2
Avaliação e Currículo

Avaliações diagnósticas em larga escala, objetivando monitorar os sistemas educacionais e dar subsídios para a implementação de reformas e direcionar políticas públicas com vistas à efetividade e eficácia dos sistemas, especialmente em relação à educação básica, são cada vez mais comuns. Essas reformas são orientadas, geralmente, com base na produtividade e na necessidade de se alcançarem metas que favoreçam a inserção dos países na lógica da competitividade, imprescindível em um mundo cada vez mais globalizado e regido pelo livre mercado (Candau, 1998). Na ação dessas reformas, os temas prioritários e as propostas a serem colocadas em prática variam em função dos diferentes contextos. Contudo, há certo consenso nos discursos que apontam a necessidade de um currículo mais sintonizado com o mundo atual, cujas competências exigidas são radicalmente distintas das exigências existentes em épocas passadas.

No contexto desta tese, cabe uma reflexão acerca de como nossas escolas vêm ocupando-se desse tema, o que está sendo ensinado e o que os alunos estão conseguindo aprender. Da mesma forma que Forquin (1993), considero que, no cotidiano escolar, o currículo toma diferentes feições e, mesmo que possa não existir enquanto documento, está presente nos espaços que se destinam à prática educativa. Desta forma, currículo é tudo o que se ensina nas salas de aula, e que, muitas vezes, é diferente daquilo que foi prescrito, a princípio.

Este capítulo, o qual inicio com as idéias associadas ao conceito de currículo, é destinado à revisão da literatura. Trabalhando com os resultados das avaliações educacionais, direcionei o foco no sentido de melhor compreender o significado de "currículo aprendido" que, naturalmente, remete-nos às considerações acerca da apropriação, por parte dos alunos, do que é ensinado em Matemática, nas salas de aula. Em seguida, mostro como as avaliações de sistemas educacionais podem ser utilizadas na estratégia de apreensão desse currículo. Por fim, apresento os conceitos e aspectos mais relevantes dos estudos envolvendo a identificação de DIF, com ênfase específica em mostrar o quanto essa ferramenta estatística é capaz de extrair dos resultados dos testes os padrões

de similaridades que possibilitam a comparabilidade dos sistemas no que os sistemas podem ser comparados e, com a mesma eficiência, captar as diferenças.

2.1

Os Currículos: oficial, ensinado e aprendido

Nas últimas décadas, a discussão em torno da seleção do conhecimento escolar colocou em destaque a relação entre dominação econômica e cultural e o currículo escolar, inserindo-o no interior da discussão político-sociológica. Estudos críticos do currículo passaram a enfatizar que a seleção do conhecimento escolar não é um ato desinteressado e neutro, é sempre resultado de uma seleção feita (por aqueles que desejam manter seu *status quo*) sobre o conhecimento de um todo; sendo, por conseguinte, culturalmente determinado e historicamente situado, não podendo ser desvinculado da totalidade social.

No campo do currículo, notadamente a partir da emergência das teorias crítico-pedagógicas, o conhecimento escolar tem-se constituído em objeto de preocupação de diferentes autores. O currículo - entendido como reflexo do momento histórico em questão, e diretamente vinculado às relações de poder; à organização e à estruturação da sociedade; e, por fim, à visão de mundo do grupo social dominante - tem papel fundamental na educação, pois é nele que estão definidos os pontos-chave a serem abordados no desenvolvimento do indivíduo-cidadão. Os processos de seleção, organização, distribuição e estratificação dos conteúdos curriculares têm sido foco de inúmeros estudos na tentativa de melhor identificar e discutir os interesses subjacentes, buscando-se alternativas, principalmente nas relações entre conhecimento escolar e poder, ou seja, buscando-se entender como o currículo contribui para reforçar divisões sociais referentes à classe social, etnia e gênero (Moreira, 2000).

Nessa abordagem crítica, questionam-se a cultura erudita, as disciplinas tradicionais e seus conteúdos, chegando-se, mesmo, a colocar em xeque a própria racionalidade com que a escola vem trabalhando. Definidos pelos que detêm o poder, os currículos são vistos como construções históricas e como instrumentos de controle das classes sociais que não estão na dominante.

