamado "efeito escola" ainda tenha um peso bastante evidente e, nesse sentido, políticas educacionais tenham amplo espaço para melhorar a qualidade da educação nacional.

Por sua vez, o uso da TRI permite a comparação dos resultados ao longo do tempo, e a escala do SAEB passou a ser o termômetro da avaliação da qualidade brasileira. Ou seja, os ciclos avaliativos puderam ser comparados a partir de 1995 nos estratos amostrais semelhantes, mesmo que os testes aplicados ao longo dos anos sejam realizados com provas diferentes e sobre diferentes alunos. Os resultados dessas avaliações acabaram por colocar a questão da

[14] Ver Bonamino e Franco (1999).

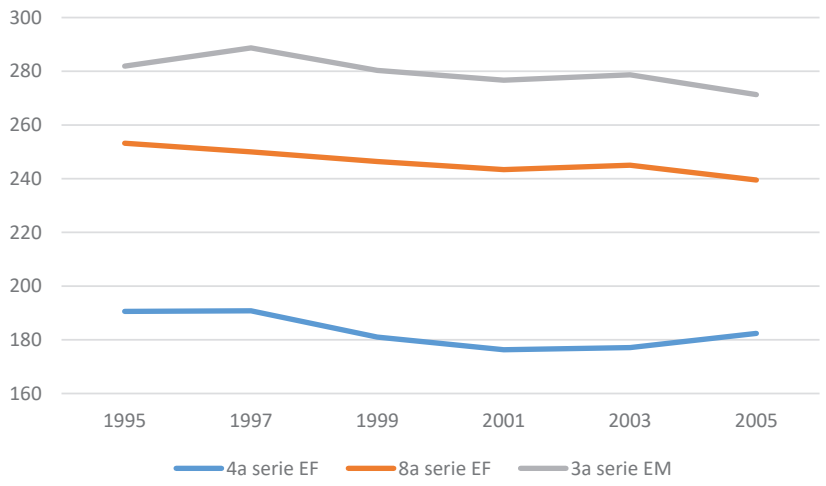
aprendizagem nos debates educacionais ao lado das históricas discussões sobre fluxos (repetência, evasão etc.). Mesmo que existam indefinições sobre quais os resultados que deveriam efetivamente ser esperados para considerar, na escala do SAEB, a educação brasileira proficiente, adequada ou em um nível razoável, era claro que os indicadores mostravam que estes resultados não estavam sendo plenamente alcançados, o processo de aprendizagem deixava a desejar e, no período, estavam em queda. Pelos Gráficos 1 e 2 abaixo, esta comparação das médias de proficiências dos alunos brasileiros nas respectivas séries e disciplinas ao longo desta fase do SAEB (1995 -2005) mostra, de uma forma geral uma queda no desempenho. A própria queda das proficiências é em parte atribuída à ampliação dos fluxos, ao crescimento do acesso ao sistema, e a diminuição da qualidade é atribuída, em boa medida, à expansão do sistema e à dificuldade de se manter um padrão estável de qualidade.

Gráfico 1 – Brasil: a evolução das proficiências médias do SAEB de Língua Portuguesa



Fonte: dados básicos INEP (2006).

Gráfico 2 – Brasil: a evolução das proficiências médias do SAEB de Matemática



Fonte: dados básicos INEP (2006).

Ainda em relação à avaliação nesse período, três outros elementos podem ser destacados: a) o desenvolvimento de sistemas avaliativos em outros entes da federação brasileira, especialmente alguns Estados; b) a criação de outras estratégias nacionais avaliativas e; c) a participação do país em avaliações internacionais.

Nos primórdios do SAEB, como já observado, houve participação ativa de diversos Estados e alguns municípios no processo. Parte dessas iniciativas deu origem a sistemas estaduais de avaliações, alguns dos quais se mantiveram até hoje, enquanto outros sofreram descontinuidades. Estados como São Paulo, com o SARESP, Minas Gerais, com o SIMAVE e o Ceará, com o SPAECE, são exemplos de sistemas estaduais de avaliação que se iniciaram no final do século passado e se mantiveram ativos até hoje. Um aspecto importante é que tais sistemas estaduais possuem a mesma escala de proficiência do SAEB.

Nacionalmente, dois outros sistemas de avaliação foram criados: o Exame Nacional para a Certificação de Competência de Jovens e Adultos (ENCCEJA) e o Exame Nacional do Ensino Médio (ENEM). O primeiro se tornou um exame que possibilitava a certificação para os níveis fundamental e médio de ensino para jovens e adultos que não tiveram a oportunidade de concluir seus estudos no período ideal.

Quando criado, em 1998, o ENEM tinha como objetivo principal fornecer um autodiagnóstico aos estudantes que terminavam o ensino médio. O exame, de caráter voluntário, permitia a comparação dos resultados obtidos por um

participante com a média dos demais participantes na edição daquele ano. Um dos problemas metodológicos do ENEM era que ele, diferentemente do SAEB, não permitia a comparabilidade ao longo do tempo dos seus resultados, impossibilitando qualquer afirmação sobre a evolução de desempenho dos estudantes ao final do ensino médio. Ao longo dos anos, várias instituições de ensino superior passaram a se valer dos resultados do ENEM como critério, parcial ou exclusivo, de seleção para ingresso nos seus cursos universitários. Em 2005, o governo federal também passou a utilizar o ENEM como critério para a concessão de bolsas no âmbito do Programa Universidade Para Todos (ProUni). Nesse sentido, o ENEM acabou também cumprindo as funções de seleção e credenciamento. Tais elementos, junto com a gratuidade do exame para aqueles oriundos de escolas públicas, fizeram com que o ENEM, apesar de continuar sendo voluntário, tenha atingido mais de 3 milhões de inscritos em 2005 com forte crescimento da participação dos concluintes do ensino médio no exame, que passou a ter uma boa cobertura nesse segmento.

Neste período, o Brasil passou a participar também de avaliações internacionais, o MEC decide pela participação do Brasil nos estudos internacionais do Laboratório Latino-americano de Avaliação da Qualidade da Educação (LLECE). O LLECE foi criado em 1994 com outros 14 países latino americanos, sob a coordenação do Escritório Regional sobre Educação da UNESCO para a América Latina e o Caribe (OREALC/UNESCO). Em 1997, o Laboratório aplicou o Primeiro Estudo Regional Comparativo e Explicativo (PERCE) sobre o desempenho da aprendizagem entre os alunos da 3ª e 4ª séries do ensino fundamental, em leitura e matemática. Anos depois, em 2006, foi aplicado o segundo estudo e, em 2013, o terceiro estudo.

