2
Avaliação das diferentes abordagens para a associação entre itens e níveis ou pontos de uma escala de proficiência

A primeira parte deste trabalho tem como objetivo comparar cinco abordagens utilizadas para a interpretação de níveis de escalas de proficiência. Quatro delas se referem a programas de avaliação educacional consolidados em âmbito nacional ou internacional, das quais uma é utilizada pelo *Programa Internacional de Avaliação de Alunos* (*Program for International Student Assessment* – PISA); outra, pelo *Trends in International Mathematics and Science Study* (TIMSS); duas outras, pelo *National Assessment for Educational Progress* (NAEP) e pelo *Sistema Nacional de Avaliação da Educação Básica* (SAEB). Recentemente, a quinta abordagem começou a ser utilizada pelo *Estudo Longitudinal sobre a Qualidade e Eqüidade no Ensino Fundamental Brasileiro* (Projeto GERES – 2005).

Os resultados dos testes de Proficiência em Matemática do Sistema Nacional de Avaliação da Educação Básica (SAEB) 2003, da 4ª série do Ensino Fundamental, constituem a base experimental do estudo proposto, que focaliza a associação entre itens e níveis ou pontos de proficiência das escalas.

Esta investigação trata de procedimentos relativos aos dois principais processos ligados às funções e características de uma escala de proficiência: sua construção e sua interpretação. Elege-se, como norte, o debate de seus principais problemas e o enfrentamento dos desafios postos pela pesquisa de alternativas que visam a sua superação. Como meta, o desenvolvimento de metodologias mais eficazes de produção de escalas de proficiência em avaliações educacionais em larga escala.

Ambos os processos são dependentes da estrutura básica de qualquer escala de proficiência baseada na Teoria da Resposta ao Item (TRI). Apesar desses processos serem interdependentes, eles tratam de etapas e elementos característicos do desempenho escolar.

O processo de construção de uma escala de proficiência diz respeito às opções metodológicas subjacentes a procedimentos estatísticos, pedagógicos e operacionais e apresentam resultados de testes em uma única métrica, expressos em escores de proficiência dispostos em uma escala unidimensional passível de ser interpretada educacionalmente. Dentre esses e outros procedimentos, destacam-se os critérios de seleção, a utilização de parâmetros estimados por meio da TRI, as opções de equalização de escalas. Tais procedimentos se conectam com os principais elementos que nortearão a etapa seguinte, a tradução dos resultados da medida da habilidade em termos de seu significado cognitivo e educacional.

O processo de interpretação da escala de proficiência realiza essa tarefa. Uma vez garantidas a fidedignidade, a unidimensionalidade e a equalização da escala, importa a consideração de tais escores, de modo a viabilizar-se a tradução da medida de habilidade em uma especificação que resuma o conteúdo cognitivo educacional da medida. A interpretação da escala baseia-se na descrição dos níveis de proficiência em que se mostra o que os alunos, cujas proficiências localizam-se em cada nível, são capazes de fazer, ou seja, quais são as habilidades por eles desenvolvidas. Isso envolve a descrição e a interpretação pedagógica dos resultados, adequadas aos principais interessados neles, tendo como leitores prioritários os educadores, mas dirigidas também a gestores, famílias, especialistas, entre outros. Essa etapa de comunicação e publicidade dos resultados é de fundamental importância, para que a escala cumpra seus objetivos principais. Portanto, a escala deve estar organizada e disposta de modo a refletir os desafios de cada etapa da aprendizagem, de cada série avaliada, de cada etapa do desenvolvimento cognitivo típico do conteúdo (dimensão) que avalia.

A metodologia utilizada é a análise comparativa de procedimentos. A comparação de mérito relativo é feita por duas vias: em primeiro lugar, analisando-se os procedimentos que operacionalizam cada abordagem; e, em seguida, aplicando-se as cinco abordagens a um

mesmo conjunto de dados, ou seja, à base de dados do resultado do teste de Matemática – SAEB 2003, 4ª série do Ensino Fundamental, de modo que sejam exploradas as implicações de cada abordagem sobre a interpretação da escala. A análise comparativa entre os critérios da associação entre itens e níveis ou pontos de proficiência nas escalas será realizada, tendo-se como referência os critérios adotados pelo SAEB a partir de 1999. É importante assinalar que, com a abordagem comparativa, não se pretende estabelecer uma hierarquia rígida entre as qualidades de cada abordagem, mas, sim, entender em que condições cada abordagem pode oferecer uma melhor oportunidade de interpretação pedagógica do desempenho escolar característico de cada nível de proficiência. Para tal, dividiu-se esta seção em três partes: (i) primeira parte – uma breve apresentação dos conceitos básicos relacionados a escalas no âmbito da Teoria de Resposta ao Item; (ii) segunda parte – uma revisão do tratamento do tema da interpretação educacional de escalas no âmbito de importantes exercícios de avaliação em larga escala; (iii) terceira parte – uma apresentação de propostas sobre o delineamento e a testagem de abordagens alternativas para a interpretação de escalas.

2.1
Conceitos básicos: dos itens à Teoria da Resposta ao Item

Basicamente, os testes padronizados usados em avaliação educacional em larga escala são compostos de itens. Cada item tem o objetivo de avaliar uma única habilidade apresentada pelos descritores que compõem a matriz de referência para a avaliação. O descritor é uma associação entre conteúdos curriculares e operações mentais desenvolvidas pelos alunos, traduzindo determinadas habilidades e competências. Essa associação apresenta um resultado que orienta todo o processo de construção dos itens do teste de proficiência escolar. O conjunto de itens do teste visa a avaliar um conjunto de habilidades que

se quer medir, característico da competência do aluno em um determinado ciclo ou período de escolaridade.

A construção desses itens é tarefa que requer um elevado grau de complexidade técnica e exige conhecimentos específicos quanto à formulação do enunciado, do comando preciso para a resposta e das opções de resposta, conforme se encontram especificados, por exemplo, em Haladyna (1997), Kubiszyn (1990), Vianna (1993), bem como no Guia de Elaboração de itens do SAEB (BRASIL MEC-INEP, 2003).

Esses itens são pré-testados, ou seja, previamente aplicados a amostras de examinandos com o objetivo de estudar o comportamento dos itens. Isso é feito porque pode suceder que um determinado item não tenha um bom comportamento, decorrente de problemas em sua estruturação, tais como a incapacidade de distinguir, de modo claro, os avaliados que desenvolveram daqueles que não desenvolveram as habilidades que o item pretende mensurar. Quando esses itens são pré-testados, é possível que os problemas que eles apresentam sejam melhor observados, de modo que os itens mais problemáticos são eliminados, por via de um processo de seleção que determina um número fixo de itens a serem utilizados na avaliação.

A decisão sobre o número de itens é um ponto importante na composição do instrumento de medida. Por um lado, o teste deve conter tantos itens quantos necessários para que se produza uma medida abrangente de habilidades essenciais do período de escolaridade a ser avaliado. Por outro lado, o teste não pode ser excessivamente longo, pois inviabiliza sua resolução pelo examinando. Para solucionar essa dificuldade, tem-se utilizado um tipo de planejamento de testes denominado Blocos Incompletos Balanceados – BIB[1] – e a construção da medida baseada na metodologia da Teoria da Resposta ao Item – TRI.

[1] O planejamento dos cadernos de testes em BIB permite a organização dos itens em blocos, que são agrupados em cadernos, de tal modo que quaisquer dos cadernos tenham um, e somente um, bloco em comum. Com isso, consegue-se que um conjunto de alunos avaliados responda a um grande número de itens, enquanto cada um, individualmente, responderá apenas a um número razoavelmente pequeno de itens.

2.1.1
As principais características dos itens utilizados nas avaliações de larga escala

Os itens que compõem um teste de proficiência devem ser observados em relação a duas características importantes: seu grau de dificuldade e seu poder de discriminação. Essas características também devem ser observadas nas avaliações realizadas em sala de aula, pelos próprios professores.

Dificuldade. Naturalmente, a dificuldade de um item diz respeito à quantidade de proficiência que capacita o aluno avaliado a acertá-lo. Itens mais fáceis requerem menos proficiência e são acertados por um maior número de alunos; itens mais difíceis requerem maior proficiência. Segundo a Teoria Clássica do Teste (TCT), a dificuldade de um item é medida pela proporção ou porcentagem de alunos que o acertam. Portanto, na verdade, trata-se de uma medida de "facilidade", visto que, quanto maior a proporção de acertos, mais fácil tende a ser considerado o item.

Discriminação. Observando-se o comportamento da resposta do aluno avaliado em relação a um item específico, o poder de discriminação de um item é a característica que lhe permite oferecer informação sobre a proficiência desse aluno, ou compará-la com a de outro aluno que também está sendo avaliado. Certamente, deseja-se que os itens que compõem um teste tenham um elevado poder de discriminação nas respectivas habilidades mensuradas. Isso porque teria pouca validade um item que, por exemplo, tivesse um alto índice de acerto tanto pelos alunos de maior desempenho quanto pelos de pior desempenho, assim definidos com base no resultado que obtêm no teste como um todo. Igualmente, seria de pouca valia um item com índices baixos e semelhantes de acerto tanto entre os alunos com alto desempenho, quanto entre os alunos com baixo desempenho. Logicamente, teria de ser excluído da avaliação de uma determinada proficiência um item no qual os alunos de pior desempenho apresentassem um percentual de acerto maior que os alunos de melhor

desempenho. Em um caso desses, diz-se que o item e o teste têm uma correlação negativa, pois os melhores resultados no teste, correspondentes aos alunos de maior proficiência, geralmente, fazem-se acompanhar de erros no item, o que é um absurdo. Vale observar, entretanto, que, ocasionalmente, constatam-se itens com esse comportamento anômalo, sendo essa uma das razões pelas quais é importante fazer a pré-testagem dos itens, antes de aplicá-los em um teste de proficiência.

Na Teoria Clássica do Teste, utiliza-se a correlação item-teste para avaliar a discriminação do item. A correlação é uma medida estatística que varia entre –1 e 1. Como já mencionado, obviamente, não é interessante que os itens tenham correlação negativa, visto que esse problema costuma ser o responsável pela eliminação sumária de itens de um banco. Por outro lado, deseja-se que os itens utilizados tenham elevadas correlações positivas, sendo esses, geralmente, os selecionados para comporem os testes de proficiência.

2.1.2
Aspectos básicos da Teoria da Resposta ao Item (TRI)

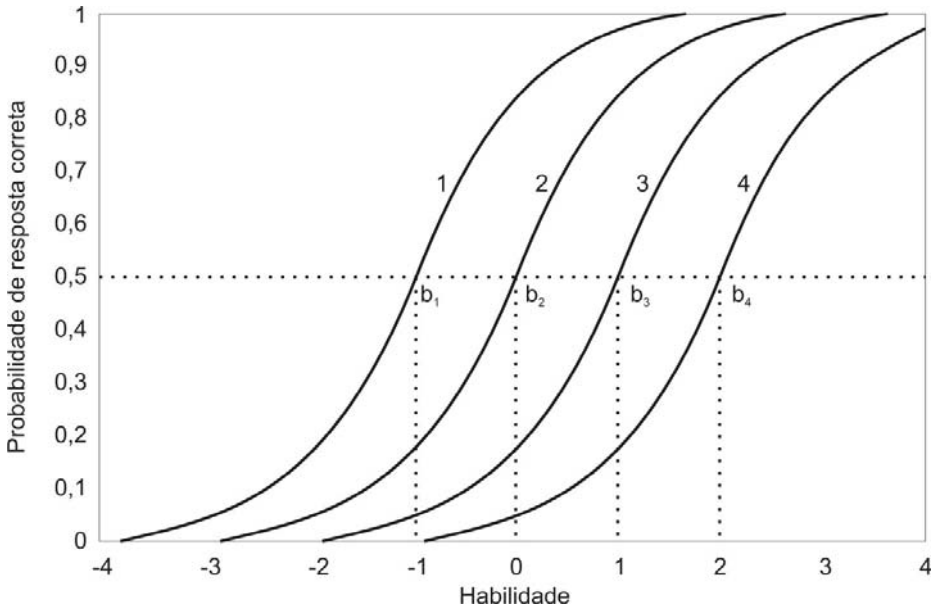
Se comparada à Teoria Clássica do Teste, segundo Lord & Novick (1964), a utilização da TRI – conjunto de modelos matemáticos no qual a probabilidade de resposta a um item é modelada em função da proficiência do aluno, variável não-observável – baseia-se em pressupostos fortes quanto ao comportamento de um indivíduo que responde aos itens de um teste, o que confere a ela algumas vantagens na elaboração de modelos de teste de avaliação de proficiência escolar. Dentre essas, destacam-se: (i) a possibilidade de comparação longitudinal de resultados de diferentes avaliações, como, por exemplo, as da avaliação de sistemas estaduais de ensino e os resultados do SAEB, desde que se incluam itens comuns aos testes e se conservem os mesmos critérios na construção e organização dos testes e na análise dos resultados; (ii) a possibilidade de avaliar com alto grau de precisão e

abrangência uma determinada área do conhecimento, sem que cada aluno precise responder a longos testes; (iii) a possibilidade de comparação entre diferentes séries, por exemplo, 4ª e 8ª séries do Ensino Fundamental e 3ª série do Ensino Médio, viabilizada pela construção de uma escala única de resultados para essas três séries (Hambleton, Swaminathan e Rogers 1991; Hambleton, 1993; Valle,1999).

Em particular, duas restrições de especial relevância para os modelos da TRI são a unidimensionalidade e a independência local. A primeira postula a homogeneidade do conjunto de itens que, supostamente, devem estar medindo um único traço latente (LORD, 1980). Isto é, postula que há apenas uma habilidade responsável pelos resultados dos alunos em um conjunto de itens, ou, mais provavelmente, que ela seja significativamente dominante entre todas as possíveis habilidades requeridas para a realização do teste. A segunda pressupõe que, para uma dada habilidade, as respostas aos diferentes itens do teste não se influenciam, ou seja, mantidas as habilidades que afetam o teste, as respostas dos alunos a quaisquer dos itens são estatisticamente independentes.

A relação entre a proficiência e a probabilidade de o aluno acertar o item pode ser descrita por uma função matemática monotônica crescente, denominada Curva Característica do Item – CCI, como mostra a Figura 1.

Figura 1 – CCIs de quatro itens de diferentes níveis de dificuldade, segundo o modelo de um parâmetro da TRI



Nesse exemplo, a CCI é apresentada graficamente, mostrando-se a relação entre proficiência e probabilidade de acerto para quatro itens, numerados de 1 a 4. Geometricamente, essa relação se comporta como uma curva, que tem uma fórmula específica e é conhecida tecnicamente pelo nome de curva característica do item (CCI). Algumas características importantes dessa curva podem ser inferidas diretamente da figura, entre as quais é possível destacar: (i) a correlação positiva entre a habilidade (proficiência) e a probabilidade de acerto, ou seja, observa-se o aumento da probabilidade de acerto, à medida que a proficiência também aumenta. Esse aumento da probabilidade não é o mesmo para todos os valores de habilidade, sendo mais intenso no centro da curva do que nos seus extremos; (ii) a não-obediência da curva a um padrão linear; por exemplo, se a habilidade dobrar, não necessariamente dobra a chance de acerto. Isso é particularmente visível, por exemplo, nas extremidades direitas de cada curva, correspondentes aos valores maiores de habilidade, quando as alturas de cada curva vão-se fixando em torno da probabilidade P = 1, indicando que, em um determinado item, a chance de acerto é próxima de 100% para um determinado nível de habilidade, e não tem como crescer muito mais com referência a

habilidades ainda maiores, de modo que, nessa região, a curva passa a subir de modo muito suave, quase imperceptivelmente.

2.1.3
Os parâmetros da TRI

Na Teoria da Resposta ao Item, as curvas características dos itens podem ser especificadas por meio de três parâmetros. A seguir:

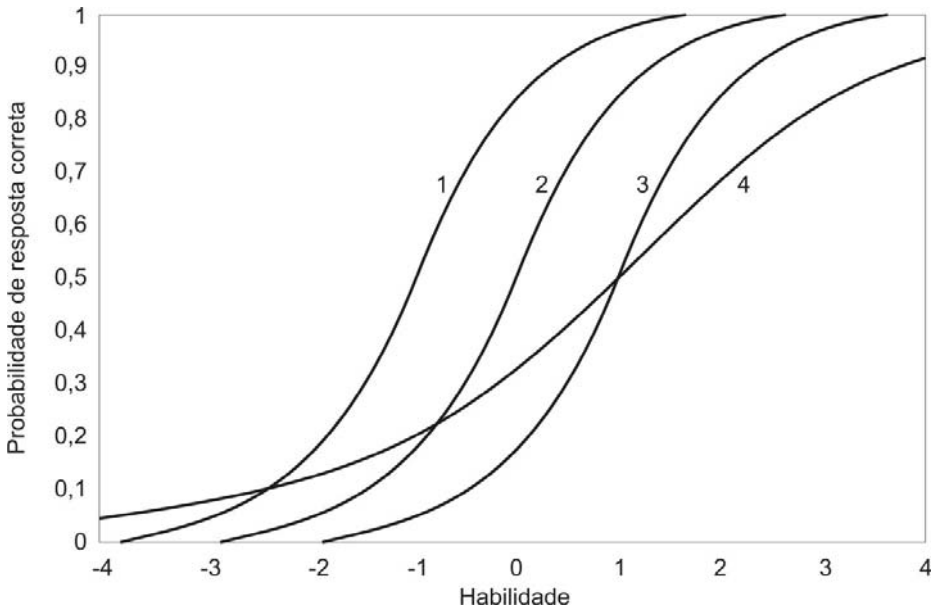
Parâmetro de dificuldade. Freqüentemente conhecido pela letra **b**, esse parâmetro mede a dificuldade de um determinado item, correspondendo à proficiência necessária, para que o percentual de acerto de um item seja de 50% (considerando-se que não existe chance de acerto casual), como ocorre nos itens de múltipla escolha. Conseqüentemente, itens de maior dificuldade apresentam um maior valor de **b**. Quando itens de diferentes dificuldades são representados em um mesmo gráfico, como na Figura 1, observa-se que os itens formam um bloco de curvas idênticas, excetuando-se o fato de que elas estão horizontalmente deslocadas umas em relação às outras. Assim sendo, as curvas mais deslocadas à direita correspondem às CCIs de itens mais difíceis, que requerem maior proficiência para garantir 50% de chance de acerto, ao passo que os itens mais fáceis situam-se à esquerda, sendo menor o nível de proficiência requerido, respeitada a probabilidade de acerto de 50%. Na Figura 1, pode-se, portanto, perceber que o item mais fácil é o 1, e o mais difícil, o 4.