Outra corrente crítico-teórica defende a transmissão, na escola pública, dos conteúdos que compõem o saber sistematizado, usualmente mais restrita às escolas dos grupos sociais privilegiados. Nesse enfoque, o domínio de tais

conteúdos é visto como indispensável à luta por melhores condições de vida e por ascensão social (Moreira, 2000).

Esses distintos pontos de vista em relação ao caráter universal ou relativo, científico ou ideológico, do conhecimento escolar geram posicionamentos diferentes quanto ao que se deve incluir nos currículos, o que, consequentemente, reflete-se nos processos de escolha, organização, ensino e avaliação dos conteúdos nas escolas e nos sistemas escolares.

As decisões sobre currículo, seja no nível das políticas e das propostas curriculares oficiais, seja no nível de sua materialização nas escolas, merecem especial atenção, particularmente naquilo que é ensinado e aprendido em sala de aula.

O currículo pode designar não somente aquilo que é formalmente prescrito, oficialmente inscrito no programa, mas aquilo que é realmente ensinado nas salas de aula, e que está, às vezes, muito distante do que é oficialmente prescrito (Forquin, 1995). Logo, constitui-se não só como o programa das matérias, mas também como o percurso de formação na escola dos conteúdos e conhecimentos escolares.

O currículo pode denotar, também, o "currículo latente" do ensino ou da socialização escolar. Em outras palavras, o conjunto de competências ou disposições que se adquirem na escola, por experiência, impregnação ou familiarização, e que não estão previstas no currículo oficial. Este sentido mais abstrato do conceito (denominado por "currículo oculto") completa o que Forquin (1995, p. 188) chama de toda a "dimensão cognitiva e cultural da escolarização". A escola não é apenas um local onde se estabelecem relações de poder e relações interpessoais, mas, por excelência, é o espaço institucional privilegiado, por onde circulam saberes e símbolos da sociedade moderna.

Dentro desse contexto, o papel do currículo é naturalizar a seleção cultural, tornando-a senso comum, fazendo com que esse conhecimento representado seja resultante da tradição, ou seja, os valores mais reconhecidos são os valores mais cultivados. Essa naturalização, que deveria ser fruto de um processo, não de interferência de indivíduos, mas sim de uma conseqüente participação e integração de toda a sociedade, acaba sendo o resultado da tradição seletiva de uma classe dominante que escolhe os aspectos da cultura mais convenientes dentro da sua concepção, visando privilegiar seus interesses. Por conseguinte, tais

peculiaridades culturais acabam por serem transmitidas à sociedade através da escola, em virtude desta ter a capacidade de tornar popular o conhecimento.

Segundo Forquin (1993, p14) *"toda educação, e em particular toda educação de tipo escolar, supõe sempre na verdade uma seleção no interior da cultura e uma reelaboração dos conteúdos da cultura destinados a serem transmitidos às novas gerações"*. Essa nova geração será, portanto, apresentada ao mundo com base em valores e princípios determinados por um grupo específico, através do currículo.

> A cada geração, a cada "renovação" da pedagogia, e dos programas, são partes inteiras da herança que desaparecem da "memória escolar", ao mesmo tempo que novos elementos surgem, novos conteúdos e novas formas de saber, novas configurações epistêmico-didáticas, novos modelos de certeza, novas definições de excelência acadêmica ou cultural, novos valores". (Forquin,1993, p.15)

A escola tem sido, ao longo desses anos, espaço de legitimação de uma cultura hegemônica, oriunda de um processo perpétuo de seleção do conhecimento, o qual depende do momento que está sendo vivido e da classe social que está no poder.

Diferentes escolas podem fazer diferentes tipos de seleção no interior da cultura. Os docentes podem ter hierarquias de prioridades divergentes, mas todos eles e todas as escolas fazem seleções de um tipo ou de outro, no interior da cultura (Forquin, 1992, p.31). Forquin utiliza o termo currículo para designar essas seleções. Evidencia-se, aqui, o questionamento sobre quais seriam esses aspectos da cultura, quais seriam esses conhecimentos, atitudes, valores, que justificam as despesas de toda a natureza que supõe um ensino sistemático e sustentado por um aparelho institucional complexo. Uma sociologia comparada dos programas escolares revelará o caráter, instável, aleatório, e até mesmo arbitrário dessa seleção.