Em 2000, o Brasil foi convidado a participar do Programa Internacional de Avaliação de Alunos (PISA). O PISA é uma avaliação internacional (como o TIMSS, PIRLS, CIVICS) que se desenvolve sob a coordenação da OCDE e que, ao longo do tempo, ganhou destaque internacional. A pretensão do PISA é medir a capacidade dos jovens de 15 anos de usar seus conhecimentos e habilidades de leitura, matemática e ciência para enfrentar os desafios da vida real. No primeiro ciclo avaliativo, em 2000, o Brasil foi o único país não membro da OCDE a participar daquele ciclo e também é o único país não OCDE a participar de todos os ciclos que passaram a ocorrer de três em três anos.

Com os dados do PISA, fica claro que a aquisição e o desenvolvimento de habilidades e competências por parte dos jovens brasileiros, independente das diferenças de interpretação em torno da escala do SAEB, estava muito abaixo daquelas desenvolvidas na média dos países da OCDE. Outro elemento que

ganha destaque com a participação no Brasil no PISA é a volta da questão dos fluxos. O Brasil chama a atenção por ter ainda vários de seus jovens de 15 anos cursando a 6ª série do ensino fundamental, ao invés da 1ª série do ensino médio (ou, ao menos, a 8ª série do ensino fundamental) como seria o esperado.

A terceira fase do SAEB se inicia no ano de 2005. A normatização é alterada, e o SAEB passa a englobar um teste com aplicação amostral – inicialmente chamada de Avaliação Nacional da Educação Básica (ANEB) – e um teste com aplicação censitária – a Prova Brasil (oficialmente, Avaliação Nacional do Rendimento Escolar – ANRESC). A Prova Brasil foi aplicada pela primeira vez em 2005 e sua principal diferença em relação ao SAEB é que, dentro de seu universo de referência, que são os alunos das escolas públicas do 5º ano e do 9º ano do ensino fundamental, ela é censitária.[15]

No ano de 2005, o SAEB e a Prova Brasil foram realizados separadamente, com instrumentos diferentes (testes e questionários). Já em 2007, estas duas avaliações se fundiram e voltamos a ter uma única avaliação externa federal, em que os segmentos que compõem o universo da Prova Brasil foram avaliados censitariamente e seus resultados foram divulgados por escola, municípios, estados e por redes de ensino; enquanto que os outros segmentos que compõem o SAEB, mas não a Prova Brasil, continuaram a ser avaliados amostralmente. Esses outros segmentos são as escolas privadas e o 3º ano do ensino médio, para os quais não existem resultados divulgados por município ou por escola. Os ciclos avaliativos se mantiveram a cada dois anos. Assim, desde 2005, o Brasil passou a contar com o tradicional sistema de avaliação para diagnóstico e, também, com um programa de "*accountability* fraca", por escolas e redes de ensino.

Ainda dentro desta perspectiva, uma mudança importante no ENEM foi introduzida em 2006. Como estudantes de praticamente todos as escolas do ensino médio participavam do exame que passava de 3 milhões de examinados, viu-se a possibilidade de divulgar os resultados do ENEM agregados por escola.[16] Assim, além de autodiagnóstico e de credenciamento, passou a ser possível a utilização do ENEM como um instrumento de diagnóstico e de *accountability* para o ensino médio. Em 2009, o ENEM foi novamente modificado e nele se introduziu a TRI, permitindo sua comparabilidade ao longo do tempo,

[15] A preocupação com a avaliação de outros níveis de ensino marcou também o período e, em termos federais, duas iniciativas foram introduzidas em relação ao processo de alfabetização: a Provinha Brasil e a Avaliação Nacional de Alfabetização (ANA).

[16] Das mais de 24.250 escolas de ensino médio que constam do Censo Escolar, cerca de 23.000 tinham alunos inscritos no Enem em 2007.

e expandido os currículos avaliados, buscando, assim, diminuir os problemas de estreitamento de currículos (que para o ensino médio eram evidentes). Um dos principais objetivos dessa mudança foi o de atrair as Universidades Federais para utilizar o ENEM em seus processos seletivos. Em 2019, o ENEM teve mais de 5 milhões de inscritos, sendo que 77% desses realizam, de fato, o exame.

O ENEM possui algumas vantagens em relação ao SAEB para efeitos de *accountability*: vai além de matemática e leitura (avalia escrita, ciências naturais e ciências humanas), inclui escolas particulares e, por ser instrumento de seleção das universidades, os alunos tendem a realizá-lo com mais comprometimento. O fato de o ENEM ser de adesão voluntária não deveria ser impedimento para sua ampla divulgação, pois: (a) isso traz pouco impacto na ordenação das escolas e (b) a literatura especializada dispõe de diversos corretores de participação que poderiam ser usados.[17]

Apesar disso, o INEP anunciou que não mais divulgaria os resultados do ENEM por escolas e, a partir de 2017, mudou a Prova Brasil, que atualmente voltou a ser chamada de SAEB, a qual passou a ser universal para as escolas públicas também no 3º ano do ensino médio. Assim, a base do sistema de *accountability* no Brasil partir de 2017 é apenas o SAEB, que por enquanto, para o ensino médio, voltou a enfrentar um problema de estreitamento de currículo.

Uma outra questão dentro do processo de *accountability* é a ligação entre as questões de aprendizagem dos alunos e as tradicionais e sempre presentes questões de fluxos na educação brasileira. Como a proficiência em exames padronizados e o fluxo escolar não são independentes, restringir a cobrança aos resultados da Prova Brasil poderia incentivar os professores, diretores e gestores a adotarem medidas dentro das escolas que aumentassem tanto o desempenho médio dos estudantes nos testes padronizados quanto às reprovações, por exemplo, endurecendo dos critérios para aprovação. Assim, foi introduzida uma nova estatística educacional, o IDEB. Este surge com o objetivo de ancorar a *accountability* em um sistema de metas educacionais, sem que este estivesse baseado apenas nos resultados da Prova Brasil e, portanto, com o risco o de contribuir para agravar o já dramático quadro de repetência e evasão escolar. Acerca do IDEB, podemos acompanhar a descrição feita por Fernandes (2016, p. 103) [18]:

[17] Para as escolas da amostra do SAEB 2011, a correlação entre as notas do SAEB e do ENEM foi de 0,87 e 0,91 para leitura e matemática, respectivamente. Para aplicação de corretores de participação no SAT dos Estados Unidos ver Dynarski (1987); Dynarski e Gleason (1993); Behrendt, Eisenach e Johnson (1986); e Clark, Rothstein e Schanzenbach (2009).

[18] Para uma discussão sobre as propriedades do IDEB, ver Fernandes (2007).