Parâmetro de discriminação. Geometricamente, esse parâmetro corresponde à inclinação da CCI no ponto em que há 50% de chance de acerto do item, considerando-se nula a chance de acerto casual, também nesse caso. Quanto maior for esse parâmetro, denominado parâmetro **a**, maior será a inclinação da curva nesse ponto, e melhor a capacidade do item discriminar alunos de proficiências diferentes. A princípio, o parâmetro de discriminação **a** pode assumir qualquer valor real, positivo ou negativo. Uma condição, entretanto, necessária à validade do item é que o parâmetro **a** seja positivo, pois um valor

negativo para **a** indicaria que, quanto maior a proficiência do aluno avaliado menor sua chance de acertar o item, o que configura uma situação absurda. Em relação aos valores positivos de **a**, quanto maiores esses valores, maior é a capacidade de discriminação do item (naquele ponto), e maior sua qualificação para figurar no teste. Geometricamente, uma inclinação positiva e acentuada indica que, naquele ponto, ocorre um drástico aumento na probabilidade de acerto de um item, quando se verifica um pequeno aumento na habilidade dos alunos avaliados. Tal fato corresponde a uma maior sensibilidade do item em distinguir os alunos que sabem dos que não sabem, ou seja, os alunos que desenvolveram a habilidade requerida pelo item dos que não a desenvolveram.

Pode-se perceber na Figura 2, por exemplo, que os itens de 1 a 3 têm discriminações aproximadamente iguais entre si, visto que, para a probabilidade de acerto de 0,5, suas inclinações são praticamente iguais. Já o item 4 tem um menor poder de discriminação, pois sua inclinação para esse mesmo valor de probabilidade (0,5) é menor que nos casos anteriores.

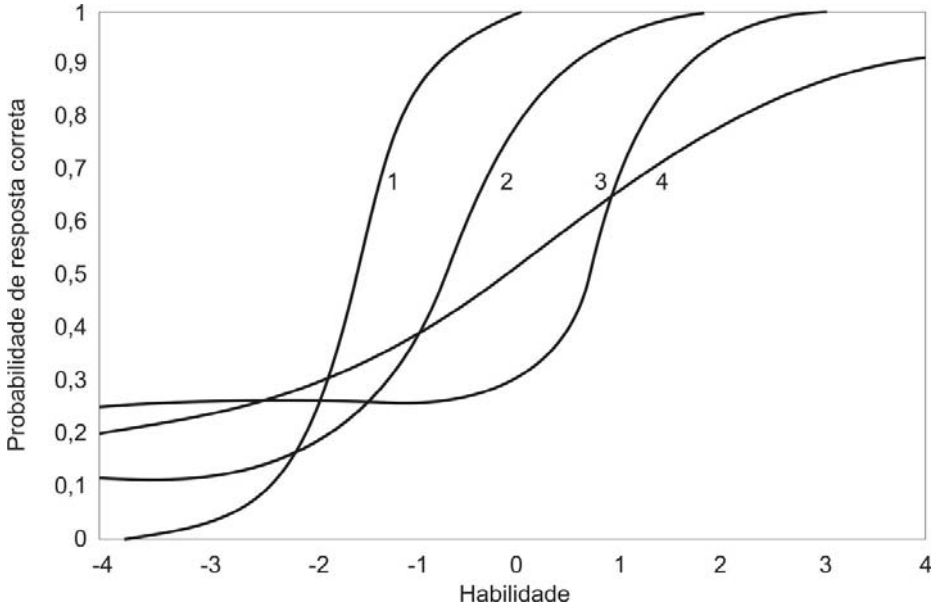
Figura 2 – CCIs de quatro itens com variações de dificuldade e de discriminação, segundo o modelo de dois parâmetros da TRI



Um aspecto importante é que a discriminação de um item não é a mesma para todo o intervalo de proficiência. Ou seja, um item que tem uma grande capacidade de discriminar os alunos de proficiência baixa dos alunos de proficiência média pode não ser muito útil para discriminar alunos de proficiência alta de alunos de proficiência ainda maior, visto que o item seria acertado por esses últimos grupos, de modo que não se poderia distinguir um grupo do outro. Um item assim poderia ser o de número 1, na Figura 2, correspondendo ao fato de que seu parâmetro de dificuldade **b** é menor. Certamente, um dos objetivos das avaliações de proficiência em larga escala é fazer distinções de proficiência entre alunos situados ao longo de todo o nível de proficiência de interesse. Assim sendo, quando os testes são elaborados, procura-se fazer com que eles se componham de itens com diferentes parâmetros de dificuldade, de modo a abranger de maneira uniforme todo o espectro relevante de proficiência.

Parâmetro do acerto casual. Esse parâmetro, denominado de parâmetro **c**, é utilizado para tratar do freqüente caso de itens de múltipla escolha, nos quais o aluno pode acertar um item, mesmo tendo uma proficiência relativamente baixa: trata-se do conhecido acerto casual, popularmente conhecido como "chute". Nas CCIs que incluem o parâmetro **c**, a extremidade esquerda corresponde a assíntotas, curvas que se aproximam de uma reta, sem, entretanto, tocá-la, que, por sua vez, referem-se a probabilidades de acerto superiores a zero. Por exemplo, em um item de múltipla escolha com cinco alternativas para respostas que parecem igualmente possíveis de ser escolhidas por um aluno, a chance de acerto casual é de 0,20 ou 1/5, uma vez que o aluno teria que acertar casualmente uma alternativa dentre as cinco apresentadas. Um caso desses é ilustrado pelo item de número 4 da Figura 3.

Figura 3 – CCIs de quatro itens com variações de dificuldade e de discriminação, levando-se em conta também o acerto casual, segundo o modelo de três parâmetros da TRI



Nem sempre, entretanto,o parâmetro **c** corresponde exatamente à divisão de 1 pelo número de alternativas de resposta ao item. Isso ocorre porque alternativas podem facilitar o desempenho de quem sabe menos, por exemplo, alternativas com respostas absurdas ou obviamente falsas, casos em que alunos de menor proficiência são "empurrados" na direção da resposta certa, aumentando assim sua chance de sucesso. Geometricamente, na CCI, isso se traduziria em uma assíntota esquerda mais alta, por exemplo, em um item de múltipla escolha com cinco alternativas para a resposta; essa assíntota poderia corresponder a 0,25, como é o caso do item 3 da Figura 3. Nesse caso, a proporção dos que acertaram o item, mesmo conhecendo muito pouco o conteúdo, foi maior do que aquele que seria de se esperar, levando-se em conta somente o número de alternativas no item. Por outro lado, em alguns itens, pode haver a ocorrência de distratores (que são as alternativas de respostas falsas ou não-preferíveis dos itens), capazes de convencer um grande número de alunos a considerá-los como a opção certa. Em casos assim, itens de múltipla escolha com cinco opções poderiam ter um parâmetro c inferior a 0,2, como, por exemplo, o item 2 da Figura 3.

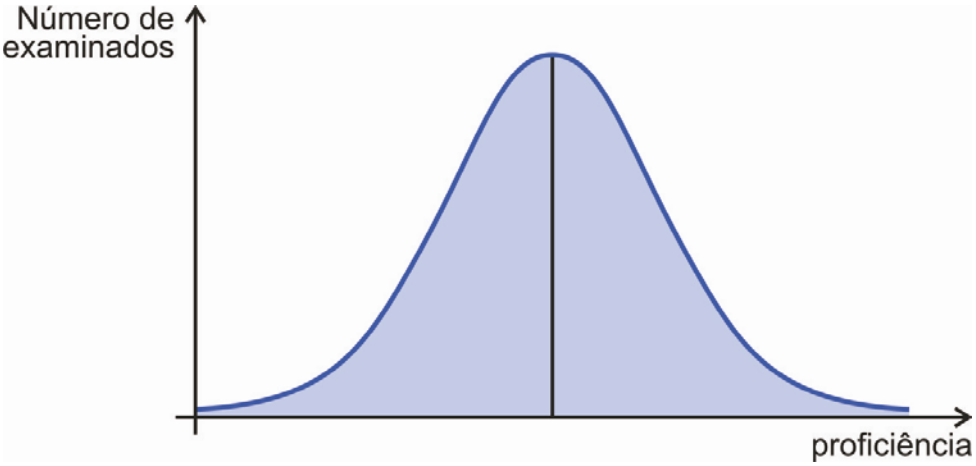
A invariância dos parâmetros da TRI. Pode-se constatar que a principal e mais importante distinção entre a Teoria da Resposta ao Item e a Teoria Clássica é a propriedade de invariância, característica da TRI. Os parâmetros do item não dependem da distribuição dos alunos avaliados segundo o nível de proficiência, e os escores de proficiência dos alunos avaliados não dependem do conjunto de itens utilizados para estimá-los. Quando o modelo da TRI ajusta-se aos dados, a curva característica de um item é a mesma, independente do grupo de alunos avaliados submetidos à estimativa dos parâmetros.

2.2
As Escalas de Proficiência

2.2.1
Considerações gerais

Segundo a Teoria da Resposta ao Item – TRI, para cada item, é calculada sua respectiva curva característica, que relaciona a habilidade do aluno avaliado com a sua probabilidade de acertar o item. Dessa forma, com base no padrão de acertos do examinado, é possível estimar sua habilidade. Nos programas computacionais usualmente utilizados na produção das medidas da TRI (como o BILOG MG e outros), as proficiências dos examinados formam um conjunto numérico com média igual a zero e desvio-padrão igual a 1. A média é o "ponto de equilíbrio" da distribuição das proficiências. Em geral, supõe-se que essa distribuição seja simétrica, podendo ser representada por uma curva chamada de "normal", ou curva em forma de sino, como a mostrada na Figura 4. Nessa curva, a proficiência é medida no eixo horizontal, de modo que os níveis de proficiência aumentam da esquerda para a direita. Por outro lado, a altura da curva, medida no eixo vertical, diz respeito à quantidade, ou à proporção de examinados que têm uma determinada proficiência.

Figura 4 – A curva normal de proficiência



Cabe mencionar que, para grandes números de alunos examinados, como ocorre nos testes de larga escala, que chegam a ter centenas de milhares de alunos avaliados, as distribuições verdadeiras dos resultados aproximam-se bastante bem da curva normal. Dessa forma, como se pode ver na figura 4, a maioria dos alunos avaliados tem sua proficiência situada em níveis médios. Isso se traduz na maior altura da curva no centro e em torno dele. Porém, à medida que se vai caminhando para as proficiências maiores situadas à direita do gráfico e também para as proficiências menores situadas à esquerda, vai diminuindo o número de casos encontrados. Assim, os resultados dos testes tendem a concentrar-se em torno da média; por outro lado, resultados muito melhores ou muito piores do que a média são mais raros, e quanto maior a discrepância em relação à média, maior é essa raridade, ou seja, menor é a altura da curva. Em uma distribuição normal padronizada de resultados, atribui-se o número zero à proficiência do meio da curva, ou seja, a média corresponde ao número 0; e o número 1, ao seu desvio-padrão.

2.2.2
Transformações lineares

Para fins de divulgação de resultados, é mais conveniente que se exprimam os resultados de um teste em uma escala diferente da normal padronizada. Em geral, ao verem e interpretarem uma escala, as pessoas preferem lidar com números positivos e inteiros a lidar com números negativos e fracionários. Essa conversão de dados normais padronizados para os valores de uma escala arbitrária é usualmente feita por um procedimento matemático simples chamado transformação linear. Segundo esse procedimento, um número *x* pode ser convertido em um número *y* por meio da equação $y = a + bx$, onde a e b são constantes quaisquer, geralmente determinadas de modo a fazer com que os valores de *y* sejam expressos em uma escala conveniente.

Portanto, as transformações lineares permitem que os resultados sejam expressos em escalas arbitrárias. Essa liberdade na escolha das escalas possibilita a escolha de números de uso e memorização mais fáceis por parte de quem utilizará os resultados dos testes, como professores e gestores escolares. Por isso, é comum que essas médias tenham valores como, por exemplo, 150 ou 200, e que os desvios-padrão valham algo como 20 ou 50 pontos.

Então, ao *continuum* de proficiência são atribuídos os números da transformação linear adotada, com uma subseqüente divisão desse *continuum* em níveis de proficiência. Esses níveis são arbitrários, uma vez que, no *continuum* de habilidades considerado, não existe nenhum ponto privilegiado onde uma determinada habilidade se separa drasticamente de outra adjacente. Mas, além disso, deve-se obedecer a critérios, de modo que esses níveis não sejam definidos num número excessivamente pequeno, tornando-se pouco informativos (por exemplo, uma escala com apenas dois níveis, baixo e alto), nem num excessivamente grande (por exemplo, escalas com número alto de intervalos, como 40), inviabilizando-se a análise gradativa dos níveis em relação às habilidades desenvolvidas.

2.2.3
Dando significado às Escalas de Proficiência

Considerando-se os pressupostos da TRI, as escalas de proficiência são obtidas a partir de um tratamento estatístico desenvolvido em três etapas: (i) calibração dos itens do teste, o que ocorre por meio da identificação de seus parâmetros, uma vez que os parâmetros dos itens e as proficiências são invariantes. Uma hipótese habitual, e necessária para a identificação da proficiência, supõe, por um lado, que os parâmetros dos itens sejam invariantes para os diferentes grupos de examinandos, e, por outro, que a proficiência dos alunos seja invariante para o conjunto de itens aplicados, exceto por uma transformação linear, especificamente afim, da escala; (ii) após a calibração e análise do funcionamento diferencial dos itens, realizam-se os procedimentos básicos de equalização da escala, de acordo com critérios pré-estabelecidos; e (iii) em uma única etapa, são calculadas, por exemplo, a proficiência média do alunato por redes de ensino, por municípios, por unidades federativas, por regiões do país, por países.

Construir e dar significado aos números de uma escala de proficiência significa escolher alguns pontos ou escolher alguns níveis e descrever as habilidades que os alunos demonstram possuir, quando situados em torno desses pontos ou níveis de proficiência. Ou seja, depois de identificados os itens representativos de cada nível ou ponto, especialistas da área de conhecimento avaliada procuram explicar o significado pedagógico das respostas dadas aos itens do teste, a partir de uma descrição das habilidades desenvolvidas e consolidadas por meio da análise das respostas dadas aos diferentes itens característicos de cada nível de proficiência.

Os resultados obtidos são interpretados por meio das escalas de proficiência que apresentam ordenadamente, em um *continuum*, o desempenho dos avaliados, do nível mais baixo ao mais alto. Então, os resultados da avaliação em larga escala são demonstrados em uma escala de proficiência apresentada em níveis, como, por exemplo, a escala de proficiência do SAEB, que varia aproximadamente de 0 a 500 pontos, de

modo a conter, de forma bem distribuída, em uma mesma métrica, os resultados do desempenho escolar dos alunos da 4ª e 8ª séries do Ensino Fundamental e 3ª série do Ensino Médio.

Assim, os alunos situados em um nível mais alto da escala revelam dominar não só as habilidades do nível em que se encontram, mas também aqueles níveis anteriores. Quem está no terceiro nível de proficiência domina também as habilidades características no segundo nível e no primeiro; quem está no último nível revela também as habilidades de todos os níveis anteriores. Analisar cuidadosamente a descrição das habilidades características de cada nível de proficiência produz um diagnóstico do desempenho escolar

As escalas de proficiência também permitem situar a "unidade" avaliada, seja ela aluno, escola, município, estado federativo ou país, em função de seu desempenho, possibilitando a comparação dos resultados obtidos. Também podem ser comparadas as médias de proficiência alcançadas entre os períodos de escolaridade avaliados. Por exemplo, os alunos da 8ª série devem, necessariamente, revelar habilidades mais complexas do que os da 4ª série, devendo, portanto, estar situados em pontos mais altos da escala. Ao interpretar-se a escala de proficiência, pode-se ainda observar o percentual de alunos em cada nível e comparar com o percentual dos alunos que se encontram no nível, ou acima do nível adequado, isto é, o nível em que se apresentam as habilidades básicas e essenciais para o período de escolaridade avaliado. Ou seja, por meio da interpretação da escala de proficiência, pode-se produzir um diagnóstico do desempenho escolar; como, por exemplo, em um desenho censitário, como é o caso da Prova Brasil[2] – Avaliação do Rendimento Escolar –, o desempenho dos alunos de cada escola, município, unidade da federação e do Brasil pode ser interpretado pedagogicamente em relação às médias de desempenho,

[2] A Prova Brasil, instituída pelo Sistema de Avaliação da Educação Básica, foi idealizada para produzir informações sobre o ensino oferecido por município e escola, individualmente, com o objetivo de auxiliar os governantes nas decisões e no direcionamento de recursos técnicos e financeiros, assim como a comunidade escolar no estabelecimento de metas e implantação de ações pedagógicas e administrativas, visando à melhoria da qualidade do ensino. Os resultados da Prova Brasil foram produzidos na mesma escala do SAEB 2001 (Prova Brasil – Avaliação do Rendimento Escolar – Resultados 206:1).

posicionadas na escala e comparadas entre si e com os ciclos de avaliações anteriores, bem como pela distribuição dos percentuais de alunos por nível de proficiência, o que propicia as seguintes informações: (i) quanto maior o percentual de alunos nos níveis mais altos da escala e menor o percentual nos níveis mais baixos, melhor é o resultado alcançado; (ii) se os percentuais de alunos se distribuem de modo significativo em todos os níveis da escala, com valores aproximados, essa situação configura um resultado heterogêneo, que requer um tratamento pedagógico; (iii) se os alunos concentram-se nos níveis mais baixos da escala, o resultado é insatisfatório; é preciso verificar se as habilidades apresentadas nos níveis mais altos da escala foram trabalhadas com os alunos.