Forquin (1992) utiliza as denominações "currículo formal", "currículo ensinado" e "currículo aprendido" como aspectos possíveis dessa seleção no interior da cultura, conforme nos consideremos construtores de programas e responsáveis oficiais, ou sob o prisma de meros docentes em suas salas de aula.

> Os conteúdos prescritos pelas autoridades – o currículo formal ou oficial – são o produto, ao longo do tempo, de todo um trabalho de seleção no interior da cultura acumulada; um trabalho de organização de mudanças das delimitações de abalo das hierarquias entre as disciplinas. Quanto aos conhecimentos em via de serem elaborados, os autores de programas, ao menos quando eles não se atrasam em demasia, transpõem-nos em função, principalmente, da idéia de que eles fazem dos públicos escolares. Para o autor, aquilo que é realmente aprendido, retido e compreendido pelos alunos não corresponde tampouco àquilo que os docentes ensinam (currículo ensinado) ou crêem ensinar, e que esta inadequação pode se tornar, por sua vez, o objeto de uma investigação sociológica, pois a recepção da mensagem (currículo aprendido) depende do contexto social e cultural (Forquin, 1992, p.32).

A consciência dessas diferentes concepções acerca do currículo leva a um deslocamento no modo de olhar a função redistributiva do Estado Moderno. A ele cabe promover políticas globais e articuladas, atribuindo-lhes caráter moderador de desigualdades sociais e econômicas e, ainda, responder, de forma eficaz, ao aumento das demandas, no contexto de uma maior divisão do trabalho e da expansão do mercado, na sociedade de massas. A educação é, portanto, dever do Estado e direito do cidadão, pois sendo concebida como valor social, reflete-se como instrumento da sociedade, na efetivação do processo de formação e construção da cidadania.

Não obstante, a evolução das idéias relativas à educação, e principalmente à avaliação, consolida-se em torno dos valores econômicos, como conseqüência do rápido desenvolvimento tecnológico e da nova ordem globalizada. A educação passa a ser direcionada para o novo estilo de desenvolvimento, reproduzindo as relações de poder e subordinação, presentes nesse modelo.

A compreensão de aspectos relacionados à escolha dos conteúdos do ensino e de sua incorporação aos programas escolares pode possibilitar um olhar mais crítico para questões, até então, restritas apenas ao plano pedagógico. A escola, como o local - por excelência - nas sociedades modernas, de gestão e de transmissão de saberes, valores, crenças e hábitos produzidos e acumulados pela humanidade (Forquin, 1992), constitui-se na principal ferramenta para uma educação cidadã. A partir desta, então, será viável a construção de uma sociedade onde as bases da economia e da política permitam que as relações humanas se dêem a partir dos princípios de eqüidade, justiça social e participação cidadã, nas diferentes instâncias de decisões.

O objeto do meu estudo relaciona-se com as discussões da seleção de conteúdos escolares e das diferentes dimensões do currículo, uma vez que as avaliações educacionais em larga escala comparam o rendimento dos sistemas educacionais, baseando-se, para isso, em tópicos do currículo escolar e no entendimento e na distinção entre "currículo oficial", "currículo ensinado" e "currículo aprendido". Parto, então, do pressuposto que o desempenho dos estudantes em Matemática nas avaliações de larga escala são medidas do "currículo aprendido". O interesse volta-se para a compreensão das diferentes ênfases curriculares, no sentido de quais conteúdos e áreas da Matemática são selecionados pelas escolas e pelas redes de ensino, utilizando-se como estratégia de análise os resultados dos alunos nas avaliações educacionais.

2.2

As Avaliações Educacionais em Larga Escala

O Brasil, a partir dos anos de 1990, começou a desenvolver um consistente sistema de informações educacionais, motivado pelo reconhecimento da importância desse instrumento para uma gestão eficiente dos programas e das políticas públicas em Educação. Desde então, os avanços alcançados são notórios, destacando-se a completa reestruturação do Instituto Nacional de Estudos e Pesquisas Educacionais – Inep - cuja principal missão tem sido a produção de informações quantitativas e qualitativas, a fim de subsidiar a formulação e implementação de políticas públicas, nos diferentes níveis de governo, bem como apontar tendências que sinalizem a necessidade de mudanças de rotas.