> O Ideb é obtido pela multiplicação da proficiência média dos alunos da escola (N) pela taxa média de aprovação da escola (P): Ideb = NP. Sob certas hipóteses ele pode ser interpretado como a razão entre a proficiência média dos alunos da escola (N) e o tempo médio que os alunos levam para concluir uma série (T): Ideb= N/T. Por exemplo, se, em média, os alunos precisam de dois anos para concluir uma série, o Ideb será igual à metade da proficiência média dos alunos da escola. Ele foi construído como forma de eliminar as reprovações improdutivas: reprovações que não contribuem para elevar o desempenho dos estudantes. Se as reprovações contribuírem para melhorar o desempenho dos estudantes da escola – seja porque incentivam os alunos a estudarem mais, seja porque tornam as turmas mais homogêneas –, a taxa ótima de reprovação seria diferente de zero, mas, provavelmente, muito baixa. Isso se confirmada a crença, de grande parte dos pesquisadores em educação, que reprovações são pouco produtivas.

Em 2007, com o objetivo de obter um maior comprometimento das redes e escolas com a melhoria da educação brasileira, foi pactuado, entre o Ministério da Educação e secretarias de educação de estados e municípios, um sistema de metas, e foi estabelecido o Plano de Metas Compromisso Todos pela Educação com base no IDEB. As metas do IDEB foram estipuladas para 2021, com metas intermediárias estabelecidas de dois em dois anos, a partir de 2007.

Para a meta de 2021, adotou-se um padrão externo: o desempenho educacional que, em média, era observado nos países da OCDE. Para cálculo do IDEB dos países da OCDE, supôs-se uma taxa de aprovação de 96%. A principal dificuldade foi fixar as notas, uma vez que os países da OCDE não fazem a Prova Brasil. Para tanto, admitiu-se que o PISA ordena os alunos da mesma forma que a Prova Brasil e, então, verificou-se que o percentil da distribuição de notas do Brasil no PISA era correspondente à média de desempenho dos países da OCDE. Encontrado esse percentil, obteve-se a nota correspondente a ele na Prova Brasil de 2005, a qual passou a ser referência para a meta do IDEB. Para definição de metas para redes de ensino e escolas individuais, considerou-se que todas as redes e escolas deveriam contribuir para que o país atingisse a meta estipulada, mas quem partisse de uma situação melhor no início também teria que obter melhores resultado ao final. Ainda que as diferenças de desempenho não fossem eliminadas, as metas consideram uma redução da desigualdade entre redes de ensino e escolas quando comparado com o ano base de 2005. A metodologia adotada considerou que a trajetória do IDEB ao longo do tempo, tanto para o Brasil como para os demais níveis de abrangência, segue o comportamento de uma função logística. Des-

sa forma, foi possível calcular o "esforço" que o Brasil e cada uma das redes e escolas teria que fazer para atingir a meta em 2021, partindo do desempenho observado em 2005.

O plano contemplava diferentes incentivos para que as diferentes escolas e redes de ensino acolhessem tais metas e se comprometessem com elas. As escolas que atingissem as metas eram beneficiadas com o aumento de seus recursos no Programa Dinheiro Direto na Escola (PDDE), mas a principal atenção do governo foi às redes que tinham piores índices. O MEC estabeleceu convênios com estados e municípios, por meio da elaboração local de um Plano de Ações Articuladas (PAR). Pelo PAR, os gestores municipais e estaduais se comprometiam a promover um conjunto de ações, responsabilizando-se pelo alcance das metas estabelecidas no âmbito federal. Em contrapartida, passavam a contar com transferências voluntárias e assessoria técnica da União.

Quando se observa a ação dos Estados e Municípios brasileiros, deve-se destacar que apesar da adesão das redes estaduais e municipais à Prova Brasil ser voluntaria, apenas alguns pouco municípios não participaram de algumas das avaliações.[19] Esta elevada participação é confirmada por pesquisa recentemente realizada (BAUER et al, 2017)[20]. Elemento importante revelado pela pesquisa que pelo menos 20 Estados desenvolveram indicadores e sistemas próprios de avaliação e aproximadamente 30% dos municípios brasileiros também desenvolveram indicadores e processos avaliativos próprios.

Quanto ao movimento de *accountability,* este também não ficou restrito ao governo federal. Diversos estados e alguns municípios, além de manterem ou desenvolverem sistemas próprios de avaliação, introduziram diferentes programas de *accountability*. Conforme destaca Brooke (2006), mesmo antes do governo federal, estados como o Ceara, já em 2001, estabeleceram uma conexão entre o seu sistema de avaliação – o SPAECE – e o Projeto Melhoria da Escola, com prêmios e recompensas em dinheiro para escolas e suas equipes que obtivessem os melhores resultados. Esta experiencia foi, ao longo dos anos, modificada, mas a ideia de *accountability* segue até os dias de hoje no Ceará. Indicadores de qualidade educacional estão atualmente presentes, inclusive na própria transferência de parte do ICMS do Estado para os municípios cearenses.[21]

[19] Na primeira aplicação da Prova Brasil, o Estado de São Paulo produziu apenas resultados por regional administrativa, mesmo que todas as escolas da rede estadual tenham participado.

[20] Segundo Bauer et ali (2017) os municípios não apenas participam da Prova Brasil (96% dos respondentes da pesquisa), mas 97% destes municípios também utilizaram a Provinha Brasil e 90% participaram da ANA.

[21] Um sistema semelhante está sendo implementado em Pernambuco atualmente.

Além do Ceará, Rio de Janeiro e Paraná, segundo Brooke (2006), também desenvolviam sistemas de responsabilização no mesmo momento em que o governo federal desenvolvia o seu. Depois dos passos dados pelo MEC, outros indicadores semelhantes ao IDEB foram desenvolvidos, por exemplo, em São Paulo, e avaliações em outras séries e disciplinas foram produzidas em diferentes municípios e/ou Estados. Políticas de difusão e análise de resultados foram implementadas em várias localidades, assim como políticas de bônus para professores com base no desempenho dos alunos nos exames e/ou em indicadores assemelhados também foram implementados. Segundo Bauer et al. (2017), existe uma parte substancial dos municípios (85%) que utilizam os resultados das avaliações (próprios ou não) em processos de difusão de resultados para diferentes públicos[22] - *accountability* fraca. Quanto à *accountability* forte, com a concessão de prêmio, bônus ou outras consequências, pouco menos de 30% dos municípios respondentes disseram fazer uso deste tipo de iniciativa.

AVALIAÇÃO EDUCACIONAL EM LARGA ESCALA E *ACCOUNTABILITY*: UMA ANÁLISE DA EXPERIÊNCIA BRASILEIRA.

Ainda que não unânime, a implantação e a ampla divulgação das avaliações censitárias, a exemplo da Prova Brasil e do IDEB, tiveram uma boa aceitação por parte da opinião pública. As divulgações têm tido grande repercussão na mídia e têm despertado o interesse de professores e gestores públicos. Botelho et al. (2014) mostram que 80,4% dos professores da rede pública de ensino conheciam o IDEB de sua escola. Há também evidência de que os resultados das avaliações são levados em consideração pela população e impactam a eleição de prefeitos (FIRPO; PIERI; SOUZA, 2017).