2.2.4
Construção da escala de proficiência: critérios de seleção de itens representativos

A interpretação das escalas de proficiência tem início usualmente com a definição de critérios de seleção de itens representativos, que serão abordados a seguir, tendo como referência os principais programas de avaliação, a saber: o *Sistema Nacional de Avaliação da Educação Básica* – SAEB; o *National Assessment for Educational Progress* – NAEP; o *Programa Internacional de Avaliação de Alunos* – PISA; o *Trends in International Mathematics and Science Study* – TIMSS; e os critérios utilizados recentemente pelo Projeto GERES/2005 – *Estudo Longitudinal sobre a Qualidade e Eqüidade no Ensino Fundamental Brasileiro*, que é uma pesquisa longitudinal na qual uma amostra de alunos e escolas de cinco importantes cidades brasileiras está sendo observada ao longo de quatro anos iniciais do Ensino Fundamental.

Neste estudo serão adotados como referência os itens e os resultados de um teste de proficiência em Matemática para a 4ª série do Ensino Fundamental do SAEB, que utilizou itens de múltipla escolha, pré-testados, validados e integrantes do Banco Nacional de Itens – BNI/SAEB.

2.3
Principais referências de abordagens para a associação entre itens e níveis ou pontos característicos de uma escala de proficiência

2.3.1
O SAEB

2.3.1.1
Objetivo e ciclo de avaliação

No início de 1990, o Ministério da Educação e do Desporto (MEC), através do Instituto Nacional de Estudos e Pesquisas Educacionais (INEP), iniciou a implementação do Sistema de Avaliação da Educação Básica (SAEB), com o objetivo de gerar e organizar informações sobre a qualidade, a eqüidade e a eficiência da educação nacional, de modo a permitir o monitoramento das políticas educacionais brasileiras, fornecendo subsídios, para que gestores de políticas públicas, diretores e professores efetuassem mudanças necessárias à melhoria da qualidade da educação e garantia da igualdade de oportunidades educacionais (MEC/INEP, 1995; Pestana, 1998; MEC/INEP 2002). E ainda, de acordo com Castro (1999), com essa iniciativa, o MEC assumia, efetivamente, seu papel de coordenação e monitoramento das políticas educacionais, já que caberia à instância federal o monitoramento de políticas e a orquestração de políticas de estados e municípios, para o que as informações quantitativas e qualitativas eram absolutamente essenciais.

Desde seu início até hoje, o SAEB realizou nove ciclos de avaliação amostral, em 1990, 1993, 1995, 1997, 1999, 2001, 2003, 2005 e 2007, com base em amostras probabilísticas complexas das diversas unidades federativas. São testados alunos das escolas públicas e privadas das 4ª e 8ª séries do Ensino Fundamental e da 3ª série do Ensino Médio. A partir de 1995, em seu terceiro ciclo de avaliação, o SAEB passa a utilizar novas metodologias de elaboração de testes e técnicas estatísticas de análise e interpretação de resultados, com base na Teoria da Resposta ao Item (TRI), que deu origem à construção de um banco de itens.

Até 1997, os alunos testados, de 4ª e 8ª séries, fizeram testes de Língua Portuguesa, Matemática e Ciências; para a 3ª série do Ensino Médio, as Ciências se desdobravam em Biologia, Física e Química. Em 1999, os testes avaliam também História e Geografia. A partir de 2001, entretanto, os alunos fizeram apenas testes de Língua Portuguesa e Matemática.

Além dos testes, o SAEB aplica, concomitantemente, quatro questionários contextuais: (i) sobre a escola, (ii) dirigido ao diretor, (iii) dirigido ao professor e (iv) dirigido ao aluno, os quais têm como objetivo produzir informações referentes ao perfil sócio-econômico e à trajetória escolar dos alunos, a práticas na escola e seu impacto sobre a aprendizagem, a fatores sociais que afetam a probabilidade de repetência, ao estilo pedagógico dos professores e à modalidade de gestão e liderança na escola, entre outros.

2.3.1.2
A seleção de itens representativos utilizada pelo SAEB a partir de 1999

Na escala de proficiência do SAEB, a definição dos níveis de proficiência relaciona-se com a seleção de um conjunto de itens, onde um dado item caracteriza um ponto ou um nível da escala em que a grande maioria de alunos, situados nesse nível, acerta o item, enquanto um percentual considerável de alunos, situados no nível abaixo da escala, erra o item.

Nas avaliações realizadas pelo SAEB, a partir de 1999, um item é selecionado para um determinado nível se atende aos seguintes critérios: (i) o número de alunos no intervalo que está sob análise deve ser maior que 50; (ii) o percentual de acertos no item no intervalo anterior ao intervalo sob análise é menor que 65%; (iii) o percentual de acertos no item no nível considerado e nos níveis acima é maior ou igual a 65%; (iv) o ajuste aos dados do modelo da TRI, na estimativa das estatísticas do item, é bom. O Quadro 1 apresenta sucintamente os critérios mencionados.

Quadro 1

SAEB 1999 a 2007	Posicionamento dos itens na escala/ critérios de seleção: • O número de alunos no intervalo que está sob análise deve ser maior que 50. • O percentual de acertos do item no intervalo anterior ao intervalo sob análise é menor que 65%. • O percentual de acertos do item no nível considerado e nos níveis acima é maior ou igual a 65%. • O ajuste da TRI do item é bom.

Considerando-se os critérios estabelecidos pelo SAEB de 1999 a 2005 e as estatísticas calculadas para a análise dos itens e dos testes de Matemática do SAEB – 2003, 4ª série do Ensino Fundamental, foi produzida a Tabela 1 a seguir, em que se destacam os itens característicos dos níveis de proficiência correspondentes. Embora fossem aplicados 169 itens, neste estudo serão considerados 168, uma vez que o item de número 25060 foi excluído por apresentar comportamento estatístico inadequado, como pode ser observado no Anexo VIII.

Identificados os itens selecionados para cada nível, passa-se à construção da Escala de Proficiência em Matemática, tendo como referência os critérios de seleção utilizados pelo SAEB a partir de 1999, que se encontram no Anexo I.

Quanto à determinação das fronteiras desses níveis, vale ressaltar que a mesma foi realizada com certo grau de arbitrariedade, visto não haver, no *continuum* de proficiência, quaisquer pontos específicos que deveriam ser necessariamente considerados como fronteiras naturais dos níveis definidos. Entretanto, as escolhas feitas levaram em conta o fato de que, na escala do SAEB, a média de proficiência para a 8ª série do Ensino Fundamental corresponde a 250 pontos, e a partir deste valor foram construídos os níveis com largura correspondente a valores de um desvio-padrão (50 pontos) ou à sua metade (25 pontos). Essa variação de largura de acordo com os níveis, resultante de um procedimento igualmente imbuído de certa arbitrariedade, deveu-se ao fato de que é muito maior a concentração de alunos nos níveis mais centrais de

proficiência para a série considerada, razão pela qual se achou necessário fazer uma divisão mais fina entre esses níveis centrais.

Pode-se observar na Escala de Proficiência em Matemática do Anexo I que todos os 168 itens puderam ser selecionados, o que é uma conseqüência de os critérios de seleção terem sido pouco exigentes. Apenas deixariam de ser selecionados itens excessivamente difíceis, que não atingissem 65% de acerto nem mesmo no nível superior de proficiência (350 a 375 pontos). Não foram observados, entretanto, casos desse tipo, como mostra a distribuição do número de itens selecionados obtidos por nível de proficiência, apresentada na Tabela 1.

Tabela 1 – Número de itens selecionados por nível de proficiência, segundo os critérios do SAEB 1999-2007

Nível de Proficiência	N. de Itens selecionados	%	% Acum.
Até 125	0	0	0
125 – 150	3	1,8	1,8
150 – 175	15	8,9	10,7
175 – 200	24	14,2	24,9
200 – 250	48	29	53,8
250 – 300	51	30,2	84
300 – 350	19	11,2	95,3
350 – 375	8	4,7	100
Total	168	100	

Pode se constatar a presença de itens selecionados em todos os níveis de proficiência, excetuando-se o primeiro (correspondente ao nível mínimo de proficiência, de até 125 pontos). Essa ausência de itens selecionados no nível inferior é, entretanto, esperada, visto que a definição dos itens selecionados, segundo o presente critério, requer que o percentual de acerto no nível inferior ao da seleção seja menor que 65%, e, naturalmente, não pode existir um nível inferior ao mínimo considerado.

O fato de todos os níveis serem contemplados com itens selecionados é desejável, pois isso permite uma interpretação da escala de proficiência em todos os níveis considerados, Nesse caso, fica

evidente que o número de itens selecionados, distribuídos ao longo dos níveis de proficiência, obedece a um padrão aproximadamente simétrico, segundo uma distribuição grosso modo normal: há uma concentração maior de itens selecionados nos níveis centrais de proficiência, com uma diminuição paulatina do número deles, à medida que se consideram os níveis extremos, tanto na direção dos níveis maiores, quanto dos níveis menores de proficiência.

2.3.1.3
A seleção de itens representativos utilizada pelo SAEB em 1995 e 1997

Nas avaliações SAEB 95 e 97, um item foi considerado representativo de um dado nível, quando: (i) ele foi respondido corretamente por um grande percentual de alunos avaliados (de pelo menos 65% dos alunos no intervalo sob análise); e (ii) foi respondido corretamente por uma pequena porcentagem (não mais de 50%) de alunos no intervalo imediatamente anterior; (iii) a diferença entre os percentuais de alunos entre esses dois intervalos, que acertam e erram o item, deve ser de pelo menos 30 pontos percentuais. Por exemplo, para construir-se essa escala, um item pode ser considerado representativo do nível 250 se for satisfeito o seguinte critério: no nível 250, que 65% ou mais dos respondentes acertem o item; que menos de 50% dos alunos posicionados no nível anterior (175) acertem o item; e que a diferença entre os percentuais dos que acertaram seja maior que 30 pontos percentuais (BRASIL/MEC/INEP/SAEB,2001). O Quadro 2 apresenta sucintamente os critérios de seleção acima mencionados.

Quadro 2

SAEB 1995 e 1997	Posicionamento dos itens na escala/ critérios de seleção: • itens com 65% ou mais de acerto no intervalo sob análise; • percentual de acerto no intervalo anterior ao intervalo sob análise deve ser abaixo de 50%; • diferença do percentual de acerto entre esses dois intervalos deve ser de pelo menos 30 pontos percentuais.

Considerando-se os critérios estabelecidos pelo SAEB 95 e 97 e as estatísticas produzidas para a análise dos itens e dos testes de

Proficiência em Matemática do SAEB 2003, foi elaborado o quadro II, integrante do Anexo VIII, em que se destacam os itens característicos dos níveis de proficiência correspondentes.

Identificados os itens característicos de cada nível, passa-se à construção da Escala de Proficiência em Matemática, tendo como referência os critérios de seleção utilizados pelo SAEB em 95 e 97, que se encontra no Anexo II.

Pode-se observar que os critérios usados para a seleção dos itens, tendo em vista a interpretação da Escala de Proficiência em Matemática, do Anexo II, são muito restritivos e rigorosos, e apenas itens com discriminação bastante alta preenchem os critérios estabelecidos. Por outro lado, itens que satisfazem a esses critérios são bastante característicos do nível que descrevem, pois tais critérios levam a percentuais expressivamente menores de acertos do item no nível imediatamente abaixo. Conseqüentemente, o número de itens aproveitados é muito baixo, como mostra a distribuição do número de itens selecionados obtidos por nível de proficiência apresentada na Tabela 2.

Tabela 2 – Número de itens selecionados por nível de proficiência, segundo os critérios do SAEB 1995-1997

Nível de Proficiência	N. de Itens selecionados	% Total	% Válido	% Vál. Acum.
Até 125	0	0	0	0
125 – 150	2	1,2	5,2	5,2
150 – 175	1	0,6	2, 6	7,8
175 – 200	0	0	0	7,8
200 – 250	10	6	25,6	33,4
250 – 300	19	11,3	48,7	82,1
300 – 350	6	3,6	15,4	97,5
350 – 375	1	0,6	2,6	100
Total Válido	39	23,2	100	
Itens não-selecionados	129	76,8		
Total	168	100		

Essa distribuição evidencia que os critérios de seleção do SAEB (1995-1997) são bastante rigorosos, e isso se traduz na observação de que apenas 39 dos 168 itens puderam ser selecionados, o que corresponde a aproximadamente 23% do total. Além disso, o número de itens representativos dos diversos níveis de proficiência não se distribui uniformemente ao longo do espectro de habilidade: os itens selecionados concentram-se nos níveis em torno de 250 pontos da escala arbitrária; observa-se que, dos 39 itens que foram selecionados, 19 situam-se no nível 250-300, e outros 10 itens, no nível inferior (200-250). Juntos, portanto, esses dois níveis centrais de proficiência abarcam mais de 70% dos itens selecionados segundo esse critério. Além disso, percebe-se certa assimetria na quantidade de itens selecionados ao longo dos níveis de proficiência, pela qual essa concentração é menor nos níveis inferiores. Com efeito, verifica-se que, para proficiências abaixo de 200, foram selecionados somente três itens, sendo que o nível 175-200 não teve nenhum item que satisfizesse aos critérios de seleção.

Vale mencionar, entretanto, que essa concentração de itens selecionados em torno da média de proficiência não é fenômeno exclusivo dos critérios do SAEB 1995-1997, pois também foi observado em grau significativo na construção da escala do SAEB 1999-2007, bem com em outras escalas consideradas a seguir, como as do NAEP, TIMSS e PISA.

2.3.2
Análise comparativa entre os critérios de seleção utilizados pelo SAEB a partir de 1999 e os utilizados em 1995/1997

Como a exposição dos processos de seleção de itens, segundo esses dois critérios, deixa claro, um dos objetivos da adoção dos novos critérios de seleção pelo SAEB 1999-2007 foi a possibilidade de inclusão de um maior número de itens selecionados, bem como a tentativa de fazer com que os itens selecionados cobrissem de modo mais completo e uniforme todas os níveis de proficiência considerados.

No SAEB 1999-2007, foram propostas duas maneiras de relaxar algumas das exigências impostas anteriormente, com o intuito de aumentar-se a quantidade de itens satisfazendo aos critérios de seleção.

A primeira proposta envolveu o relaxamento da segunda condição, segundo a qual os examinados tinham que apresentar um percentual de acerto inferior a 50% no nível imediatamente anterior ao de seleção. Dessa forma, passou-se então a considerar itens com um percentual de acerto igual ou maior que 50% nesse nível inferior (ao mesmo tempo em que a primeira e terceira condições foram mantidas inalteradas).

Percebeu-se que essa mudança possibilitou a seleção de 22 itens, 21 dos quais eram itens novos em relação aos obtidos anteriormente. Apenas um item (16905) satisfez aos dois critérios de seleção[3]. Isso permitiu elevar o número de itens selecionados de 37 para 58, o que fez subir para 34% a taxa de seleção.

A segunda proposta de mudança referiu-se à eliminação da exigência de que deveria haver uma diferença de no mínimo 30% de acerto entre o nível escolhido para a seleção e o nível que lhe ficava imediatamente abaixo (mantendo-se a primeira e a segunda condições inalteradas). Isso resultou na inclusão de outros 25 itens, fazendo o total subir para 82 itens selecionados, ou 49% dos 168 itens do banco. Nenhum dos itens selecionados assim obtidos já havia sido determinado pelos dois primeiros critérios, visto que, naturalmente, esse terceiro critério exclui os dois primeiros.

Observa-se também que essas duas formas de mudança de critérios amenizaram o problema já mencionado da concentração de itens selecionados nos níveis centrais de proficiência. A primeira mudança fez surgirem 8 itens selecionados no nível de 300-350[4] (contra 6 itens selecionados que haviam sido determinados nesse mesmo nível segundo os critérios originais). E ele também possibilitou a inclusão de 5 itens

[3] Esse item, entretanto, foi selecionado para diferentes níveis segundo esses dois critérios: conforme o primeiro critério, sua seleção se deu no nível 250-300; conforme o segundo critério, a seleção foi num nível imediatamente superior ao primeiro: 300-350. Ao reunirmos, posteriormente, todos os itens que foram selecionados segundo os critérios do SAEB 1995-1997 e seus dois modos de relaxamento, optou-se, aqui, por destinar esse item ao nível de 250-300, devido ao fato de essa seleção estar de acordo com as exigências originais.