A efetivação de sistemas padronizados de avaliação de larga escala é, portanto, um fenômeno relativamente novo no Brasil. O Sistema Nacional de Avaliação da Educação Básica – SAEB - uma das primeiras realizações com a finalidade de conhecer os resultados da aprendizagem dos alunos em nível nacional foi implementado em 1990. Após quase duas décadas de existência, já se consolida como o mais amplo instrumento de avaliação externa da qualidade do desenvolvimento de habilidades e competências dos estudantes do país, e como um dos mais sofisticados e amplos sistemas de avaliação em larga escala da América Latina (Araújo e Luzio, 2005). Aplicado em alunos das séries finais de ciclos do Ensino Fundamental – 4ª e 8ª séries e da 3ª série do Ensino Médio, tem como objetivos medir o desempenho escolar e levantar informações sobre o perfil

socioeconômico e cultural dos alunos, bem como seus hábitos de estudo. Por meio de questionários dirigidos aos professores e diretores, é produzido um conjunto de variáveis relacionadas ao perfil e às práticas pedagógicas daqueles, e ao perfil e às práticas de gestão escolar destes. São coletadas, também, informações sobre equipamentos e infra-estrutura das escolas selecionadas na amostra.

Ainda, a partir de meados da década de 1990, em uma demonstração evidente de que a avaliação dos sistemas educacionais passa a assumir um papel estratégico na orientação de políticas educacionais, novos levantamentos foram sendo implementados, envolvendo diferentes níveis de ensino, dentre os quais, destacam-se: o Exame Nacional do Ensino Médio – ENEM; o Exame Nacional de Cursos – ENC e a reformulação do sistema de avaliação da pós-graduação. Essa nova concepção, oficialmente expressa nas mudanças introduzidas pela nova Lei de Diretrizes e Bases da Educação Nacional (LDB Nº 9394/96)[1], justifica-se pela constatação de que não há país no mundo preocupado em aumentar a eficiência, a eqüidade e a qualidade do seu sistema educacional que tenha ignorado a importância da avaliação como mecanismo de acompanhamento dos processos de reforma. Cada vez mais se atribui relevância - tanto à *avaliação institucional* em suas diferentes dimensões (condições da infra-estrutura das instituições escolares; processos de gestão; formação, qualificação e produtividade dos recursos humanos; etc.) - como em relação à *avaliação de resultados* (o que e como os alunos aprendem; quais os fatores associados ao rendimento escolar; impactos de fatores extra e intra-escolares na aprendizagem; etc.) (Castro, 1999).

Embora recentes, como já ressaltado acima, sobretudo quando comparados à tradição de países desenvolvidos, os sistemas nacionais de avaliação já atendem às expectativas de melhorias nas informações disponíveis acerca do processo de aprendizagem dos alunos, nos diferentes níveis e, tão importante quanto, tais sistemas lograram êxito ao introduzir uma cultura avaliativa nas diferentes instâncias educativas. Além disso, a repercussão alcançada pelos resultados dessas avaliações na mídia tem contribuído para introduzir na agenda pública a discussão e o debate sobre a melhoria da qualidade da educação. Nesse contexto, percebe-se

[1] A Lei de Diretrizes e Bases da Educação Nacional, Lei nº 9.394, de 20 de dezembro de 1996, estabelece que ao Ministério da Educação cabe o monitoramento dos sistemas e a administração das unidades públicas; a coordenação do Fundo de Manutenção e Desenvolvimento do Ensino Fundamental e de Valorização do Magistério; o financiamento e a implementação de políticas nacionais de elevação de qualidade e inclusão educacional: a distribuição de livros didáticos, a merenda escolar e o transporte escolar.

uma importante mudança cultural entre os dirigentes e gestores dos sistemas de ensino, que passam a reconhecer as avaliações externas das escolas como ferramentas úteis ao monitoramento de suas ações. Assim, as redes estaduais e municipais, além de se engajarem na implementação dos projetos de âmbito nacional, passaram a desenvolver e a colocar em prática seus próprios sistemas de avaliação. Entenderam que estes são instrumentos estratégicos e eficazes, tanto para subsidiar a formulação de políticas, como para o acompanhamento da qualidade da educação pela possibilidade de comparabilidade dos resultados ao longo do tempo.