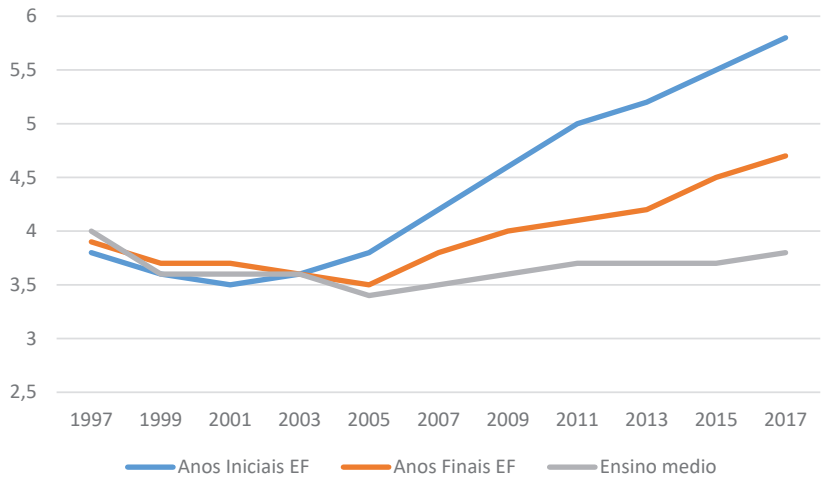
Entretanto, saber qual o impacto de todo esse movimento de avaliação e *accountability* educacional sobre a qualidade da educação do Brasil não é algo que possua uma resposta simples. O gráfico 3 sugere uma inflexão nos IDEBs, justamente em 2005, ano de implantação da Prova Brasil e a partir do qual o IDEB passou a ser divulgado.[23] Evidentemente, não podemos atribuir, automaticamente, a melhora observada no IDEB à política de *accountability*

[22] Umas destas formas de divulgação de resultados que foi objeto de polêmicas é a colocação de placas nas escolas com os seus resultados.

[23] Na realidade, a inflexão observada para o ensino fundamental 1 ocorre um pouco antes, em 2001. O IDEB oficialmente é divulgado a partir de 2005, mas é possível calculá-lo desde 1995, com base nos dados do SAEB e do Censo da Educação Básica.

em questão. A melhora poderia ser consequência de outros fatores que se deram independente e simultaneamente à Prova Brasil e ao IDEB. Por outro lado, o crescimento do IDEB é bastante distinto entre as diferentes etapas de ensino. Ele é bastante expressivo na primeira etapa do ensino fundamental, menos pronunciado na segunda etapa do fundamental e pequeno para o ensino médio.

Gráfico 3 – Brasil: evolução do IDEB 1997-2017



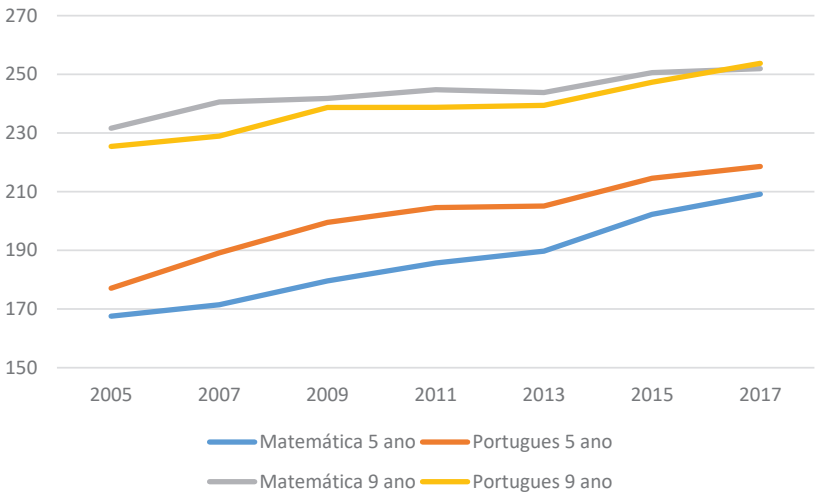
Fonte: elaboração dos autores dados originais do INEP

O gráfico 3 levanta três questões importantes a respeito da evolução do IDEB. Primeira, o aumento significativo do IDEB do ensino fundamental a partir de 2005 é genuíno ou está contaminado por inflação de notas? Segunda, se genuíno, pode ser explicado, ao menos em parte, pela política de *accountability*? Por fim, por que a elevação do IDEB no ensino fundamental não se verifica para o ensino médio? Sem ter a pretensão de dar uma resposta definitiva a essas questões, esta seção tem por objetivo tecer algumas reflexões sobre elas.

Crescimento do IDEB do Ensino Fundamental a partir de 2005: Melhoria no Aprendizado ou Inflação de Notas?

Pelo gráfico 3, percebemos a elevação do IDEB no ensino fundamental, especialmente a partir do segundo lustro da primeira década do século XXI. Se decompormos os dados do IDEB, podemos perceber que a melhoria se deve muito aos avanços nas proficiências obtidas na Prova Brasil, ou seja, o avanço do IDEB se deve mais às melhoras na proficiência da Prova Brasil/SAEB do que aos avanços nos dados de aprovação. Os avanços nas proficiências podem ser observados no gráfico 4.

Gráfico 4 – Brasil: proficiências Prova Brasil ensino fundamental 2005-2017



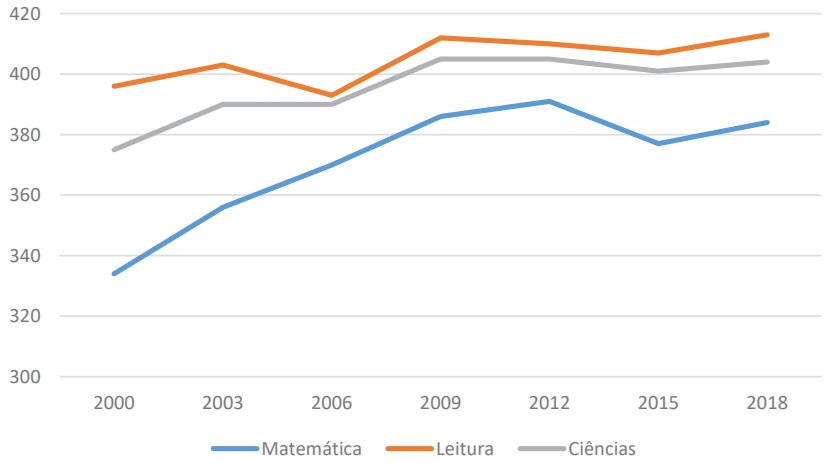
Fonte: elaborado pelos autores com dados do INEP.