[4] Levando-se em conta o item 16905, mencionado na nota anterior.

selecionados no nível máximo de 350-375 (contra apenas 2 que haviam sido originalmente determinados nesse mesmo nível segundo os critérios anteriores). Isso se verificou porque o primeiro dos critérios anteriores favorecia a seleção de itens mais difíceis, ao exigir que, nos níveis imediatamente anteriores aos da seleção, o percentual de acerto fosse inferior a 50%. Isso produziu um deslocamento do espectro de habilidades na direção da extremidade superior. Com efeito, observa-se que, dos 22 itens selecionados conforme o primeiro relaxamento, nada menos que 21 deles corresponderam ao nível 250-300 ou mais (e apenas um item correspondeu ao nível 200-250).

Por outro lado, a segunda mudança permitiu um relativo aumento do número de itens selecionados na direção oposta, ou seja, na região dos menores níveis de habilidade. Isso foi conseguido não só porque se manteve, nesse caso, a exigência de que o percentual de acerto dos examinados situados no nível anterior ao nível de seleção fosse inferior a 50%, mas também porque não se exigia mais uma diferença tão grande (30%) de acerto entre os dois níveis adjacentes.

Com efeito, observou-se, nesse caso, que 18 dos 25 itens assim obtidos situavam-se no nível 200-250 ou antes. Mesmo nessa direção inferior, entretanto, o número de itens selecionados foi paulatinamente diminuindo, à medida que se caminhava para a extremidade inferior do espectro, de modo que o menor nível em que surgiram itens selecionados, conforme esse segundo relaxamento, foi de 150-175, onde surgiram 2 itens (contra apenas um que havia sido determinado segundo os critérios originais). No nível mínimo de habilidade (125-150), porém, esse segundo relaxamento não foi capaz de produzir novos itens selecionados.

Em suma, a adoção simultânea dos três conjuntos de critérios do SAEB (o conjunto original e seus dois relaxamentos) resultou em 84 itens selecionados, de um total de 168 itens disponíveis. Dessa forma, foi possível obter itens selecionados em todos os níveis de proficiência existentes (com exceção do nível inferior, de até 125 pontos, ao qual naturalmente não se atribuem itens característicos, por não ser possível

comparar o desempenho dos alunos de proficiência situado nesse nível mínimo com o desempenho de alunos com proficiência hipoteticamente ainda inferior). Algumas características marcantes desse novo conjunto de critérios (que já eram típicas da seleção baseada no conjunto original e que, de certo modo, foram atenuadas pelos dois relaxamentos) são:

a) A concentração de itens selecionados nos níveis centrais de proficiência: 63% deles situam-se nos níveis 200-250 e 250-300.
b) A assimetria à esquerda da distribuição: 27% deles concentram-se nos dois níveis superiores (300-350 e 350-375), enquanto que 10% se espalham pelos três níveis inferiores (125-150, 150-175 e 175-200).

2.4
O NAEP: objetivo e ciclo de avaliação

O objetivo do *National Assessment for Educational Progress* – NAEP – é produzir informações sobre o que alunos de escolas americanas, públicas e privadas, conhecem e podem fazer com o que sabem nas diversas áreas do conhecimento: leitura, matemática, ciências, história e civismo.

Os testes são produzidos por um grande comitê de especialistas, educadores e cidadãos diretamente interessados, que especificam os objetivos da avaliação segundo o que os alunos devem aprender durante sua educação escolar. Os objetivos por área são definidos em termos de uma matriz de conteúdos para o processo de avaliação. Para satisfazer aos objetivos da avaliação e assegurar que os tópicos selecionados para medir cada objetivo cubram uma extensão de níveis de dificuldade, as amostras do NAEP se fazem com um grande número de itens dentro de cada área a ser avaliada. Originalmente, o NAEP produzia seus resultados apresentando os percentuais de resposta correta para itens individuais. Posteriormente, foram apresentadas porcentagens médias de grupos de itens. Em 1983, quando o *Educational Testing Service* – ETS – tornou-se o gestor do NAEP, introduziram-se as escalas de proficiência, com média 250 e desvio-padrão 50, baseadas na Teoria da Resposta ao Item – TRI.

2.4.1
A ancoragem de itens[5] utilizada pelo NAEP

A ancoragem é um meio utilizado pelo NAEP para a construção da escala de proficiência, que se realiza por meio de um procedimento simples e direto, baseado apenas no aumento da proporção de alunos que respondem corretamente ao item entre pontos-âncora adjacentes. Esse processo permite verificar o que a maioria dos alunos em um ponto-âncora sabe e pode fazer, e o que a maioria dos alunos situados nos níveis imediatamente mais baixos não sabe e não pode fazer. O NAEP utiliza dois métodos para decidir que itens serão selecionados para a descrição dos pontos-âncora.

O primeiro método é denominado método direto, porque o que os alunos sabem e conseguem fazer relaciona-se diretamente a um ponto-âncora. Como poucos alunos têm (ou nenhum aluno tem) proficiência exatamente no ponto-âncora na escala do NAEP, as proficiências são agrupadas em um pequeno intervalo acima e abaixo do ponto-âncora, de modo a haver pelo menos 100 alunos no intervalo representativo do ponto em questão. Esse método se fundamenta em dois critérios:

i) 80% dos alunos em um ponto-âncora respondem corretamente ao item, e, no ponto imediatamente inferior, menos de 50% dão a resposta correta ao item.

ii) 65% dos alunos no ponto mais alto respondem corretamente ao item, e menos de 50% dos alunos, no nível imediatamente mais baixo, são capazes de responder corretamente ao item, sendo que a diferença de acerto entre esses níveis é de, no mínimo, de 30 pontos percentuais.

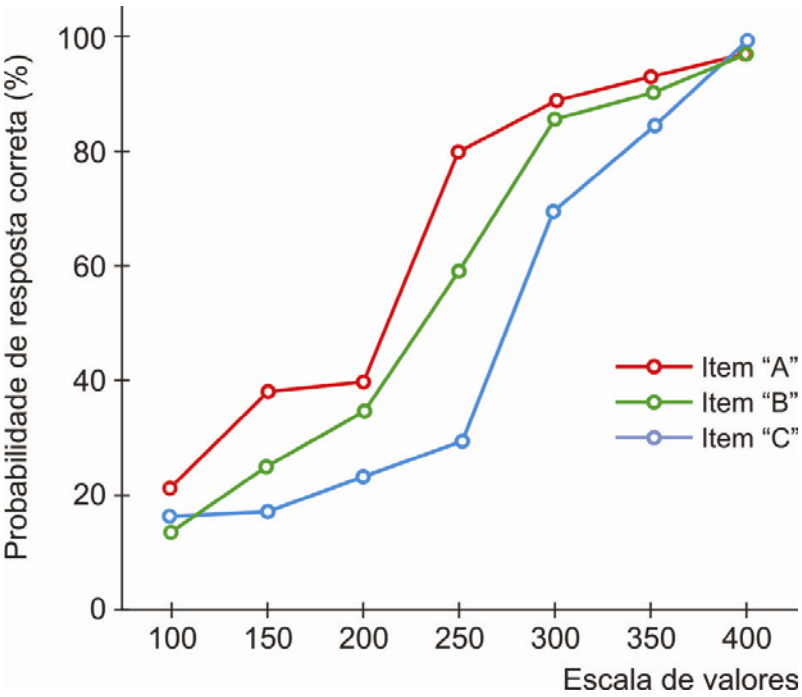
Para os escores que estão no ponto-âncora mínimo (o primeiro da escala), ou próximo dele, o critério para a seleção do item é: (i) 80% ou (ii) 65% de alunos respondendo corretamente o item. O primeiro critério é menos aplicável e pode não produzir itens

[5] No NAEP, o processo de seleção de itens para níveis específicos de proficiência é explicitamente chamado de ancoragem (lit. *anchoring*), razão pela qual se optou por manter essa expressão para se referir à seleção dos itens significativos dos níveis de proficiência nesse programa.

suficientes para a interpretação; por isso, o segundo critério é o mais utilizado.

O segundo método é chamado de método refinado, porque se enquadra no modelo logístico de três parâmetros para a escalonagem de escores usa as Curvas Características dos Itens – CCIs – para aproximar, nos pontos de ancoragem selecionados, a porcentagem de acerto em cada item. Os parâmetros do item tornam-se facilmente disponíveis, se a escala tiver sido produzida segundo os métodos da TRI. Esse método utiliza o mesmo critério de seleção de itens, exceto pelo fato de que os tamanhos da amostra nos pontos-âncora não estão disponíveis, e, portanto, o tamanho mínimo da amostra não é utilizado. Ambos os métodos produzem resultados semelhantes. Beaton e Allen (1992) apresentam uma representação gráfica desse processo, conforme se vê a seguir.

Figura 5



Nessa representação, os itens "A", "B", "C" são ilustrativos do processo. Seis possíveis pontos-âncora são selecionados, correspondendo aos valores 100, 150, 200, 250, 300, 350 na escala. O item "A" é selecionado para o nível 250 segundo qualquer dos dois critérios de

seleção, visto que a probabilidade de acerto para alunos com proficiência em torno de 250 é de 80%, enquanto que, para os alunos do nível imediatamente mais baixo (200), a probabilidade é de 40%. O item "C" é selecionado para o nível 300, usando-se o segundo critério, porque as probabilidades de sucesso nos dois níveis satisfazem aos valores mínimos dos pontos iniciais. O item "B" não é selecionado em nenhum nível, pois a discriminação entre níveis adjacentes não é grande o suficiente segundo as exigências consideradas (o percentual de acerto chega aos 80% no ponto 300 da escala de proficiência, mas o percentual de acerto no ponto de seleção inferior – 250 – supera os 50%).

Usando-se um ou outro método, os itens que compõem os testes são agrupados, de modo que um conjunto de itens possa discriminar entre diferentes pontos-âncora. Os itens que não discriminam são agrupados separadamente e passam por uma revisão. Um comitê de especialistas por área, educadores e outros interessados, estuda e interpreta os itens selecionados. Também compete a esse comitê generalizar, a partir de itens específicos que discriminam entre níveis-âncora, as habilidades desenvolvidas pelos alunos nesses níveis. Assim sendo, o processo de ancoragem é um meio de estudar-se, empiricamente, o desenvolvimento das habilidades dos alunos avaliados pelo NAEP, identificando-se itens que discriminam entre os pontos-âncora, ou seja, itens que permitem que os pontos de seleção descrevam o que a grande maioria de alunos em um nível pode fazer a mais em relação à maioria dos alunos do nível imediatamente inferior. O Quadro 3 apresenta de forma sucinta os critérios de ancoragem.

Quadro 3

Programa de Avaliação	Critérios/Posicionamento dos itens na Escala de Proficiência
NAEP	• Os pontos-âncora são arbitrários, correspondendo a intervalos relativamente estreitos e dispostos intercaladamente ao longo de um *continuum*. • Posicionamento dos itens na escala/Critérios de seleção: Primeiro Critério: Mais de 80% dos alunos em um ponto-âncora respondem corretamente ao item e, no nível imediatamente inferior, menos que 50%, dão a resposta correta ao item. Segundo Critério: Itens selecionados: itens com 65% ou mais de acerto no nível-âncora e com menos de 50% no nível imediatamente inferior, sendo que a diferença entre os percentuais de alunos posicionados nesses níveis seja de, no mínimo, 30 pontos percentuais.

Em ambos os métodos de seleção do NAEP, vários pontos da escala, chamados de pontos de seleção ou pontos-âncora, são relacionados. Como não há, ao longo da escala NAEP, pontos observáveis, nos quais o desempenho dos alunos cresça dramaticamente, a seleção dos pontos é arbitrária. A escala NAEP trabalha com números redondos. Os pontos-âncora devem estar suficientemente distantes para retratar desempenhos diferentes, mas não tão distantes a ponto de se tornarem triviais. Na avaliação de matemática do NAEP, os pontos-âncora foram 200, 250, 300 e 350, para toda a escala.

Deve-se observar, entretanto, que, no NAEP, a escala completa envolvendo a 4ª, a 8ª e a 12ª séries estende-se, aproximadamente, de 0 a 500 pontos, sendo que a maioria dos alunos dessas séries têm seus escores variando de 100 a 400 pontos. Portanto, os quatro pontos de seleção acima mencionados referem-se ao intervalo total de proficiência, de modo que, quando se restringe a análise para uma série específica (como a 4ª série, no presente caso), nem todos esses pontos-âncora devem ser considerados, visto não haver, para essa série específica, variações relevantes de percentuais de alunos entre alguns desses pontos.

Por exemplo, como a 4ª série diz respeito a séries iniciais, é a menos avançada das três séries consideradas, observa-se que dois pontos de seleção de interesse são os valores de 200 e de 250 da escala, que, na avaliação de 1990, corresponderam, para a 4ª série, aproximadamente, aos percentis 30 e 90 respectivamente. Dito de outro modo, nessa avaliação, cerca de 30% dos alunos obtiveram um resultado de até 200 pontos, e cerca de 90% deles obtiveram um resultado de até 250 pontos. A escolha desses dois pontos para se fazer a seleção dos itens-âncora parece ser adequada, visto que, de fato, percebe-se uma diferença significativa de proficiência entre esses pontos: no primeiro deles, encontram-se estudantes de proficiência baixa, enquanto que, no segundo, estão os estudantes de proficiência elevada.

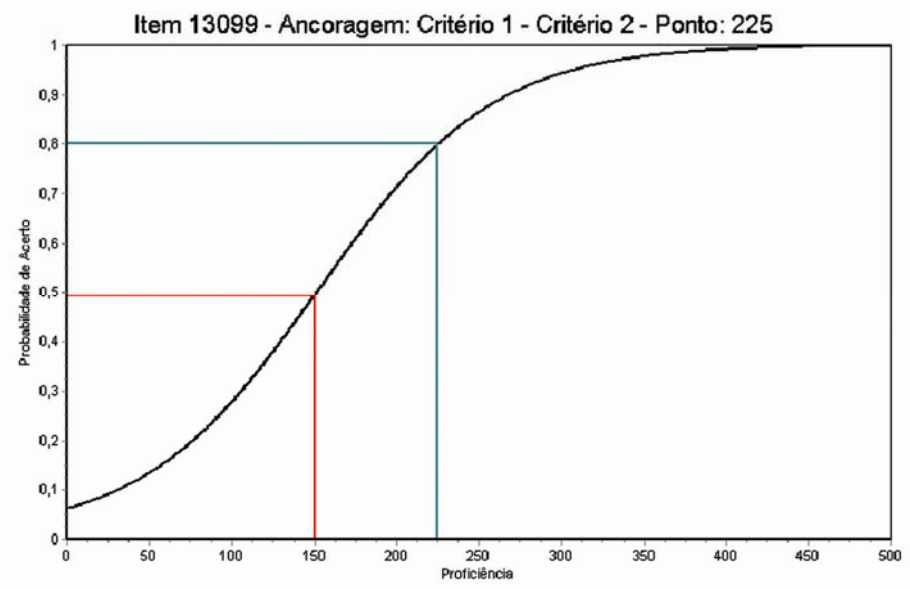
Se se incluir, entretanto, por exemplo, o ponto de seleção correspondente ao valor de 300 na escala, pode-se perceber que o salto de proficiência será muito pouco significativo, visto que esse salto ocorrerá entre alunos de percentis aproximadamente correspondentes a 90 e a 100 na escala de proficiência (uma vez que quase todos os alunos da 4ª série têm proficiência igual ou inferior a 300 nessa escala). E a inclusão do ponto de 350 seria ainda menos pertinente, visto não haver praticamente diferença entre os níveis de 300 e de 350 quanto aos percentis de proficiência (sendo ambos aproximadamente iguais a 100).

Portanto, para os itens considerados no presente estudo, referentes à 4ª série do Ensino Fundamental, foram utilizados somente dois pontos de ancoragem, de modo a haver uma correspondência adequada com os procedimentos empregados no NAEP. Um outro aspecto relevante nessa etapa foi a escolha dos valores da escala do SAEB, que seriam utilizados como pontos-âncora. Isso teve que ser considerado porque, naturalmente, a escala do SAEB não é a mesma escala que a do NAEP, de modo que uma tarefa importante foi decidir quais pontos da escala do SAEB poderiam ser considerados como correspondentes aos pontos de 200 e 250 da escala do NAEP.

A solução para isso foi obtida escolhendo-se os dois pontos-âncora da escala do SAEB como sendo aqueles correspondentes ao trigésimo e ao nonagésimo percentil da proficiência dos examinados no SAEB, visto que esses mesmos percentis eram os que, na escala do NAEP de 1990, localizavam-se nos pontos de 200 e de 250 de sua escala de proficiência. Na escala do SAEB, esses pontos corresponderam, aproximadamente, às proficiências de 150 e de 225 respectivamente, de modo que esses foram os dois pontos-âncora escolhidos.

O gráfico a seguir ilustra o processo de ancoragem de um dos itens do banco conforme o primeiro critério (e também conforme o segundo, visto que, se um item satisfaz ao primeiro critério de ancoragem, ele, na verdade, satisfaz a ambos). Observa-se, então, que o item em questão – de número 13099 – ancora no nível de proficiência de 225, e cerca de 80% dos respondentes nesse nível acertam o item, ao passo que o percentual de acerto entre os respondentes situados no nível anterior (com proficiência de 150) corresponde a aproximadamente 50%.