O grande desafio, todavia, ainda está em fazer com que essas informações cheguem à ponta da linha e o uso que se faz delas. Não basta aceitar e reconhecer a importância da avaliação, ainda é preciso difundir melhor as informações geradas em tantos levantamentos, procurando facilitar suas interpretações, por meio de uma linguagem adequada, para que, nas escolas, sejam adequadamente analisadas e debatidas, envolvendo professores, diretores, pais, alunos, acadêmicos e membros da sociedade civil na participação do processo em seus vários momentos (Ravella, 2000; Locatelli e Andrade, 2001).

Essas análises e discussões sobre os resultados dos dados colhidos mediante os processos de avaliação ajudam a buscar respostas para algumas das principais indagações com que se deparam os gestores e demais agentes inseridos no processo educativo. Através delas é possível identificar prioridades e alternativas, a fim de que se eleve a eficácia das ações e se otimizem os investimentos no setor. Além disso, indicam o que se deve esperar que os alunos aprendam em sua trajetória escolar à luz dos currículos propostos e identificam quais são os fatores escolares ou extra-escolares que favorecem ou limitam a aquisição das competências esperadas. Por fim, elas, ainda, imprimem aos sistemas de avaliação o mérito de garantirem transparência à sociedade em geral, ao difundirem, em números e indicadores, os resultados do que acontece em sala de aula e em que condições.

Adicionalmente, assumem um papel de destaque, quando conseguem estabelecer uma estreita relação com o esforço das redes de ensino pela oferta de melhores serviços educacionais, nos processos de planejamento e na adequação de currículo, dentre outras ações. Para tanto, da mesma forma que, dentro do contexto escolar, a avaliação feita pelo professor se integra naturalmente ao

processo ensino-aprendizagem, as avaliações externas devem ser encaradas como um componente complementar ao trabalho desenvolvido nas próprias redes de ensino. Enquanto as avaliações dos professores, de caráter mais formativo, visam oferecer ao docente - e ao próprio aluno - elementos para se desvendarem os fatores e condições que podem ajudar ou dificultar o processo de construção do conhecimento, as avaliações de sistemas educacionais, mais padronizadas e precisas, tendem a avaliar o produto da aprendizagem e, através da possibilidade da comparabilidade, apresentar diagnósticos.

Uma vez vencidas as críticas e as resistências iniciais, pelo menos em parte, a concordância de idéias em torno da importância estratégica de se investigarem, detalhadamente, os níveis de qualidade da educação, bem como das variáveis que impactam os resultados do processo educativo, tem ganhado terreno. Esse avanço tem feito com que a avaliação educacional se firme como um campo de pesquisa privilegiado, na busca pelo desenvolvimento educacional, em uma perspectiva de cooperação multilateral. Apoiados por associações e organismos internacionais, têm prosperado diversos projetos que promovem estudos internacionais comparados. Em 1997, o Brasil participou do Primeiro Estudo Internacional Comparado, realizado pelo Laboratório Latino-americano de Avaliação da Qualidade da Educação, vinculado à Oficina Regional de Educação para América Latina e Caribe (OREALC/Unesco). Com o objetivo de avaliar os níveis de aprendizagem em Linguagem e Matemática de alunos das 3ª e 4ª séries da escola primária, e os fatores a ela associados, participaram desse estudo treze países da região. Essa cooperação tende a ampliar seus horizontes, a partir da implementação do Plano de Ação em Educação, aprovado em reunião dos chefes de Estados da Cúpula das Américas, realizada em Santiago, Chile (1998).

Além disso, recentes acordos de cooperação bilateral na área educacional, firmados pelo Brasil, Estados Unidos e a Inglaterra, também definem a avaliação como uma das áreas de maior interesse para o desenvolvimento de parcerias e de cooperação técnica. Outra iniciativa na área de avaliação internacional da qual o Brasil já participa é o PISA – *Programme for International Student Assessment* , cujos resultados são o foco de minha pesquisa e cujas características serão detalhadas mais adiante.