Como visto anteriormente, uma possibilidade é que esta melhoria nas proficiências se deva a fenômenos como os de inflação de notas (*score inflation*), quando diretores e professores das escolas adotam medidas que elevam a pontuação nos exames sem que a aprendizagem seja afetada como, por exemplo, motivando e treinando os estudantes para o teste. Esta é efetivamente uma possibilidade, mas não temos conhecimento de análises que tenham conseguido estabelecer quanto da ampliação das notas demostradas no gráfico 4 se devem a este fenômeno ou quanto é um crescimento genuíno da aprendizagem.

Por outro lado, é importante destacar duas coisas. A primeira é que a elevação das notas, especialmente as do 5º ano, mesmo que pequenas se olhadas de um ciclo avaliativo para outro; quando se olha todo o período, é de mais de 40 pontos (ou por volta de 25 pontos no 9º ano), isto se aproxima a um desvio padrão na escala do SAEB e pode significar algo como o equivalente a quase dois anos adicionais de aprendizagem (um ano no caso do 9º ano), o que não é algo pequeno e dificilmente se obteria apenas com sucessivos processos de treinamento para a Prova Brasil. O segundo elemento a configurar a hipótese que este crescimento não se deve apenas ao fenômeno da "inflação de notas" é o desempenho do Brasil em um outro teste, diferente e aplicado amostralmente no Brasil, no caso os dados do PISA. Estes dados poderiam ser comparados especialmente com os resultados do 9º ano (apesar do resultado do PISA ser negativamente impactados pelos alunos de 15 anos que frequentam o 7º e 8º anos). No gráfico 5, o resultado dessa avaliação (PISA) é

mostrado e possui um comportamento relativamente parecido com os dados da Prova Brasil, reforçando, assim, a hipótese de que, se não podemos afastar totalmente fenômeno do "*score inflation*" no caso brasileiro, ele claramente também não pode explicar todo o ganho de proficiência e efetivamente existe aprendizagem genuína nos dados sobre a evolução do SAEB/Prova Brasil.

Gráfico 5 – Brasil: resultados do PISA 2000-2018



Fonte: elaborado pelos autores com base em dados da OCDE.

Os Programas de Avalição e Accountability Contribuíram para O Aumento do IDEB a partir de 2005?

Como vimos na subseção anterior, a elevação das pontuações do SAEB e do IDEB do ensino fundamental (especialmente na 1ª fase) foi expressiva e não pode ser justificada apenas por inflação de notas. Assim, somos levados a concluir que houve melhora das condições de ensino-aprendizagem após 2005. Evidentemente, não temos como afirmar que tal fato não teria ocorrido se a política de *accountability* relacionada à Prova Brasil e ao IDEB não tivesse sido implantada. Uma melhor interpretação seria, talvez, que tal crescimento decorre de uma maior mobilização de diversas esferas de governo (federal, estadual e municipal) e da sociedade em geral para a melhoria da educação, onde a política relacionada à Prova Brasil e o IDEB seja apenas um elemento desse movimento mais geral.

Vários fatores poderiam ser relacionados para ajudar a explicar o crescimento das pontuações obtidas nas avaliações. Em primeiro lugar, o grande aumento na taxa de frequência à pré-escola ocorrido entre 1985 e 2005 pode ter contribuído para que as crianças que ingressaram no ensino fundamental no pe-

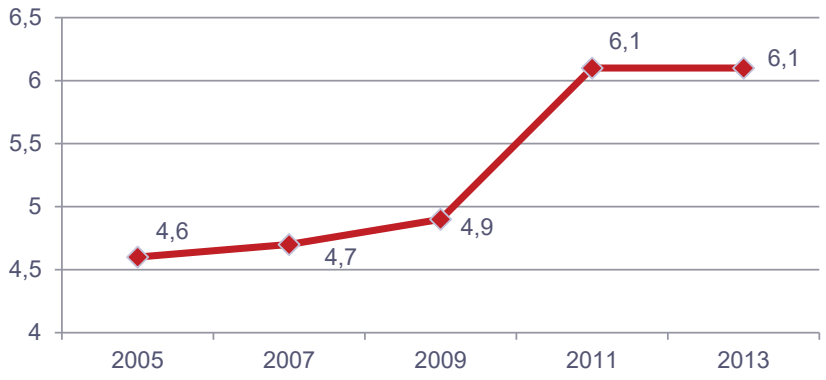
ríodo subsequente tivessem maior capacidade de aprendizado. Além disso, o aumento educacional das mães, que ocorreu entre 1995 e 2005, também pode ter contribuído para aumentar os investimentos familiares nas crianças.

Uma medida que merece destaque diz respeito à ampliação do ensino fundamental de oito para nove anos.[24] Os gráficos 6 e 7 mostram a evolução do IDEB da primeira fase do ensino fundamental para duas redes de ensino: rede estadual de Minas Gerais e rede municipal de Ribeirão Preto (SP).

Gráfico 6 – Minas Gerais IDEB – anos iniciais

Fonte: elaborado pelos autores com base em dados do IBGE.

Gráfico 7 – Ribeirão Preto IDEB – anos finais



Fonte: elaborado pelos autores com base em dados do IBGE.

[24] Em 06/02/2006, foi sancionada lei Federal regulamentando o ensino fundamental com 9 séries, dando prazo até 2010 para que todas as redes de ensino do país se enquadrassem na nova regra. Entretanto, algumas redes já haviam adotado o ensino fundamental de nove anos antes de 2006.

Em Minas Gerais, o ano 2009 é o primeiro ano em que os alunos sem repetência chegam ao final da primeira fase do ensino fundamental com cinco anos de escolaridade, ao invés de quatro. Em Ribeirão Preto, isso ocorre no ano de 2011. Assim, o impacto da implantação do ensino fundamental de nove anos parece notório. Eleva o IDEB em cerca de um ponto. Como diferentes redes implantaram o ensino fundamental de nove anos em anos distintos, o impacto da medida na média do IDEB tende a ser mais suave no tempo. Peña (2014) avalia o impacto do ensino fundamental de nove anos na primeira fase do fundamental entre 2007 e 2011, concluindo que a medida elevou o SAEB de matemática em 4,43 pontos (0,09 DP) e de leitura em 5,11 pontos (0,1 DP). Isso explicaria apenas 11% e 14% da variação observada no período, respectivamente, na pontuação de matemática e leitura.

Por fim, iniciativas para melhorar a gestão em alguns municípios, tais como Sobral no Ceará, também trouxeram bons resultados em termos de aprendizado.[25]

Note que as políticas de *accountability* não concorrem com várias medidas adotadas pelas escolas e sistemas educacionais na explicação da melhora do desempenho observado. A *ccountability* não afeta diretamente as práticas de ensino. Ela altera a estrutura de incentivos e, se funcionar, leva os responsáveis a ofertar educação a adotarem medidas que afetam diretamente o processo de ensino-aprendizagem: controle de faltas docentes, redução do tamanho das turmas, alteração do currículo, tutoria etc.