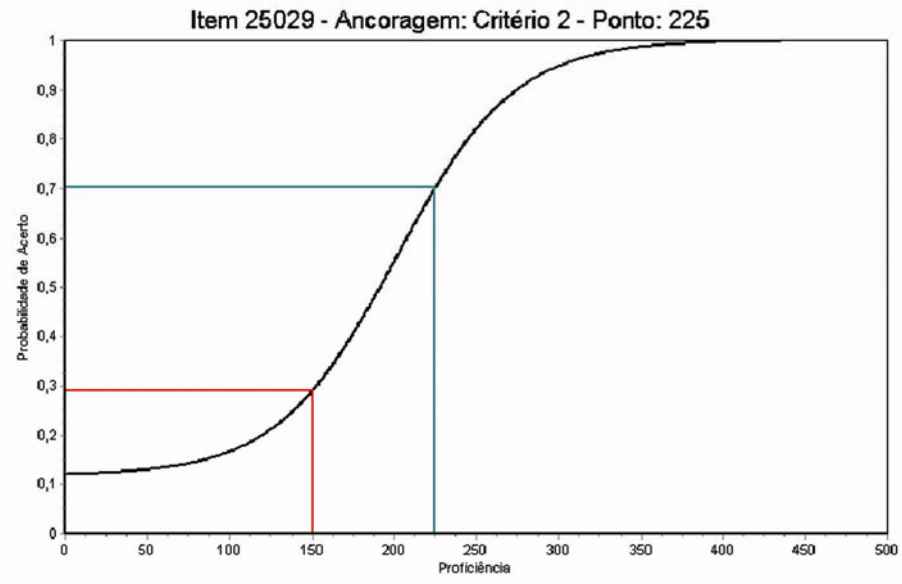
Figura 6 – Item 13099



Já a observação do segundo critério de ancoragem é ilustrada pelo gráfico a seguir, onde se mostra um item – 25099 – também ancorando no nível 225 apenas pelo segundo critério, de modo que, nesse nível, o

percentual de acerto é de cerca 70% (superior, portanto, ao valor mínimo para esse nível, que é de 65%), ao mesmo tempo em que o percentual de acerto no nível anterior é de 30% (inferior, portanto, ao valor máximo para esse nível, que é de 50%), e a diferença entre os percentuais de acerto desses dois níveis é de aproximadamente 40% (superior, portanto, à diferença mínima entre os dois níveis, que é de 30%).

Figura 7 – Item 25099



Considerando-se os critérios estabelecidos pelo NAEP e as estatísticas produzidas para a análise dos itens e dos testes de Matemática do SAEB 2003 da 4ª série do Ensino Fundamental, produziu-se, então, o Quadro III, integrante do Anexo VIII, em que se destacam os itens selecionados característicos dos níveis de proficiência correspondentes e, no Anexo IV, as curvas características dos respectivos itens da Escala de Proficiência constante do Anexo III.

A Tabela 3 apresenta a distribuição do número de itens-âncora por nível de proficiência, segundo o primeiro critério adotado pelo NAEP:

Tabela 3 – Número de itens-âncora por nível de proficiência conforme o primeiro critério do NAEP

Nível	N. de Itens-âncora	% Total	% Válido
1	3	1,8	7,7
2	36	21,3	92,3
Total Válido	39	23,1	100
Itens não-selecionados	130	76,9	
Total	169	100	

Nessa distribuição, constata-se que, de acordo com o primeiro critério utilizado pelo NAEP, foram selecionados 39 itens-âncora, uma quantidade que perfaz apenas 23,1% do total de itens. Além de ser um número consideravelmente pequeno, esse conjunto também se distribui de maneira bastante desigual entre os níveis-âncora: verifica-se que apenas 3 itens ancoraram no primeiro nível, o que correspondeu a somente 7,7% dos itens que ancoraram. Essa falta de itens no primeiro nível deveu-se à exigência de que, nesse nível, o percentual mínimo de acerto ao item deveria ser de 80%, o que fez com que somente itens relativamente muito fáceis pudessem ser selecionados.

Naturalmente, a assimetria observada quanto ao número de itens selecionados por nível é indesejável, visto que seria interessante que houvesse um número significativo de itens distribuídos uniformemente em ambos os espectros de proficiência.

Com a adoção do segundo critério, menos rigoroso que o primeiro, há um natural acréscimo do número de itens selecionados. Observa-se que parte dos itens definidos conforme o primeiro critério corresponde a um subconjunto dos encontrados conforme o segundo critério, ou seja, todos os itens que ancoraram conforme o segundo critério também o fizeram conforme o primeiro critério, embora o oposto não necessariamente tenha ocorrido.

A Tabela 4 expõe a distribuição dos itens selecionados por níveis de proficiência, de acordo com o segundo critério adotado pelo NAEP.

Tabela 4 – Número de itens-âncora por nível de proficiência conforme o segundo critério do NAEP

Nível de seleção	N. de Itens selecionados	% Total	% Válido
1	14	8,3	16,9
2	69	40,8	83,1
Total Válido	83	49,1	100
Itens não-selecionados	86	50,9	
Total	169	100	

Nessa distribuição, pode-se observar que, com a adoção desse segundo critério, o número de itens selecionados aumenta para 69; ainda assim, o percentual de itens selecionados em relação ao total de itens continua baixo (40,8%). Também permanece o problema de poucos itens no primeiro intervalo considerado, com apenas 16,9% dos itens ancorados nesse nível.

Outro aspecto relevante sobre a ancoragem dos itens no NAEP é a constatação de que ela se verifica quase que exclusivamente para os itens com índice de dificuldade média ou baixa conforme se percebe no Quadro III do Anexo VIII. Com efeito, pelo quadro, percebe-se que, por exemplo, dos 63 itens mais difíceis do banco (assim medidos pelos seus respectivos coeficientes b, que estão apresentados em ordem crescente na tabela), nenhum deles ancora. A razão de isso ter ocorrido se deve à exigência de que, no nível de ancoragem, o percentual de acerto deveria ser de pelo menos 80% no primeiro critério, ou 65% no segundo. Como, nos itens mais difíceis, esses percentuais de acerto não foram atingidos no pontos-âncora considerados, esses itens não puderam ser selecionados.

2.4.2
Análise comparativa entre os critérios de seleção utilizados pelo SAEB a partir de 99 e o NAEP

Comparando-se os resultados da seleção realizada segundo os critérios do SAEB 1999-2007 com a seleção obtida segundo os critérios do NAEP, observa-se uma série de diferenças significativas entre os dois métodos.

Uma diferença facilmente perceptível refere-se ao número de itens selecionados: conforme os critérios do SAEB 1999-2007, todos os 168 itens foram selecionados, ao passo que, no NAEP, somente 39 itens satisfizeram ao primeiro critério, e 83 itens satisfizeram ao segundo. Essa disparidade é parcialmente explicada pela diferença de rigor entre os dois critérios, que foi muito maior no caso do NAEP que no do SAEB 1999-2007: enquanto que, no SAEB, para ser selecionado, um item devia apenas satisfazer o critério de ter um percentual igual ou superior a 65% em um nível, e menos do que isso no nível imediatamente anterior, o NAEP, por sua vez, especificava a necessidade de haver diferenças maiores de acerto entre os níveis, de modo que apenas os itens que experimentaram um mínimo de 30% de diferença entre eles foram selecionados conforme os critérios do NAEP.

Para o SAEB, os níveis correspondem aos intervalos compreendidos entre os pontos-limite, abrangendo todos os estudantes com proficiência situada entre esses extremos. Portanto, para o SAEB, o conjunto formado pelos níveis corresponde a um intervalo contínuo de proficiências.

O NAEP, por outro lado, considera apenas subamostras de estudantes com proficiência situada em estreitos níveis em torno dos pontos de seleção[6]. Logo, no NAEP, os intervalos de seleção são descontínuos.

Essa restrição dos níveis de seleção a intervalos estreitos do espectro de proficiência é outro fator que, efetivamente, limitou a

[6] Como se viu, a seleção dos itens do NAEP somente se deu em quatro níveis relativamente estreitos do espectro de proficiência (correspondendo aos pontos de seleção de 150, 200, 250 e 300, ocupando centros de níveis de 25 pontos de largura).

ocorrência de itens selecionados segundo esse sistema. Isso nos leva à constatação de que o NAEP concentrou-se em analisar diferenças de acerto entre dois níveis previamente definidos para a quarta série do Ensino Fundamental.

2.5
O TIMSS: objetivo e ciclo de avaliação

A *International Association for the Evaluation of Educational Achievement* – IEA – é uma entidade pioneira em processos internacionais de avaliação educacional. Seus primeiros trabalhos foram realizados em 1964, quando se concluiu o primeiro estudo internacional em matemática (*First International Mathematics Study* – FIMS). Desde então, já foram realizadas, aproximadamente, 20 avaliações em linguagem, civismo, leitura, matemática e ciências. Em relação a essas duas últimas disciplinas, a IEA implementou o TIMSS (*Trends in International Mathematics and Science Study*), que propicia a oportunidade de aferir, comparar e explicar as tendências em matemática e ciências para a 4ª e a 8ª séries nos países envolvidos no processo avaliativo. Desde a avaliação de 1995, vem-se utilizando a TRI nessas avaliações, cujo ciclo é quadrienal, e cujas mais recentes edições foram:

1995 – *Third International Mathematics and Science Study*, a maior e mais complexa avaliação realizada pela IEA, em ciências e matemática, com alunos das 3ª, 4ª, 7ª e 8ª séries do Ensino Fundamental e 3ª série do Ensino Médio.

1999 – TIMSS – renomeado para *Trends in International Mathematics and Science Study*, avaliou apenas os alunos da 8ª série nas áreas de matemática e ciências, sendo esses os mesmos da 4ª série da avaliação de 1995.

2003 – Esta versão do TIMSS avaliou os alunos de 4ª série em 26 países, e da 8ª série em 46 países[7], em matemática e ciências, utilizando questões de múltipla escolha e questões abertas em três domínios cognitivos: conhecer, aplicar e inferir.

2007 – Versão mais recente do TIMSS, avaliou também as mesmas séries da edição anterior, estendendo-se para mais de 60 países, muitos dos quais já haviam sido avaliados em 2003. Vários de seus resultados ainda não vieram a público, estando com sua divulgação prevista para o final de 2008 e o início de 2009.

Além dos testes de matemática e ciências, são aplicados questionários aos diretores e professores das escolas, buscando produzir informações extensivas sobre o ensino e sobre a aprendizagem nas áreas do conhecimento avaliadas. É importante ressaltar, ainda, que o TIMSS concluiu recentemente uma inovadora análise de "oportunidades de aprendizado" dos alunos, que categoriza e compara os currículos de matemática e ciências, os livros didáticos e os materiais pedagógicos utilizados entre os países participantes. Portanto, os resultados, ao invés de se limitarem à identificação dos países com desempenho mais alto ou mais baixo, são também utilizados como um instrumento capaz de aferir o progresso educacional de um país e redefinir suas metas curriculares e práticas pedagógicas.

2.5.1
A seleção de itens representativos utilizada pelo TIMSS[8]

Os processos de equalização conduzidos pelo TIMSS, com respeito às avaliações dos anos 1995, 1999 e 2003, resultaram na construção de escalas de proficiência independentes para ciências e matemática. Para

[7] No total, os países participantes da edição de 2003 foram: África do Sul, Arábia Saudita, Argentina, Armênia, Austrália, Autoridade Nacional Palestina, Bahrein, Bélgica, Botswana, Bulgária, Canadá, Chile, Chipre, Cingapura, Coréia do Sul, Egito, Escócia, Eslováquia, Eslovênia, Espanha, Estados Unidos, Estônia, Filipinas, Gana, Holanda, Hong Kong, Hungria, Iêmen, Indonésia, Inglaterra, Irã, Israel, Itália, Japão, Jordânia, Letônia, Líbano, Lituânia, Macedônia, Malásia, Marrocos, Moldávia, Noruega, Nova Zelândia, Romênia, Rússia, Sérvia, Síria, Suécia, Taiwan e Tunísia. Em alguns países, avaliou-se apenas uma das séries consideradas no estudo; em outros países, ambas as séries foram avaliadas.

[8] O Relatório Técnico do TIMSS 2003 foi a referência utilizada para o estudo desse tópico.

cada uma dessas disciplinas, não houve equalização entre a 4ª e a 8ª séries, de modo que ambas adotaram escalas independentes e numericamente iguais, tendo como referência a avaliação de 1995, com média 500 e desvio padrão 100.

Uma vez estabelecida a escala de proficiência, o TIMSS utiliza, para a descrição das habilidades dos alunos, o processo de seleção em cinco intervalos, especificados a partir de quatro pontos dessa escala. Esses pontos, tanto para a 4ª, quanto para a 8ª série, foram definidos nas avaliações de 1995 e 1999, por meio dos percentis 25, 50, 75 e 90, não sendo considerado nenhum critério pedagógico explícito.

Em 2003, porém, adotou-se outro critério, com o objetivo de definir pontos fixos que se tornassem uma referência para as avaliações seguintes, pois os pontos de percentis estavam mudando a cada ano de aplicação, devido, principalmente, a dois fatores: (i) novos países participantes; e (ii) variações na proficiência dos países participantes das avaliações anteriores.

Os novos pontos foram selecionados de modo a não ficar muito distantes dos pontos de 1999 para a 8ª série, mantendo-se distâncias iguais entre os intervalos. Os pontos selecionados são apresentados na Tabela 5.

Tabela 5 – Pontos selecionados na escala TIMSS

Percentil	Proficiência	Referência internacional
90	625	Desempenho Avançado
75	550	Desempenho Alto
50	475	Desempenho Intermediário
25	400	Desempenho Baixo

Para a definição dos níveis de proficiência, foram considerados intervalos de 10 pontos (5 pontos abaixo e 5 pontos acima) em torno de cada ponto de referência, conforme apresentado na Tabela 6.

Tabela 6 – Níveis selecionados na escala TIMSS

Referência internacional	Níveis de proficiência
Desempenho Avançado	620-630
Desempenho Alto	545-555
Desempenho Intermediário	470-480
Desempenho Baixo	395-405

Após calcularem o percentual de acerto para cada item em cada um dos níveis de proficiência, os especialistas dedicaram-se à definição de itens selecionados, utilizando três critérios. O primeiro se refere à escolha de itens com ancoragem perfeita, ou seja, (i) para itens de múltipla escolha, o critério utilizado foi de 65% de acerto no nível de ancoragem e 50% no nível anterior. Visando-se a não descartar itens no processo de ancoragem, mais dois critérios foram utilizados para a incorporação dos itens de múltipla escolha não selecionados anteriormente, quais sejam: (ii) itens que quase ancoraram – para itens de múltipla escolha: pelo menos 60% no nível de seleção e menos de 50% no nível anterior; e (iii) itens difíceis de serem ancorados – para itens de múltipla escolha: pelo menos 60% ou 65% no nível de ancoragem, independente do desempenho dos estudantes no nível anterior. Para itens abertos, como não há possibilidade de acerto ao acaso, o critério foi de 50% no nível de ancoragem, não se considerando o percentual no nível anterior. Após a identificação dos itens em seus pontos de ancoragem, os especialistas passaram a construir a escala de proficiência, objetivando diagnosticar as habilidades desenvolvidas pelos alunos em cada ponto de seleção.

Aplicando-se os critérios de seleção do TIMSS à construção da escala de proficiência em Matemática do SAEB 2003 para a 4ª série do Ensino Fundamental, serão levados em consideração os critérios de seleção de itens representativos abordados anteriormente e apresentados de forma sucinta no Quadro 4.

Quadro 4

Programa de Avaliação	Posicionamento dos itens na escala de proficiência/critérios de seleção
TIMSS	A Escala de Proficiência foi arbitrada como tendo uma média de 500 e um desvio padrão de 100. Os níveis de proficiência foram definidos a partir de pontos fixos, tendo como referência os percentis 25, 50, 75 e 90 e intervalos de 10 pontos (5 pontos abaixo e 5 pontos acima) em torno de cada ponto de referência. Posicionamento dos itens na escala/Critérios de seleção: (i) Seleção perfeita; pelo menos 65% de acerto no nível de seleção e 50% no nível anterior. (ii) Itens que quase ancoraram: pelo menos 60% no nível de seleção e menos de 50% no nível anterior. (iii) Itens difíceis de serem ancorados: pelo menos 65% no nível de seleção independentemente do percentual no nível anterior.

A fim de se utilizarem os critérios do TIMSS na determinação dos níveis de proficiência em Matemática para os itens considerados no presente estudo, foram adotados intervalos de 5 pontos, correspondentes a 10% do desvio padrão (2,5 pontos abaixo e 2,5 pontos acima) em torno de cada um dos percentis: 25, 50, 75 e 90. Esse critério foi escolhido em função de a escala SAEB possuir uma média de 250 pontos e um desvio padrão de 50 pontos.

Tabela 7 – Correspondência entre os níveis de proficiência do TIMSS e do SAEB nos níveis de seleção definidos pelos critérios do TIMSS

Referência Internacional	Níveis de Proficiência do TIMSS	Níveis de Proficiência Correspondentes na Escala do SAEB
Desempenho Avançado	620-630	235,1 a 240,1
Desempenho Alto	545-555	203,6 a 209,1
Desempenho Intermediário	470-480	170,7 a 175,7
Desempenho Baixo	395-405	141,7 a 146,7

O Quadro IV, integrante do Anexo VIII, apresenta o demonstrativo do percentual de acerto em cada item, segundo os níveis de proficiência estabelecidos conforme os critérios adotados pelo TIMSS.