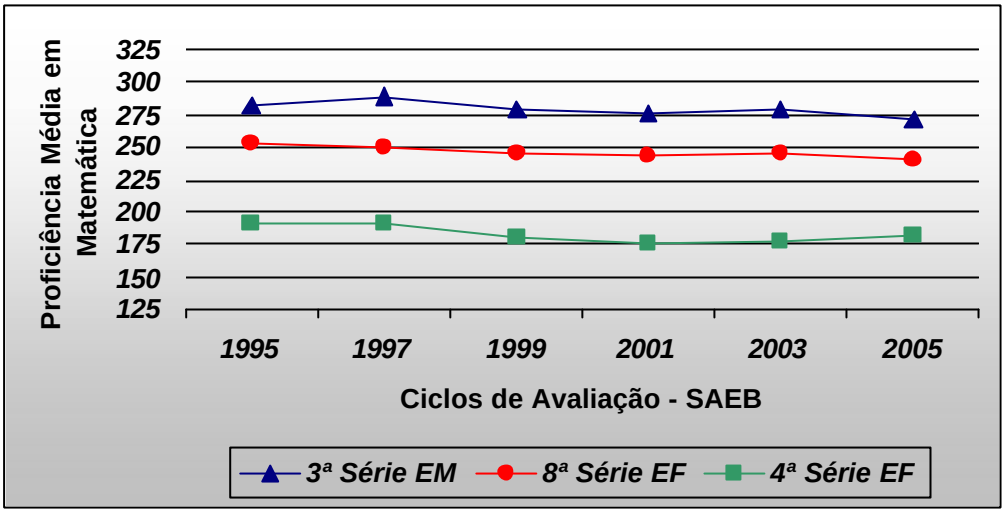
No entanto, resistem ainda, algumas críticas às avaliações de larga escala. Essas se apóiam, notadamente, nos argumentos de que as grandes avaliações

servem para aumentar o controle governamental sobre as escolas, ou ainda, por trazerem em si uma tendência homogeneizadora para a educação, em virtude do estabelecimento de padrões. Dois contra-argumentos podem ser apresentados para refutar essas críticas: - a definição de padrões de proficiência, desejáveis ao término de cada etapa da escolarização, pode ser a bússola de que necessitam as escolas e professores, orientando-os para o que deve ser ensinado e o que deve ser aprendido, o que resolveria, com maior eficiência, o problema da eqüidade; - embora nem sempre os números e indicadores sejam os que gostaríamos de ver divulgados, há que se ressaltar a relevância dos dados gerados na orientação das reformas educacionais e, sobretudo, na introdução da preocupação com a melhoria da educação, nos compromissos assumidos pelos governos.

O fato é que, ao longo dessas duas últimas décadas, há uma atmosfera nova em relação a nossa Educação Básica. Não só pelo fato de que agora temos maior quantidade e precisas informações para anunciar, mas também porque a Educação passou a ser assunto de primeira página, revelando ter assumido uma posição de destaque nas prioridades do país. Infelizmente, o que não mudou muito foi a qualidade da educação fornecida pelo Estado. A cada anúncio dos resultados de desempenho escolar de nossos alunos fica a certeza de que a educação nacional, de um modo geral, e, em especial, a Matemática, não tem cumprido as tarefas de ensinar de forma eficiente e nem de promover a eqüidade.

Os indicadores de desempenho nacional, extraídos do Saeb, mostram, de forma clara, não apenas o baixo aprendizado dos alunos brasileiros, mas, também, a distribuição desigual desse aprendizado ao longo do tempo. A situação se torna ainda mais preocupante, se observarmos que o Saeb avalia apenas o elementar de cada disciplina, que seria o mínimo necessário para a formação de leitores competentes e para que estudantes utilizem o instrumental matemático, de forma eficiente, na resolução de problemas. Também, no PISA, o desempenho de nossos alunos é muito aquém do desejado. Os dados desse estudo internacional, objeto de investigação desta pesquisa, serão apresentados e discutidos detalhadamente numa seção à parte. No gráfico abaixo, é apresentada a série histórica do Saeb, em Matemática, onde é possível concluir, que uma parcela considerável de nossos alunos não está aprendendo o mínimo esperado para as etapas de escolarização avaliadas.