De todo modo, dado a magnitude da evolução dos indicadores e o período da reversão de tendência, é difícil argumentar que as políticas de avaliação e *accountability* não contribuíram em nada com a melhoria da aprendizagem no ensino fundamental no Brasil.

Por Que o Crescimento do IDEB no Ensino Fundamental não Atingiu o Ensino Médio?

O crescimento do IDEB observado para o ensino fundamental não se repete para o ensino médio. O crescimento do IDEB entre 2005 e 2017 foi de apenas 0,4 pontos nesse segmento. E mais, esse crescimento foi comandado pela redução da repetência. No período, a proficiência cresceu 10,9 pontos em leitura e diminuiu 0,7 em matemática. Em matemática, a proficiência atinge em 2015 (267,5 pontos) o menor valor observado para o SAEB/Prova Brasil, desde sua implantação em 1995. Esse desempenho pouco favorável do ensino médio tem gerado preocupação e controvérsia.

[25] Ver Rocha, Komatsu; Menezes-Filho (2018).

Uma parcela da diferença de crescimento entre os IDEBs do ensino fundamental e do ensino médio poderia ser explicado pela inflação de notas. Como o ensino médio não tem prova Brasil e IDEB por escolas e rede de ensino, os resultados SAEB e IDEB do ensino médio são *low-stakes* e, em princípio, não contaminados pela inflação de notas. Como vimos, entretanto, a inflação de notas não explica toda a elevação do IDEB e Prova Brasil do ensino fundamental.

A interpretação mais pessimista da evolução dos indicadores educacionais a partir de 2005 é de que, em virtude da falta de mudanças no ensino após os anos iniciais do ensino fundamental, os ganhos iniciais vão sendo perdidos ao longo dos anos finais do ensino fundamental e do ensino médio. As gerações beneficiadas pelo melhor ensino nos anos iniciais do ensino fundamental apresentariam, ao final do ensino médio, o mesmo desempenho que teriam atingido na ausência da melhoria do ensino nos anos iniciais. Todo esforço ocorrido nos anos iniciais seria perdido!

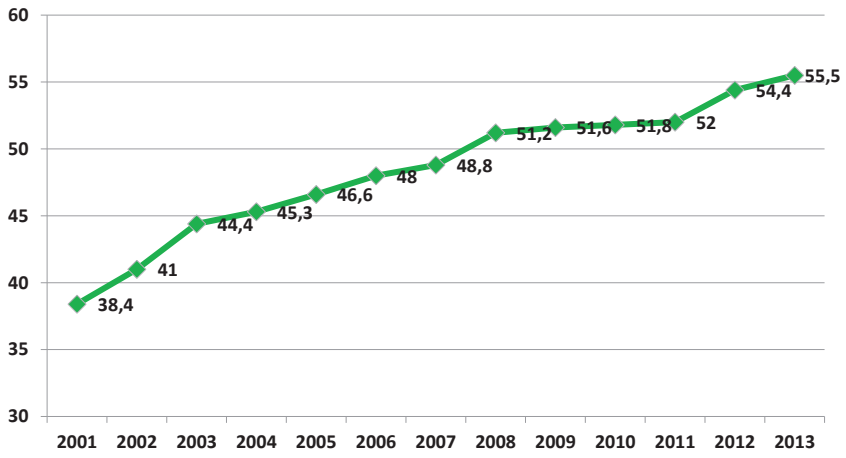
Entretanto, essa interpretação não é necessariamente correta e, do nosso ponto de vista, não parece ser a mais plausível. Em primeiro lugar, a composição daqueles que fazem a prova em determinada série pode estar mudando ao longo do tempo. Tais efeitos de composição podem ser decorrentes de alterações nos padrões de repetência e evasão. É possível também que seja mais difícil avançar na escala do SAEB quando a pontuação obtida é mais elevada. Assim, em termos de aprendizado, um crescimento de cinco pontos no final da primeira fase do ensino fundamental (por exemplo, de 210 para 215) pode ser menos significativo do que um crescimento de 2 pontos ao final do ensino médio (por exemplo, de 300 para 302).

O primeiro ponto (efeitos de composição) chama a atenção para o fato de que, na presença de repetência e evasão, não é tão claro como deveríamos avaliar se o desempenho dos estudantes está melhorando ao longo do tempo. Podemos pensar em três alternativas: (i) comparar o desempenho de determinada série ao longo do tempo, independentemente da idade dos alunos; (ii) comparar o desempenho de gerações sucessivas em determinada idade, independentemente da série cursada; (iii) comparar o desempenho de gerações sucessivas em determinada série, independentemente do ano em que o aluno é testado. Se todos os alunos ingressassem na escola na idade correta e não houvesse repetência nem evasão, essas três comparações seriam idênticas. No entanto, na presença de repetência e evasão, elas podem apresentar resultados muito distintos em caso dessas taxas sofrer alteração ao longo do tempo.

Por exemplo, uma redução da nas taxas de evasão pode levar a um aumento no desempenho entre gerações sucessivas, avaliadas em determinada idade, mas a uma redução no desempenho dos estudantes em determinada série ao longo do tempo. Isso porque uma parcela de estudantes de determinada geração (que antes não seriam esperados a atingir a série em questão) agora é observada na série e esses estudantes possuem, presumivelmente, desempenho inferior daqueles que, nas condições anteriores, seriam esperados a atingir a série em consideração.

Parece claro que tanto a repetência quanto a evasão vêm caindo ao longo do tempo. O Gráfico 8 mostra a evolução da taxa líquida de matrícula do ensino médio. Ela sugere que uma parcela de jovens que, nas condições que vigoravam 10 anos atrás, não chegariam ao final do ensino médio, hoje chegam. Como esses jovens, presumivelmente, possuem pior desempenho, esse fato poderia explicar, ao menos em parte, o porquê o crescimento observado na pontuação do SAEB e IDEB sete anos atrás na primeira fase do ensino fundamental não é observado hoje ao final do ensino médio.

Gráfico 8 – Taxa líquida de matrícula no ensino médio



Fonte: elaborado pelos autores dados básicos INEP.

Em relação ao segundo ponto (a dificuldade de avançar na escala do SAEB na medida em que pontuações mais elevadas vão sendo obtidas), ele tem sido observado em diversos exames que utilizam uma escala comum para diversas séries. Por exemplo, a tabela 1 apresenta as pontuações da amostra de normatização do *Comprehensive Test of Basic Skills* (CTBS), modelo U, desenvolvido pela CTB/McGraw-Hill, verticalmente escalado com base na TRI e utilizando um modelo logístico de três parâmetros. Em ambas disciplinas,

o ganho entre séries cai drasticamente entre as séries mais baixas e as mais elevadas. Além disso, o desvio padrão das pontuações diminui na medida em que a pontuação média sobe.