Identificados os itens selecionados de cada nível, passa-se a construir a Escala de Proficiência em Matemática do SAEB, tendo como referência os três critérios de seleção utilizados pelo TIMSS, que se encontram no Anexo IV.

2.5.2
Considerações sobre a seleção dos itens segundo os critérios do TIMSS

A seguir, a Tabela 8 apresenta o número de itens selecionados de acordo com cada um dos três critérios definidos pelo TIMSS.

Tabela 8 – Número de itens selecionados por nível de proficiência conforme os três critérios do TIMSS considerados separadamente

Níveis de Proficiência	Seleção Perfeita	2º Critério	3º Critério
Nível 1	6	11	5
Nível 2	7	14	7
Nível 3	22	27	9
Nível 4	15	17	8
Total	50	69	29

Observa-se que apenas 50 itens tiveram uma seleção perfeita, o que correspondeu a cerca de 30% do total. A adoção do segundo critério produziu 69 itens. Também se percebe, entretanto, que os itens determinados conforme o primeiro critério são um subconjunto dos itens determinados pelo segundo critério. Assim, a adoção do segundo critério representou um acréscimo de apenas 19 itens ao total de itens selecionados, visto que os outros 50 já haviam sido determinados pelo primeiro critério. Vale a pena observar, outrossim, que, nos 50 itens que são selecionados de acordo com ambos os critérios, a seleção se dá exatamente nos mesmos níveis.

Tomando como base todos os itens selecionados determinados independentemente desses três critérios, percebe-se que o total válido

(ou seja, a soma dos itens que foram selecionados) corresponde a 74 itens, ou cerca de 44% do total. Também se percebe a existência de itens selecionados em todos os quatro níveis de seleção, com maior concentração no nível 3, e uma distribuição aproximadamente simétrica em torno desse centro, como mostram os dados apresentados na Tabela 9.

Tabela 9 – Número de itens selecionados por nível de proficiência conforme os três critérios do TIMSS tomados conjuntamente

Níveis de Proficiência	N. de Itens selecionados	% Total	% Válido	% Vál. Acum.
Nível 1	11	6,5	14,9	14,9
Nível 2	14	8,3	18,9	33,8
Nível 3	28	16,6	37,8	71,6
Nível 4	21	12,4	28,4	100
Total Válido	74	43,8	100	
Itens não-selecionados	95	56,2		
Total	169	100		

2.5.3
Análise comparativa entre os critérios de seleção utilizados pelo SAEB a partir de 99 e o TIMSS

Uma semelhança entre os critérios de seleção utilizados pelo SAEB, a partir de 1999, e os utilizados pelo TIMSS é que, em ambos os critérios, o número de itens selecionados obtidos não se distribui uniformemente ao longo de todo o intervalo de proficiência considerado, concentrando-se em níveis mais centrais de proficiência, ao mesmo tempo em que esse número vai diminuindo em direção aos níveis extremos, tanto para as de maior quanto para as de menor proficiência. Nos dois casos, os saltos de percentual de acerto necessários à seleção nos dois critérios ocorrem mais freqüentemente nos níveis de proficiência próximos à média, ficando mais raros, à medida que se caminha em direção aos extremos.

Quanto às diferenças entre o TIMSS e o SAEB, uma óbvia distinção entre os dois diz respeito aos intervalos de proficiência. No SAEB, são selecionados oito intervalos que compõem todo o espectro de proficiência referente à 4ª série, enquanto que, no TIMSS, os níveis de seleção se resumem a quatro, de larguras restritas, de modo que apenas uma parte relativamente pequena do espectro de proficiência está sendo considerada.

Outra diferença significativa reside no número de itens selecionados, que é bem menor no TIMSS do que no SAEB, visto que, no primeiro, os critérios de seleção são mais restritos, fazendo com que menos da metade dos itens satisfaça aos critérios de seleção, ao passo que, no SAEB, todos os itens são selecionados.

2.6
O PISA: objetivo e ciclo de avaliação

Em uma perspectiva internacional comparativa, trinta e um países participam do Programa Internacional de Avaliação de Alunos, PISA, desenvolvido e coordenado pela Organização para Cooperação e Desenvolvimento Econômico – OCDE. São integrantes desse programa de avaliação em larga escala os seguintes países: Alemanha, Austrália, Áustria, Bélgica, Brasil, Canadá, China, Coréia, Dinamarca, Espanha, Estados Unidos, Finlândia, França, Grécia, Holanda, Hungria, Islândia, Irlanda, Itália, Japão, Letônia, Luxemburgo, México, Nova Zelândia, Noruega, Polônia, Reino Unido, República Tcheca, Rússia, Suécia e Suíça.

O PISA avalia uma amostra de jovens de 15 anos de idade, matriculados em escolas das zonas urbana e rural, das redes públicas e privada, e tem como objetivo produzir indicadores sobre a efetividade dos sistemas educacionais.

No PISA, a avaliação do desempenho escolar é realizada por meio de testes com ênfases distintas em três áreas do Conhecimento: Leitura, Matemática e Ciências. Na primeira avaliação, ocorrida em 2000, o foco foi a proficiência em Leitura, ficando Ciências e Matemática em segundo plano. Em 2003, a área principal foi a Matemática. Em 2006, a ênfase da

avaliação foi em Ciências e, em 2009, a ênfase será novamente em Leitura.

Além de testes cognitivos, o PISA aplica questionários contextuais para os alunos e o diretor de cada unidade escolar, destinados à produção de indicadores relacionados à caracterização socioeconômica, cultural e de hábitos de estudos dos alunos, bem como às características das escolas onde estudam. Assim, esses indicadores possibilitam a análise dos resultados dos testes de proficiência em função de indicadores que podem interferir no aprendizado do aluno, ou seja, permitem analisar fatores socioeconômicos e culturais que interferem de modo separado ou conjuntamente no desempenho escolar.

2.6.1
A seleção de itens representativos utilizada pelo PISA[9]

Em cada avaliação produzida, o PISA enfatizou uma área de conhecimento: em 2000, Leitura; em 2003, Matemática; e, em 2006, Ciências. Tal proposta de trabalho levou, conseqüentemente, à elaboração de escalas de proficiência para se avaliar o desempenho dos alunos envolvidos no processo.

A definição dos níveis de proficiência no PISA ocorre em duas fases. A primeira fase é baseada na análise de itens em relação às habilidades avaliadas, em cada área do conhecimento. Conseqüentemente, são produzidas descrições de habilidades em gradação de complexidade, retratando-se detalhadamente as demandas cognitivas dos itens. A segunda fase envolve decisões sobre: (i) onde fixar pontos-limite para níveis; e (ii) como associar os alunos aos níveis. Isso é uma questão de interpretação técnica e prática do significado de um determinado nível e tem conseqüências na divulgação de resultados, tanto nacional quanto internacionalmente. Nesse processo, primeiramente, é importante entender que as habilidades avaliadas pelo PISA devem ser consideradas em um *continuum*. Embora não existam pontos delimitando intervalos ao

[9] O Relatório Técnico Pisa 2003 foi utilizado como referência nesse tópico.

longo desse *continuum*, são arbitrados intervalos (níveis), com o propósito de possibilitar a comunicação sobre o escore da proficiência dos alunos e seu significado pedagógico-cognitivo. Os critérios utilizados para o arbitramento dos limites dos níveis de proficiência no PISA 2000 e reiterados em 2003 são os seguintes:

- o sucesso esperado de um aluno em um nível particular, em um teste contendo itens daquele nível, deve ser de pelo menos 50%, elevando-se para 80% para alunos com escores de proficiência próximos ao limite superior do nível;
- a amplitude dos níveis deve ser similar, garantindo-se a coerência conceitual dos níveis na escala e facilitando-se sua interpretação pedagógica;
- a probabilidade de um aluno, com escore de proficiência próximo ao meio do nível, responder corretamente a um item de dificuldade média do nível deve ser diretamente proporcional à elevação registrada em sua proficiência, definindo-se, algumas vezes, para a escala como "valor-RP", no qual "RP" indica "probabilidade de resposta correta".

Para cada nível arbitrado, espera-se que os alunos respondam corretamente, no mínimo, à metade dos itens relacionados àquele nível; ou seja, em um teste composto por itens com nível de dificuldade uniformemente dispostos ao longo daquele intervalo de proficiência, deverão acertar pelo menos 50%. Assim, alunos com escores de proficiência próximos ao limite inferior de um nível acertariam 50% dos itens, enquanto se esperaria que alunos localizados no meio e no topo desse nível alcançassem uma média de acertos bem maior. No topo de um nível, estariam os alunos que "dominam" as habilidades descritas naquele nível. Esses alunos, provavelmente, responderiam corretamente a cerca de 80% dos itens daquele nível. Ao mesmo tempo, estando no topo daquele nível, eles estariam próximos ao limite inferior do nível seguinte e, de acordo com os critérios utilizados, teriam a probabilidade de acerto em torno de 50% do nível subseqüente.

Os níveis construídos devem ter mais ou menos o mesmo intervalo, embora isso não se aplique ao nível de proficiência mais alto (acima de 669) nem ao mais baixo (abaixo de 358), pois eles não possuem bordas, ou seja, não possuem limites superior e inferior, respectivamente. Portanto, o nível abaixo de 358 foi retirado da análise por não possibilitar uma interpretação pedagógica significativa.

A cada uma das áreas de conhecimento avaliadas pelo PISA foram atribuídos os mesmos valores em suas escalas, apresentadas como subescalas, conforme os diferentes blocos de conteúdos, tendo-se como referência uma média de 500 e um desvio padrão de 100, sendo definidos, ainda, seis níveis de proficiência com intervalos eqüidistantes. Por exemplo, em 2003, na avaliação de Matemática, foram construídas 4 subescalas: (i) Análise Quantitativa; (ii) Espaço e Forma; (iii) Relações entre Gráficos, Tabelas, Textos e Fórmulas; (iv) Probabilidade e Estatística.

A Tabela 10 apresenta os níveis e os intervalos de proficiência estabelecidos para essas subescalas.

Tabela 10 – Níveis e intervalos de proficiência no PISA

NÍVEL	INTERVALOS DE PROFICIÊNCIA
6	ACIMA DE 669
5	607 A 669
4	545 A 607
3	482 A 545
2	420 A 482
1	358 A 420

Na aplicação dos critérios de seleção do PISA à construção da escala de proficiência de matemática do SAEB 2003 para a 4ª série de Ensino Fundamental, foram, primeiramente, calculados os escores alcançados pelos alunos. Para a determinação desses escores, utilizou-se o modelo de um parâmetro da Teoria da Resposta ao Item (modelo de Rash), onde

os resultados do BILOG-MG sofreram uma transformação linear de modo a apresentarem média igual a 176,89 e desvio-padrão de 48,304, valores correspondentes à média e ao desvio-padrão em matemática da quarta série do ensino fundamental na escala do SAEB. A seguir, para o processo de definição dos níveis de proficiência, adotaram-se os seguintes passos:

- Cálculo da proficiência para uma probabilidade de acerto de 50% e 80% para cada item.
- Cálculo da diferença de proficiência entre os pontos de probabilidade de acerto de 50% e de 80%. Como, no modelo Rash, todos os itens possuem a mesma discriminação (o que faz com que todas elas sejam idênticas, a não ser pelo parâmetro de dificuldade b que desloca as curvas para a direita ou para a esquerda, dependendo de sua dificuldade), essa diferença é a mesma para todos os itens e tem o valor de 72.4 pontos na escala considerada. Essa diferença de 72.4 pontos entre dois níveis consecutivos na escala do SAEB é correspondente à diferença de 62 pontos entre dois níveis consecutivos na escala do PISA, visto que, em ambos os casos, esses valores separam os grupos de estudantes avaliados conforme suas respectivas probabilidades de acerto ao item, definido anteriormente.

  Entretanto, vale observar que, na escala do SAEB, esse intervalo de 72,4 pontos equivale a cerca de 150% do desvio-padrão (48,304) de sua respectiva escala, ao passo que, na escala do PISA, o intervalo de 62 pontos corresponde somente a 62% do desvio-padrão desta escala. Considerando-se a metodologia aqui empregada, isto se explica pelo fato de que, como o SAEB tem seus itens construídos especificamente para o modelo de 3 parâmetros, a utilização de um modelo de Rash redundou em uma baixa discriminação para os itens. Essa baixa discriminação fez com que o intervalo correspondente à diferença entre a probabilidade de acerto de 80% e a probabilidade de acerto de 50% fosse, para o SAEB, muito maior do que a encontrada no caso do PISA.

No entanto, se procedimentos iterativos de seleção mais acurados dos itens fossem utilizados para obter uma escala com itens de maior discriminação, esse tipo de procedimento poderia acarretar uma perda de cerca de 50% da quantidade dos itens, o que prejudicaria a análise de ancoragem.

- Adotou-se este valor de 72.4 pontos como a amplitude dos diferentes níveis, sendo que o primeiro nível iniciou-se com o valor de 35.07, o qual corresponde à menor proficiência da base.

Na tabela 11, expõem-se os níveis calculados.

Tabela 11 – Níveis de proficiência no PISA

Nível	Níveis de proficiência correspondentes na escala do SAEB	Diferença
6	–	
5	Acima de 324.67	72.4
4	252.27 a 324.67	72.4
3	179.87 a 252.27	72.4
2	107.47 a 179.87	72.4
1	35.07 a 107.47	72.4

Definidos os níveis acima, procedeu-se à seleção dos itens em quatro subescalas, a saber: Espaço e forma; Grandezas e medidas; Números e Operações e Tratamento da informação. Em cada subescala, os itens foram selecionados nos níveis onde se encontravam seus respectivos parâmetros **b**, e os resultados obtidos estão apresentados no Anexo IX.

Identificados os itens selecionados de cada nível, passa-se à construção da Escala de Proficiência em Matemática, tendo como referência os critérios adotados pelo PISA, que se encontra no Anexo VI.

2.6.2
Análise comparativa entre os critérios de seleção utilizados pelo SAEB a partir de 1999 e o critério de interpretação de escala do PISA

A seguir, a tabela 12 apresenta o número de itens selecionados de acordo com cada um dos três critérios definidos pelo PISA, agrupados nas respectivas subescalas.

Tabela 12 – Número de itens selecionados de acordo com cada um dos três critérios do PISA

Níveis segundo o SAEB	Subescalas				Total
	Espaço e forma	Grandezas e medidas	Números e operações	Tratamento de informação	
1	1	1	3	3	8
2	8	12	31	5	56
3	13	33	41	4	91
4	1	4	5	1	11
5			2		2
Total	23	50	82	13	168

Uma diferença prontamente perceptível entre os critérios de seleção do SAEB e do PISA é a divisão dos níveis de seleção, visto que esses níveis têm extensões que dependem de cada programa. Enquanto, no SAEB, os níveis têm uma extensão de 25 ou 50 pontos, no PISA, os níveis tiveram sua largura determinada pela extensão das curvas de Rasch de cada item, tendo todas elas uma largura maior, de 72,4 pontos. Isso, naturalmente, resultou em um menor número de níveis de proficiência segundo os critérios do PISA (5 níveis), em comparação com o número de níveis de proficiência segundo o SAEB-2003 (8 níveis). A existência de um menor número de níveis tem vantagens e desvantagens. Por um lado, com menos níveis, consegue-se uma maior diferenciação entre os itens selecionados em níveis diferentes, o que é útil, por exemplo, para propósitos de interpretação do desempenho escolar. Por outro lado, um menor número de níveis dificulta a

realização de distinções mais finas de proficiência, o que pode comprometer a precisão da interpretação pedagógica dos itens e dos níveis de proficiência.

Outra diferença facilmente perceptível entre esses dois sistemas é o fato de a seleção do PISA se valer de subescalas (mais precisamente, em número de quatro, a saber: espaço e forma; grandezas e medidas, números e operações e tratamento de informação), ao passo que o SAEB não determinou a seleção especificamente para essas subescalas.

Quanto às semelhanças entre esses dois programas, observa-se que, nos dois, todos os itens são selecionados, sendo que ambos os critérios de seleção vinculam-se de forma simples e direta à proficiência atrelada a um percentual de acerto específico. Obviamente, fica, nesse caso, impossível comparar os critérios quanto aos níveis de seleção para itens iguais, visto que os níveis de seleção diferem consideravelmente entre um sistema e outro.

Observa-se também uma semelhança entre ambos os programas no que diz respeito à maior concentração de itens selecionados nos níveis centrais de seleção, com um decréscimo acentuado de itens selecionados nos níveis extremos (tanto para mais, quanto para menos) e suas vizinhas imediatas. Mais uma vez, a causa dessa diminuição de itens selecionados nas extremidades é a existência de poucos itens, cujas CCIs têm seus respectivos pontos de maior discriminação situados próximos aos extremos de proficiência.

Por fim, vale ressaltar a questão da arbitrariedade dos níveis de seleção, que já existia no caso do SAEB e persiste no caso do PISA. Nesse último caso, a determinação das extremidades de cada nível baseou-se na escolha inicial de um limite de proficiência mínima; entretanto também seria perfeitamente possível arbitrar outro valor para esse limite, bem como uma diferente largura para os níveis de proficiência, o que daria como decorrência um sistema consideravelmente diferente do atual quanto aos níveis de seleção dos itens.

2.7
O Projeto Geres: objetivos e ondas de aplicação

Iniciado em 2004, com financiamento da Fundação Ford e do PRONEX – Programa Núcleo de Excelência/CNPq, o Projeto GERES/2005 – Estudo Longitudinal sobre a Qualidade e Eqüidade no Ensino Fundamental Brasileiro – é uma pesquisa longitudinal na qual uma amostra de alunos e escolas de cinco importantes cidades brasileiras deve ser observada ao longo de quatro anos.

Tratando-se de uma pesquisa longitudinal de painel, uma mesma amostra de escolas e alunos será observada e avaliada ao longo de quatro anos, começando com alunos que entram na 1ª série em 2005 (Projeto GERES/2005 – Estudo Longitudinal sobre Qualidade e Eqüidade no Ensino Fundamental Brasileiro). Participam do GERES/2005 escolas públicas e privadas que oferecem as séries iniciais do Ensino Fundamental, nos municípios de Belo Horizonte (MG), Rio de Janeiro (RJ), Salvador (BA), Campo Grande (MS) e Campinas (SP).

Para acompanhar o progresso dos alunos na aprendizagem de leitura e matemática, propõe-se a aplicação de cinco ondas de testes cognitivos, focalizando as habilidades básicas tipicamente demandadas pela escola a alunos das séries iniciais, além de questionários contextuais, anteriormente mencionados. A medição da aprendizagem cognitiva dos alunos deve ser feita em cinco momentos diferentes, dentro das quatro primeiras séries do Ensino Fundamental. A primeira onda de aplicação, realizada em março de 2005, visou à aferição do nível de proficiência com que os alunos entraram na escola, ou seja, a detecção de habilidades e competências já desenvolvidas pelo aluno ao ingressar na 1ª série do Ensino Fundamental (ou seu equivalente, com a organização do tempo escolar em ciclos). A segunda onda, em novembro de 2005, teve como objetivo verificar a aprendizagem escolar na primeira série do Ensino Fundamental e, subseqüentemente, o mesmo se realizando em novembro de 2006, 2007 e 2008.

2.7.1
Metodologia de construção da escala de proficiência adotada pelo Projeto GERES

No Projeto Geres, a escala de proficiência em Matemática, com média 100 e desvio-padrão 25, foi construída a partir de informações fornecidas por cada item, em intervalos de 25 pontos, representativos de 6 níveis de proficiência.

Desse ponto de partida, fez-se a análise detalhada das habilidades desenvolvidas pelos alunos, considerando-se três fases particularmente importantes nesse processo, a saber:

1ª A introdução de uma habilidade/Início do desenvolvimento da habilidade.

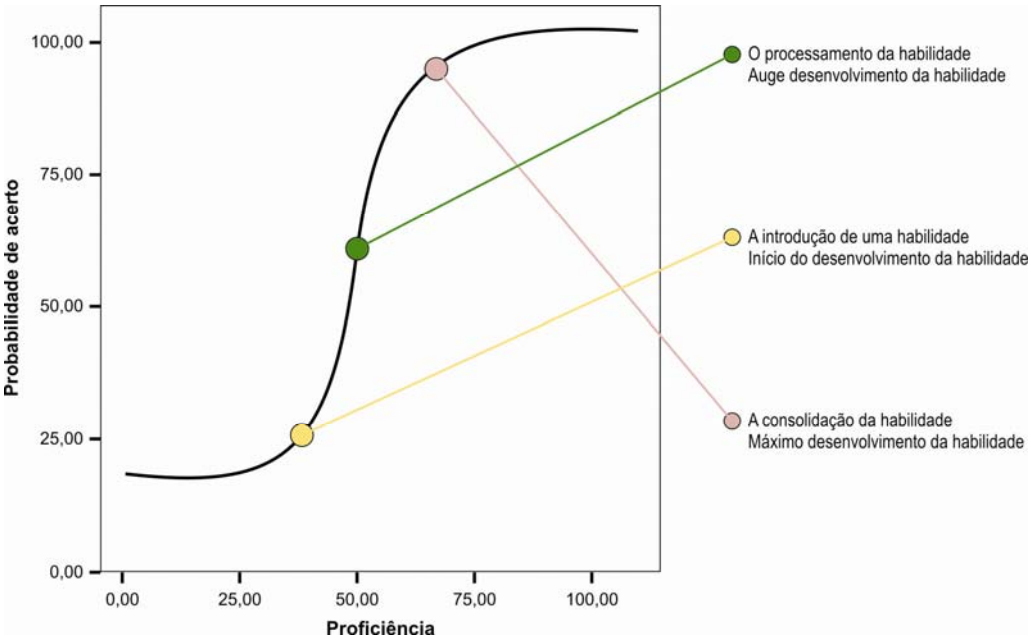
2ª O processamento da habilidade/Auge do desenvolvimento da habilidade.

3ª A consolidação da habilidade/Máximo desenvolvimento da habilidade.

A essas fases associam-se três pontos na CCI. O primeiro delimita o início do desenvolvimento de uma habilidade e é estatisticamente definido pelo ponto em que se encontra a maior taxa de crescimento da inclinação da CCI. Esse ponto, assinalado pela cor rosa claro, indica o nível de proficiência em que os alunos passam a ter maiores condições de desenvolver tal habilidade. O segundo ponto, assinalado pela cor verde, coincide com o parâmetro **b**. Ele indica o nível de dificuldade do item e diz respeito ao nível de proficiência que se espera dos alunos, levando-se em conta o acerto ao acaso (chute), ou seja, a probabilidades de acertarem ou errarem o item. A consideração da probabilidade de acerto ao acaso (parâmetro **c**) eleva a probabilidade de acerto associada ao parâmetro de dificuldade, o qual passa de um valor fixo de 0,5 (50%) para 0,5 + c/2. Esse ponto verde é, também, um marco do local em que o item fornece o máximo de informações para o modelo, ou seja, o ápice da Curva de Informação do Item (CII), onde a 2ª derivada da CII é igual a zero. Em torno desse ponto, pode-se dizer que a habilidade está no auge do desenvolvimento da habilidade, a CCI atinge a mais elevada inclinação, e, nessa região, o modelo oferece o poder de discriminação mais alto, ou seja, consegue-se distinguir melhor o grupo de alunos que desenvolveu a habilidade testada daquele que ainda não atingiu essa

etapa. O terceiro ponto sinaliza a consolidação da aprendizagem. Estatisticamente, é onde se localiza a maior taxa de decréscimo da inclinação da CCI. Pode-se considerar, nesse caso, que, a partir desse ponto, a habilidade está consolidada. A Figura 9 ilustra as três fases anteriormente descritas.

Figura 8 – Fases de desenvolvimento das habilidades

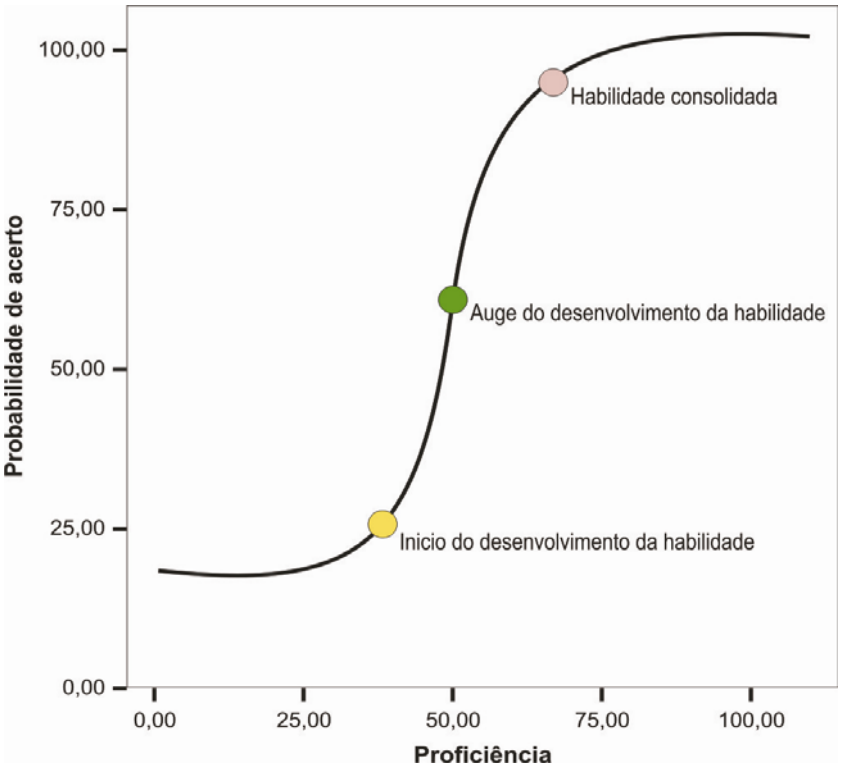


Na Curva Característica do Item (CCI), a representação estatística correspondente às três fases descritas dá-se nos pontos citados anteriormente, os quais podem ser traduzidos conforme o Quadro 5 e as Figuras 10 e 11.

Quadro 5 – Avaliação das três fases de desenvolvimento das habilidades

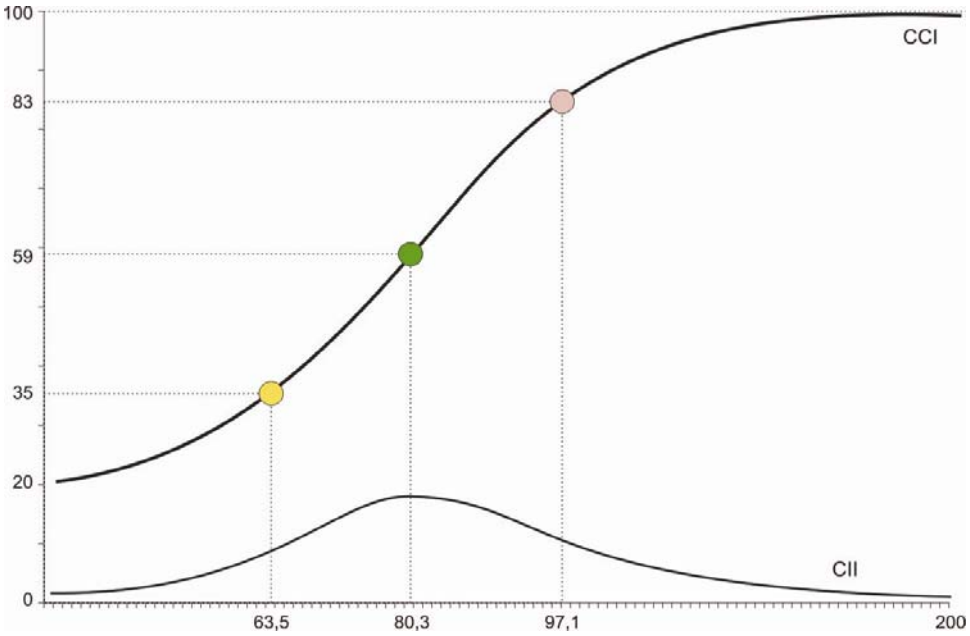
Ensino	Avaliação – Desenvolvimento da habilidade
1 – Introduz	CCI começa a subir rapidamente (início do desenvolvimento).
2 – Trabalha	Auge do crescimento da CCI (auge do desenvolvimento).
3 – Consolida	CCI começa a saturar (habilidade consolidada).

Figura 9 – Exemplo hipotético de Curva Característica do Item



Os modelos apresentados nos parágrafos anteriores se ajustam adequadamente aos dados do GERES. A seguir, apresenta-se um exemplo do comportamento das CCIs e CIIs para um mesmo item testado no GERES.

Figura 10 – Curva Característica do Item e Curva de Informação do Item (CII) para um item testado no GERES



Para representar as fases de desenvolvimento da habilidade na CCI, consideraram-se:

Início do desenvolvimento da habilidade – a **taxa de crescimento** da 1ª derivada da Curva Característica do Item (CCI) é máxima (3ª derivada da CCI = 0 (primeira raiz); teta < parâmetro **b**);

Auge do desenvolvimento da habilidade – o **AUGE da aprendizagem** ocorre no ponto em que a 1ª derivada da CCI do Item é máxima (2ª derivada da CCI = 0 = parâmetro **b**);

Habilidade consolidada – a **taxa de decréscimo** da 1ª derivada da CCI é máxima (3ª derivada da CCI = 0 (segunda raiz); teta > parâmetro **b**).

A região cor de rosa claro, antecedente à verde, indica que a habilidade requerida pelo item está em início de desenvolvimento conforme anteriormente proposto. A probabilidade efetiva de acerto situa-se em torno de 30%. Deve-se considerar, porém, para efeito de cálculo, que essa probabilidade efetiva de acerto é estimada levando-se em consideração o parâmetro **c**, ou seja, cresce em torno de 30% acima da probabilidade de acerto ao acaso. Essa primeira região tem início no

ponto onde a taxa de crescimento da 1ª derivada da CCI é máxima (3ª derivada da CCI = 0 (primeira raiz); teta < parâmetro B) e termina no limite inferior da região verde.

Após a região verde, inicia-se outra, cor de rosa escuro, que se refere à etapa de consolidação da habilidade requerida pelo item. Essa região tem início no limite superior da região verde e termina no ponto onde a taxa de decréscimo da 1ª derivada da CCI é máxima (3ª derivada da CCI = 0 (segunda raiz); teta > parâmetro B). A probabilidade efetiva de acerto gira em torno de 80%, considerado o parâmetro **c**, ou seja, 80% acima da probabilidade de acerto ao acaso. O erro padrão relativo aos pontos-limite da área de informação global do item, composta pelas três regiões descritas, foi desprezado.

Com os resultados dos testes de Proficiência em Matemática do SAEB 2003, 4ª série do Ensino Fundamental, foi produzido o quadro V, que se integra ao Anexo IX, em que se destacam o número dos itens, a posição no bloco, os parâmetros **a**, **b**, **c** da TRI e o **b** e **b** máximo.

De acordo com as informações produzidas, passa-se à construção da Escala de Proficiência em Matemática, segundo os critérios de seleção adotados pelo GERES, que se encontra no Anexo VII.

2.7.2
Análise comparativa entre os critérios de seleção utilizados pelo SAEB a partir de 99 e o GERES

Ao se compararem os critérios de seleção do GERES com os de outros sistemas de avaliação, é preciso antes lembrar que, para cada item, o GERES trabalha com dois pontos de referência, correspondentes ao ponto de auge do desenvolvimento da habilidade e ao ponto de habilidade consolidada, diferindo assim dos demais sistemas, que, em geral, consideram apenas um ponto de seleção. Em decorrência disso, serão feitas comparações entre o GERES e o SAEB 1999-2007, levando-se em conta cada um dos dois tipos de pontos considerados pelo GERES.

2.7.2.1
Comparações entre o SAEB 1999-2007 e o GERES no ponto de auge do desenvolvimento da habilidade

Observando-se as seqüências de itens selecionados segundo os critérios do SAEB 1999-2007 e os do GERES, para o auge do desenvolvimento da habilidade, percebe-se uma óbvia semelhança quanto à consideração de todos os itens como itens selecionados. Vale ressaltar, entretanto, que nem sempre os itens foram selecionados no mesmo nível segundo os dois critérios. Na verdade, dos 168 itens considerados[10], nada menos que 129 (um pouco mais de três quartos do total) foram selecionados em níveis diferentes segundo os dois critérios. Desses, a maioria (111) foi selecionada, segundo os critérios do GERES, um nível antes da seleção segundo o SAEB. Houve também um número expressivo (17) de itens separados por dois níveis de seleção, e um item (4975) para o qual a diferença foi de três níveis.

Pode-se observar, ainda, na tabela 13 que as diferenças são sempre menores ou iguais a zero, ou seja, não houve nenhum caso em que o nível de seleção do item, segundo o GERES, ficasse acima do nível de seleção do SAEB. Isso se explica pela modalidade de seleção no SAEB, baseada no percentual acumulado (65%) de acerto no item, sendo que, invariavelmente, para que esse percentual acumulado fosse atingido, era necessário avançar na CCI muito mais para a direita, em relação ao ponto de máxima derivada primeira (ponto esse usado na determinação da seleção segundo o GERES).

[10] Preferiu-se excluir um item (o de número 25060) desta análise, pelo fato de o mesmo ter apresentado um comportamento anômalo, que se traduziu em um fraco ajuste entre a curva característica do item e os dados observados dos percentuais de acerto ao longo dos níveis.

Tabela 13 – Disparidade entre os níveis de seleção segundo os critérios do SAEB (1999-2007) e do GERES (no auge do desenvolvimento da habilidade)[11]

Diferença	N. de Itens	%
-3	1	0,6
-2	17	10,1
-1	111	65,7
0	39	23,1
Total	168	100,0

O deslocamento na direção dos menores níveis de seleção, ocorrido no GERES, teve também o efeito de fazer surgirem itens selecionados no primeiro nível (de menor proficiência, correspondente a até 125 pontos) conforme mostram os dados apresentados na Tabela 14.

Tabela 14 – Número de itens selecionados por nível de proficiência, segundo os critérios do GERES (no auge do desenvolvimento da habilidade)

Nível de Proficiência	N. de itens selecionados	% Total	% Vál. Acum.
Até 125	8	4,8	4,8
125 – 150	11	6,5	11,3
150 – 175	25	14,9	26,2

[11] A coluna "Diferença" fornece a distância "nível de seleção no Geres – nível de seleção no SAEB (1999-2005)". Logo, uma diferença nula indica que o item selecionado foi posicionado no mesmo nível de proficiência segundo ambos os critérios. Uma diferença de -1 indica que o nível de seleção no GERES está imediatamente abaixo do nível de seleção no SAEB (1999-2005); uma diferença de -2, que o nível de seleção no Geres está dois níveis abaixo do nível de seleção no SAEB.