A redução do crescimento da pontuação média na medida em que a pontuação média aumenta pode ser uma consequência da forma como a escala é construída. Como destaca Ballou (2009), as escalas com base na TRI admitem que o aumento de proficiência necessário para elevar, em determinado montante, a probabilidade de responder um item corretamente (dado uma probabilidade inicial) é independente do grau de dificuldade do item. Isso significa que, considerando itens que se diferenciam apenas pelo grau de dificuldade, o aumento de proficiência requerido para um examinado com baixa proficiência elevar a probabilidade de responder corretamente um item fácil de 0,1 para 0,9 é o mesmo que o requerido para outro examinado de alta proficiência elevar a probabilidade de acerto de 0,1 para 0,9 de um item muito difícil. Ou seja, os conhecimentos e habilidades que o primeiro examinado precisaria adquirir têm a mesma medida que os conhecimentos e habilidades que o segundo examinado teria que adquirir.

Tabela 1 – Pontuação no *Comprehensive Test of Basic Skills* (CTBS/U) – 1981, Amostra de Normatização

Série	Leitura/Vocabulário			Matemática		
	Pontuação Média	Desvio Padrão	Variação	Pontuação Média	Desvio Padrão	Variação
1	488	85	-	390	158	-
2	579	78	91	576	77	186
3	622	65	43	643	44	67
4	652	60	30	676	35	33
5	678	59	26	699	24	23
6	697	59	19	713	20	14
7	711	57	14	721	23	6
8	724	54	13	728	23	7
9	741	52	17	736	17	8
10	758	52	17	739	16	3
11	768	53	10	741	18	2
12	773	55	5	741	20	0

Fonte: Yen (1986).

Alguém pode considerar que a noção de proficiência implícita nos modelos de TRI não casa com a sua noção intuitiva de aquisição de conhecimentos e habilidades. Desse modo, o fato de o crescimento da proficiência se reduzir na medida em que a proficiência aumenta, não significa, necessariamente, que o aprendizado diminui com a elevação da proficiência, uma vez que isso depende da noção de aprendizado que cada um possui.

Se for verdade que é mais difícil (envolve mais esforço) atingir uma probabilidade de acerto de 0,9 para um item muito difícil (quando os conhecimentos e habilidades adquiridos fixam uma probabilidade de acerto em 0,1), do que atingir essa mesma probabilidade para um item fácil (quando a probabilidade de acero é também 0,1), então, avançar na escala de proficiência vai se tornando mais difícil na medida em que se atingem níveis mais elevados de proficiência. Sendo assim, se ocorresse um aumento de esforço igual para todos os estudantes de todas as séries (ou uma melhora equivalente na qualidade de ensino), seria de esperar que o aumento de proficiência fosse maior nas séries iniciais (onde a proficiência é menor) do que nas séries mais elevadas. Do mesmo modo, o ganho de proficiência nas séries iniciais, em virtude de um melhor ensino, tende a se reduzir na medida em que vai se avançando no sistema, supondo que as condições de ensino tenham se mantido constante nas séries mais elevadas.

Note que em toda discussão acima foi admitido que proficiência é unidimensional, como é o caso do modelo de TRI adotado no SAEB. Enquanto a hipótese de unidimensionalidade é questionável para um exame restrito a uma única série, sua aplicação para uma escala vertical, cobrindo diversas séries, é, sem dúvida, bem mais problemática. Por exemplo, vamos admitir que o ensino de matemática nos anos iniciais do ensino fundamental se restrinja apenas à aritmética e que nos anos finais do ensino fundamental, além da aritmética, entrassem outros conteúdos, como álgebra e geometria. Assim, os itens do caderno de teste para os alunos do final da primeira fase do ensino fundamental cobririam apenas aritmética, enquanto os itens do caderno de teste para os alunos do final da segunda fase do ensino fundamental cobririam aritmética, álgebra e geometria. Vamos admitir agora que um aluno, ao final da primeira fase do ensino fundamental, apresente uma proficiência de matemática maior do que a de um aluno ao final da segunda fase do ensino fundamental. De acordo com a metodologia de ligação adotada pelo SAEB (grupos não equivalentes e itens comuns), isso significa que o aluno da primeira fase do ensino fundamental possui mais conhecimentos e habilidades em aritmética, mas não, necessariamente, em matemática. Isso porque o aluno da segunda fase do ensino fundamental pode ser capaz de acertar itens de álgebra e geometria que o aluno da primeira fase do ensino fundamental não é capaz, pelo simples fato de não ter sido exposto a esse material. Se o aluno dos anos iniciais do ensino fundamental fizesse o teste dos anos finais (ao invés do teste dos anos iniciais) apresentaria, provavelmente, um desempenho inferior ao verificado ao realizar o teste referente ao seu nível. Assim, o avanço da proficiência média entre os alunos da primeira e segunda fase do ensino fundamental “subestimaria” o “real” avanço médio de conhecimentos e habilidades matemáticas.

Em suma, mudanças na composição dos alunos que realizam o SAEB de ensino médio é a maior dificuldade de avançar na escala, na medida que a pontuação aumenta poderiam explicar, ao menos em parte, o porquê o desempenho no ensino médio não apresenta a mesma evolução do desempenho observado no ensino fundamental alguns anos antes.

CONCLUSÃO

Este artigo buscou avaliar o movimento de avaliação educacional em larga escala e *accountability* no Brasil. Realizou-se uma breve revisão da literatura internacional sobre *accountability* educacional, destacando os argumentos em sua defesa, a posição dos críticos e, principalmente, a avaliação de seus resultados. Traçou-se um histórico do movimento de avaliação em larga escala e *accountability* no país. Por fim, fez-se uma avaliação desse movimento.

O Brasil possui hoje um sistema de avaliação da educação básica que, sem exagero, poderia ser classificado entre os melhores do mundo. Mas, evidentemente, pode ser aprimorado. Além de retratar o passado, as avaliações têm o papel de sinalizar para o sistema o que se espera das escolas. Então, ao menos para a segunda fase do ensino fundamental e para o ensino médio, as avaliações deveriam ir além de leitura e matemática e incluir ciências da natureza e humanidades. No ensino médio, isso era feito com o ENEM, mas foi revertido com a decisão do INEP de não mais divulgar os resultados do ENEM por escolas, e universalizar o SAEB para as escolas públicas de ensino médio.

O artigo enfatiza as políticas de avaliação e *accountability* implantadas pelo Ministério da Educação. Em um outro trabalho, seria interessante apresentar e avaliar mais detalhadamente as experiências de avaliação e *accountability* realizadas por estados e municípios.