175 – 200	26	15,5	41,7
200 – 250	63	37,5	79,2
250 – 300	30	17,9	97,0
300 – 350	5	3,0	100,0
350 – 375	0	0	
Total Válido	168	100,0	
Itens não-ancorados	0	0	
Total	168	100,0	

Analogamente ao que ocorreu no SAEB, percebe-se no GERES alguma simetria na quantidade de itens selecionados ao longo dos níveis de proficiência, havendo um maior número deles no nível central de 200-250, com sua quantidade diminuindo paulatinamente na direção dos extremos (tanto máximo quanto mínimo) de proficiência. Também se nota que, no caso do GERES, o nível de maior proficiência (350-375) ficou vazio, uma conseqüência do deslocamento de níveis mencionado anteriormente.

2.7.2.2
Comparações entre o SAEB 1999-2007 e o GERES no ponto de consolidação da habilidade

Nesta nova comparação, percebe-se, na Tabela 15, uma tendência oposta à observada na Tabela 13, uma vez que, no GERES, os itens deslocam-se para as faixas superiores de seleção, devido à nova seleção estar relacionada a uma elevada probabilidade de acerto do item (correspondente a 80% ou mais). Isso fez os níveis de seleção dos itens, segundo o GERES, coincidirem bem mais com os do SAEB

(aproximadamente, 70% dos itens foram selecionados nos mesmos níveis segundo os dois critérios). Por outro lado, quando existentes, as discrepâncias entre os níveis de seleção são agora positivas em sua maioria, ou seja, fazendo-se a subtração (nível no GERES) – (nível no SAEB), percebe-se que, em 40 itens, a seleção para o GERES corresponde a um nível a mais em relação à seleção observada no SAEB. Houve também um item (16511), cujo ponto de seleção no GERES ficou dois níveis acima do ponto de seleção no SAEB. Além disso, dois itens (5018 e 29472) tiveram no GERES sua seleção localizada acima dos 375 pontos da escala, de modo que não puderam entrar na comparação com o SAEB, cuja escala para essa série não ultrapassa o limite dos 375 pontos. Por outro lado, ainda houve nove itens, cujo ponto de seleção no GERES ficou um nível abaixo do ponto de seleção no SAEB. Isso ocorreu porque, em seus respectivos pontos de habilidade consolidada (critério de seleção do GERES), ainda não havia um percentual acumulado de acerto grande o suficiente para selecioná-los segundo os critérios do SAEB.

Tabela 15 – Disparidade entre os níveis de seleção segundo os critérios do SAEB (1999-2007) e do GERES (consolidação da habilidade)

Diferença	N. de Itens	%
-1	9	5,4
0	116	69,9
1	40	24,1
2	1	0,6
Total	168	100,0

Tabela 16 – Número de itens selecionados por nível de proficiência, segundo os critérios do GERES (na consolidação da habilidade)

Nível de Proficiência	N. de itens selecionados	% Total	% Vál. Acum.
Até 125	0	0	0
125 – 150	1	0,6	0,6
150 – 175	9	5,4	6,0
175 – 200	17	10,2	16,3
200 – 250	61	36,7	53,0
250 – 300	48	28,9	91,9
300 – 350	26	15,7	97,6
350 – 375	4	2,4	100,0
Total Válido	166	100,0	

Comparações entre os seis critérios de seleção: considerações metodológicas

Uma vez que os itens selecionados norteiam todo o trabalho de produção e interpretação das escalas, seus limites estruturais se refletem também na escala.

Um dos limites da metodologia utilizada, seja pelo NAEP, seja pelo SAEB, seja pelo TIMSS, é a incerteza prévia de quantos itens escolhidos serão selecionados para cada nível de proficiência. Além disso, não se sabe também se existirão itens selecionados para todos os níveis de proficiência. Isso implica que o teste, obrigatoriamente, inclui um número grande de itens, uma vez que deverá haver itens selecionados suficientes e satisfatórios para todos os níveis previstos. Outro limite é a

perda de itens cujo nível de dificuldade (parâmetro b) encontra-se próximo da fronteira entre dois níveis de proficiência. Tais itens não informam se os indivíduos avaliados estão em um nível ou outro, o que provoca mais desajustes (ruído) no modelo do que agrega informação útil à interpretação, objetivo último da avaliação.

Observa-se ainda que, na tentativa de superação de alguns limites da seleção por critério único (SAEB), tanto o NAEP quanto o TIMSS recorrem a mais de um critério para a seleção de itens para construção de escala (seleção).

O TIMSS utiliza mais dois critérios para o processo de seleção. O principal objetivo é superar o nível reduzido de informação produzida em relação às habilidades desenvolvidas pelos alunos. Observa-se, portanto, a adoção de critérios adicionais em detrimento da definição de uma técnica mais acurada para análise dos itens.

O NAEP também utiliza mais de um critério. Uma conseqüência negativa dessa opção foi a possibilidade de comitês diferentes chegarem a interpretações diferentes de pontos de seleção. A dubiedade levou os comitês a abrirem mão de avanços na seleção para manter a coerência da escala (Mullis:1991).

Cabe, ainda, mencionar que a apresentação de amostras de itens nos vários pontos de seleção tem levado à interpretação não fidedigna em escalas, como no NAEP. Isto é, a interpretação distorcida da amostra de itens leva à conclusão de que apenas os alunos com escores acima dos pontos de seleção responderam corretamente aos itens, enquanto que, na verdade, alguns alunos com escores abaixo do ponto de referência também responderam ao item corretamente, especialmente quando se trata de itens de múltipla escolha. Para minimizar a distorção na interpretação, o NAEP passou a publicar, com a amostra de itens, a porcentagem efetiva de alunos, em cada grau ou nível de idade, que responderam corretamente ao item.

O PISA adota a TRI no modelo de Rasch e outra modalidade de apresentação de resultados. A dependência em relação aos pontos de seleção é grande, uma vez que são esses elementos que relacionam

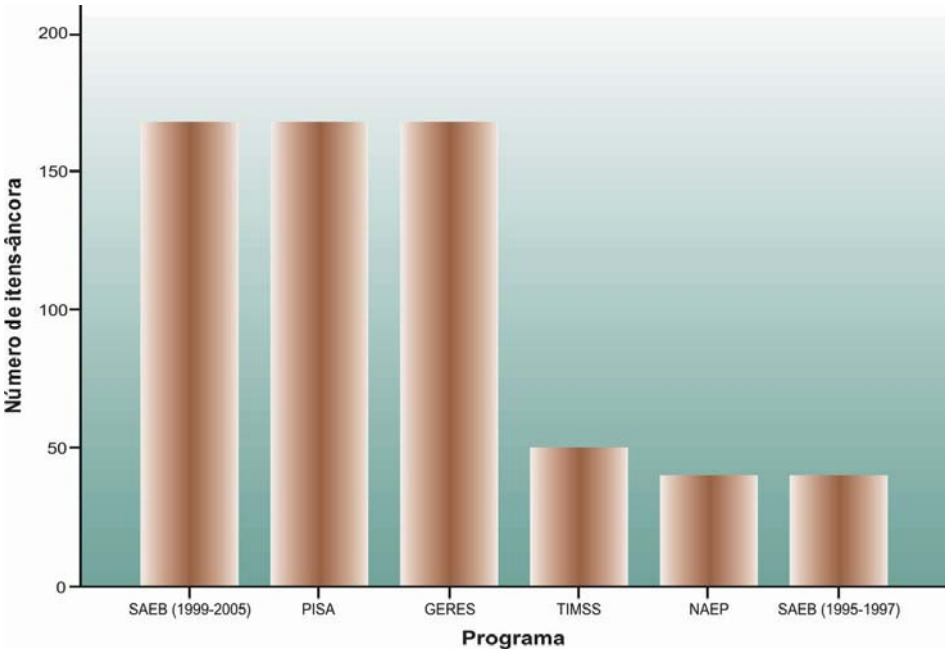
habilidades (itens) a uma tipologia (níveis) via percentual de acerto. Assim, somente podem ser incorporadas interpretações dos itens que se ajustem adequadamente ao modelo, apresentando dentro dos limites de cada nível 50% a 80% de acertos e percentuais significativamente diversos nos níveis acima e abaixo.

Pode-se observar que uma implicação direta da utilização da metodologia adotada no Projeto GERES é a diminuição do número mínimo de itens necessários ao teste. Outra implicação é que não se perde informação de nenhum item, o que resulta em um ajuste mais preciso entre o modelo, os dados e a interpretação decorrente.

Números de itens selecionados e níveis de seleção

O Gráfico 1 mostra o número de itens selecionados segundo os diferentes critérios considerados. Em relação ao TIMSS e ao NAEP, que adotaram mais de um critério de seleção, são apresentados os números referentes aos principais critérios. Em relação ao GERES, os números referem-se aos itens selecionados no ponto correspondente ao auge do desenvolvimento da proficiência, um critério segundo o qual todos os itens puderam ser selecionados.Se se tivesse considerado, porém, o ponto onde ocorre a consolidação da aprendizagem, esse número seria ligeiramente (duas unidades) menor, visto que, no ponto de consolidação da aprendizagem, dois itens ficaram além do limite de 375 pontos da escala em questão, conforme se comentou anteriormente. Acredita-se, contudo, que tal discrepância não é relevante na comparação que será feita entre o GERES e os demais programas de avaliação.

Gráfico 1



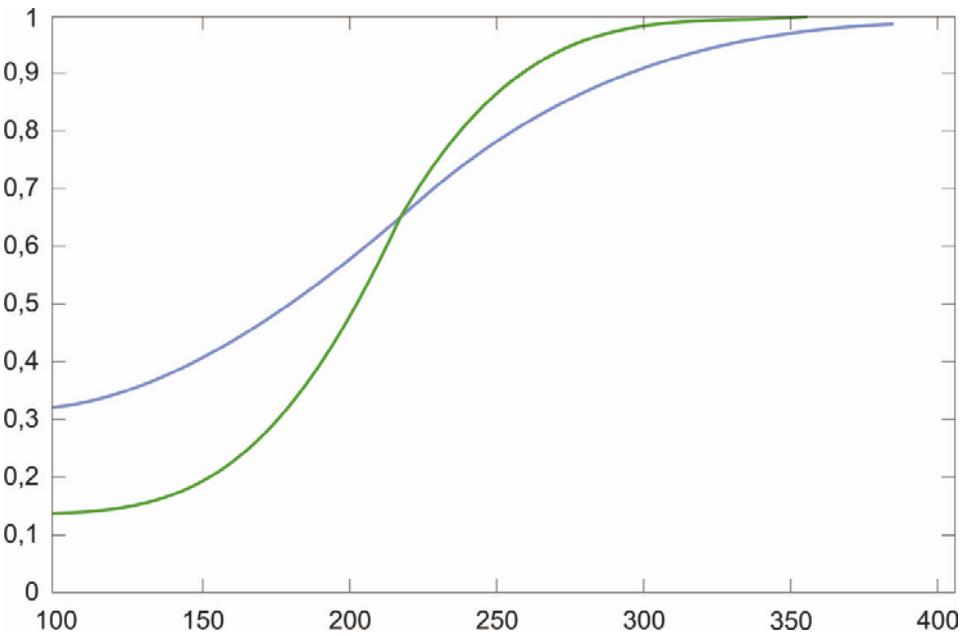
Programa	Itens-âncora
SAEB (1999-2005)	168
PISA	168
GERES	168
TIMSS	50
NAEP	39
SAEB (1995-1997)	39

Tanto o SAEB (1999-2007), quanto o Projeto GERES adotaram a TRI no modelo de três parâmetros, utilizando o conjunto de itens selecionados correspondente à totalidade dos itens do banco. Entretanto, como já se observou, os itens, em geral, foram selecionados em níveis diferentes segundo esses critérios, principalmente no caso em que, no GERES, foi utilizado o ponto da CCI correspondente ao auge do desenvolvimento da aprendizagem, situação na qual os níveis de seleção, segundo o GERES, tenderam a ser menores que os níveis de seleção segundo o SAEB. Por outro lado, no caso em que se utilizaram para o GERES os pontos correspondentes à consolidação da aprendizagem, o número de discrepâncias diminuiu, e, nas situações em que essas diferenças de fato se verificaram, ocorreu o contrário, ou seja, os níveis

de seleção, segundo o GERES, tenderam, em geral, a ser maiores que os níveis de seleção segundo o SAEB.

Uma característica indesejável do critério de seleção adotado pelo SAEB é a arbitrariedade decorrente da exigência de 65% de acerto no nível de seleção. Como se adota o modelo de três parâmetros, muitas vezes ocorre que itens, sendo selecionados em um mesmo nível (ou seja, correspondendo a 65% de acerto nesse nível), apresentam curvas bastante diferentes entre si, decorrentes do fato de terem, por exemplo, parâmetros de discriminação e de acerto casual muito diferentes. O gráfico 2 ilustra essa situação, mostrando as CCI's dos itens 25471 e 26325, que foram selecionados no nível 5 da escala de proficiência do SAEB; esses itens, entretanto, possuem grau de dificuldade muito próximos, sendo que um item possui uma boa discriminação, e o outro, não. Além disso, apresentam características distintas quanto ao acerto casual.

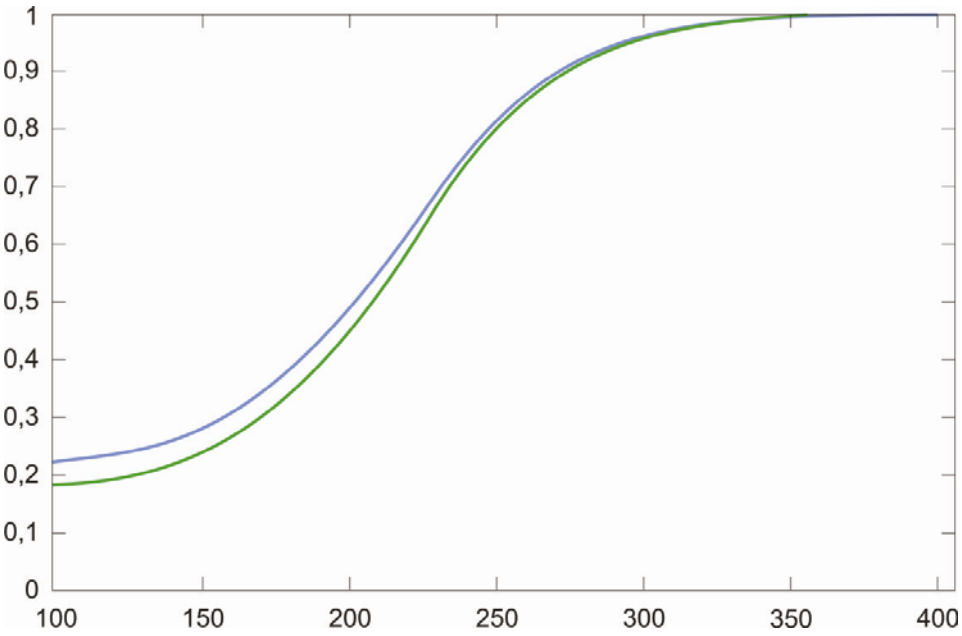
Gráfico 2



CCI	N. do item	Parâmetro **a** (discriminação)	Parâmetro **b** (dificuldade)	Parâmetro **c** (acerto casual)
Azul	25471	0,0138	212,3978	0,27498
Verde	26325	0,0252	207,4464	0,12802

Em outros casos, o inverso tem-se verificado, ou seja, há a ocorrência de itens com curvas bem semelhantes, porém sendo selecionados em níveis diferentes, porque pequenas diferenças entre seus parâmetros os levaram a apresentar os percentuais exigidos de acerto em níveis diferentes. O gráfico 3 refere-se aos itens 25205 e 24967, de baixa dificuldade, próxima ao ponto 215 na escala, selecionados, no entanto, em níveis diferentes (respectivamente, níveis 5 e 6).

Gráfico 3



CCI	N. do item	Parâmetro **a** (discriminação)	Parâmetro **b** (dificuldade)	Parâmetro **c** (acerto casual)
Azul	25205	0,021	215,59	0,21
Verde	24967	0,021	216,11	0,17

Cabe mencionar que esses problemas ocorreram no SAEB e também em outros sistemas que se basearam em percentuais de acerto (como o NAEP) para definir os critérios de seleção.

Por sua vez, a seleção dos itens feita segundo os critérios adotados pelo Projeto GERES não se baseia em percentuais arbitrários de acertos, mas, sim, na capacidade máxima de discriminação dos itens, o que confere a esse critério uma característica mais fundamental.

Finalizando, outra característica importante apresentada por alguns desses sistemas (TIMSS e NAEP) diz respeito à definição de itens selecionados apenas para alguns intervalos de proficiência, criando, por conseguinte, "vazios" no intervalo total de proficiência. Mais uma vez, o SAEB e o Projeto GERES apresentaram, quanto a esse aspecto, um comportamento semelhante, produzindo itens selecionados distribuídos por todo o *continuum* de proficiência. No caso do SAEB, entretanto, perde-se de vista o ponto exato de seleção do item, pois a análise é feita considerando o percentual de acerto no intervalo de proficiência como um todo. Por outro lado, no Projeto GERES, tanto no caso do auge do desenvolvimento da habilidade, quanto no caso de sua consolidação, é possível determinar um nível bem mais estreito (por exemplo, correspondente, no primeiro caso, a um intervalo centrado no parâmetro **b** e levando em conta o erro da medida, que, geralmente, é da ordem de frações de ponto de proficiência).