REFERÊNCIAS

AOS, S. PENNUCCI, A. K–12 **Class Size Reductions and Student Outcomes**: *a Review of the Evidence and Benefit-Cost Analysis*. Documento n. 13-01-2201, Washington State Institute for Public Policy, 2013.

BALLOU, D. Test Scaling and Value-Added Measurement. **Education Finance and Policy**, MIT Press, v. 4, n. 4, p. 351–383, 2009.

BAUER, A. et al. Iniciativas de avaliação do ensino fundamental em municípios brasileiros. **Revista Brasileira de Educação**, v. 22, n. 71, p. 1-19, out. 2017.

BEHRENDT, A.; EISENACH, J. JOHNSON, W. R. Selectivity Bias and the Determinants of SAT Scores. *Economics of Education Review*, 5 (4), p. 363-371, mês?? 1986.

BONAMINO, A.; FRANCO, C. Avaliação e Política Educacional: o Processo de Institucionalização do SAEB. **Cadernos de Pesquisa**, Rio de Janeiro, n. 108, p. 101-132, nov. 1999.

BOTELHO, F. B. et al. Sistemas de Accountability nas Escolas Públicas Brasileiras: Identificando a Eficácia das Diferentes Experiências. In: Fernandes, R., Souza, A. P. F., BOTELHO, F.; SCORZAFAVE, L. G. (orgs). **Políticas públicas educacionais e desempenho escolar dos alunos da rede pública de ensino**. Ribeirão Preto: FUNPEC-Editora, 2014, p. 59-80.

BROOKE, N. O futuro das políticas de responsabilização educacional no Brasil. **Cadernos de Pesquisa.** v. 36, n. 128, p. 377-401, maio./ago. 2006.

CASTRO, M. H. G. O Saeb e a Agenda de Reformas Educacionais: 1995- 2002. **Em Aberto**, Brasília, v. 29, n. 96, p 85-98, maio/ago. 2016.

CLARK, M.; ROTHSTEIN, J.; SCHANZENBACH, D. W. Selection Bias in College Admissions Test Scores. **Economics of Education Review**, Cambridge, n. 28, p. 295-307, aug. 2008.

COLEMAN, J. et al.; **Equality of Educational Opportunity**. Washington DC, 1966.

DYNARSKI, M. The Scholastic Aptitude Test: Participation and performance. **Economics of Education Review**, Elsevier, v. 6, n.3, p.263–273, jun. 1987.

DYNARSKI, M. GLEASON, P. Using Scholastic Aptitude Test Scores as indicators of state educational performance. **Economics of Education Review**, Elsevier, v. 12, n. 3, p. 203–211, jun. 1993.

FERNANDES, R. A universalização da avaliação e a criação do IDEB: pressupostos e perspectivas. **Em Aberto**, Brasília, v. 29, n. 96, mai./ago. 2016.

_____. Índice de Desenvolvimento da Educação Básica (Ideb*)*. Brasília: Inep, **Textos para Discussão,** Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 26p., 2007.

FERNANDES, R. GREMAUD, A.. Qualidade da Educação: Avaliação, Indicadores e Metas. In: Veloso, F., Pessôa, S., Henriques, R. e Giambiagi, F. (orgs). **Educação Básica no Brasil**: construindo o país do futuro. Editora Campus/ Elsevier, 2009, p. 213-238.

FIGLIO, D. LOEB, S. School Accountability. In Hanushek, E., Machin, S. e Woessmann, L. (Eds) **Handbook of the Economics of Education**, v. 3, North-Holland, 2011, p. 383-421.

FIRPO, S.; PIERI, R.;SOUZA, A. P.. Electoral Impacts of Uncovering Public School Quality: Evidence from Brazilian Municipalities. **Revista Economia,** v. 18, p. 1-17, 2017.

FLETCHER, P. R.; RIBEIRO, S. C. **Modeling education system performance with demographic data:** an introduction to the PROFLUXO model. Paris: Unesco, 1989.

GATTI, B. Avaliação educacional no Brasil: experiências, problemas, recomendações. **estudos em avaliação educacional**, Fundação Carlos Chagas, n.10, p. 67-80, 1994.

HANUSHEK, E. A. Grinding the antitesting ax: more bias than evidence behind NRC panel's conclusions. **Educational Next***,* v. 12, n. 2, p. 49-55, 2012.

HANUSHEK, E. A.; RAYMOND, M. Does school accountability lead to improved student performance? **Journal of Policy Analysis & Management**, v. 24, n. 2, p. 297-327, 2005.

HARBISON, R.W.; HANUSHEK, E. A. **Educational performance of the poor**. New York: Oxford University Press, 1992.

HORTA NETO, J. L. Um olhar retrospectivo sobre a avaliação externa no Brasil: das primeiras medições em educação até o SAEB de 2005. **Revista Ibero Americana de Educacion**, Brasília, v. 42. n. 5, p 1 -13, abr. 2007.

HOUT, M.; ELLIOTT, S. W. **Incentives and test-based accountability in education**. National Research Council of National Academies. Washington, DC: The National Academies Press, 2011.

KANE, T.; STAIGER, D. The promise and pitfalls of using imprecise school accountability measures. **Journal of Economic Perspectives**, v.16, n. 4, p. 91–114, fall 2002.

KLEIN, R.; RIBEIRO, S. C O Censo Educacional e o Modelo de Fluxo: o problema da repetência. **Revista Brasileira de Estatística**, v. 52, n. 197-198, p. 5-45, nov. 1991.

PEÑA, P. A. I**mpact of extension of elementary education in Brazil on test scores**. Estudo realizado para o programa Todos pela Educação, 2014. Disponível em: http://www.todospelaeducacao.org.br/biblioteca/1524/impact-of-elementary-school-extension-on-test-scores/. Acesso em: 1 de novembro de 2019

PESTANA, M. I. Trajetória do SAEB: criação, amadurecimento e desafios. *Em Aberto*, Brasília, v. 29, n. 96, p. 71-84, mai./ago/, 2016.

REARDON, S. F. RAUDENBUSH, S. W. Assumptions of value-added models for estimating school effects. **Education Finance and Policy**, v. 4, n. 4, p. 492-519, 2009.

ROCHA, R.; KOMATSU, B.; MENEZES-FILHO, N. . **Avaliando o impacto das políticas educacionais em Sobral, Texto para discussão**, Insper, n.35, out. 2018.

YEN, W. The choice of scale for educational measurement: an IRT perspective. **Journal of Educational Measurement,** v. 23, n. 4, p.299-325, 1986.

Recebido em: 03 de dezembro de 2019
Aceito em: 01 de junho de 2020
Publicado em: 30 de junho de 2020