Gráfico 1: Médias de Proficiências em Matemática – BRASIL – 1995 - 2005[2]



Fonte: Inep

O gráfico apresenta as médias de desempenho dos alunos de 4ª e 8ª séries do Ensino Fundamental e da 3ª série do Ensino Médio, em Matemática, no período de 1995 a 2005. Ao apresentar esses resultados, cabe destacar algumas considerações acerca da questão da comparabilidade. Embora o primeiro ciclo do SAEB tenha ocorrido em 1990, apenas a partir de 1995, houve a intenção explícita de garantir a comparabilidade e, assim, formar uma série histórica de resultados da avaliação. Neste ciclo foi introduzido o uso da Teria de Resposta ao Item (TRI), para a construção de instrumentos, atribuição de escores e análise, de forma a viabilizar a comparação dos resultados. Os resultados obtidos a partir da TRI são independentes de grupos e não sofrem influência quanto ao grau de dificuldade dos testes. Em avaliações educacionais, são utilizadas técnicas de ligação que permitem que dois ou mais resultados sejam comparáveis. O procedimento adotado nas análises do SAEB tem por base a metodologia da equalização e das técnicas estatísticas, para transformação dos escores, a fim de que, ao final, os resultados sejam comparáveis.

Os dados acima confirmam tendências identificadas em pesquisas anteriores, que apontavam uma baixa efetividade na aprendizagem em

[2] As medias dos anos de 1995, 2003 e 2005 foram estimadas incluindo o estrato de escolas públicas federais.
Em todos os anos, a zona rural foi avaliada e incluída para a estimativa das médias apenas na 4ª série.
Para a composição do estrato rural não foi incluída a Região Norte em 1997 e em 1999 e 2001, apenas participaram os estados da Região Nordeste, Minas Gerais e o Mato Grosso.

Matemática. Além disso, indicam que o descompasso entre o que é proposto pelos currículos e o desempenho real dos alunos se acentua, a partir das séries finais do Ensino Fundamental e do Ensino Médio. A média de proficiência satisfatória para a 4ª série do Ensino Fundamental em Matemática é de 200 pontos. Como em Matemática, mais do que qualquer outra disciplina, a progressão curricular exige a necessidade de dispor os conteúdos curriculares de maneira lógica e seqüencial, alcançar esse nível, no mínimo, garante que as competências e habilidades desenvolvidas pelos alunos irão permitir o progresso escolar com mais qualidade.

Em todos os ciclos do Saeb, no entanto, a média nacional esteve abaixo do mínimo de 200 pontos. Em 1995, a média foi de 191 pontos, portando 9 pontos abaixo. Em 2005, com média de 182 pontos, a diferença passa a ser de 18 pontos, aumentando a distância para o mínimo esperado. Observa-se, todavia, que após uma tendência de queda mais fortemente acentuada até o ciclo de 2001, a tendência atual é de crescimento.

No que se refere aos resultados da 8ª série do Ensino Fundamental, o cenário é bastante semelhante ao da 4ª série, especialmente pela comum tendência de queda até 2001. No entanto, ao contrário do que ocorre na 4ª série, após certa estabilidade entre 2001 e 2003, em função das médias desses dois ciclos estarem em intervalos de confiança com interseção, a diferença aumenta em 2005. Por outro lado, a queda da 8ª série foi mais suave. No entanto, sendo a média mínima satisfatória, para esta série, de 300 pontos, percebe-se que a distância da média em relação à média mínima na 8ª série é maior do que na 4ª série, em todos os ciclos, evidenciando-se o acúmulo dos déficits de aprendizagem, ao longo do Ensino Fundamental.

Finalmente, uma análise dos desempenhos no Ensino Médio, cuja média estabelecida como satisfatória para os estudantes concluintes deste nível de ensino é de 375 pontos, revela que as distâncias entre as médias alcançadas e a média desejável são maiores do que as verificadas nas outras duas séries investigadas pelo Saeb, o que revela um desempenho muito aquém do razoável. Considerando que esses resultados refletem o aprendizado de toda a Educação Básica, e não apenas o que foi aprendido no Ensino Médio, tal quadro demonstra que o déficit de aprendizagem, ao longo do Ensino Fundamental, pelas características da Matemática, tem um reflexo acentuado neste nível de ensino.

Uma outra forma de se analisar o rendimento escolar é verificando a freqüência de alunos em cada um dos intervalos das escalas de desempenho das áreas de conhecimento avaliadas. Desta forma, é possível observar, mais detalhadamente, o desenvolvimento de habilidades pelos estudantes, bem como a quantidade de alunos, por estágios, desse desenvolvimento. Pelas descrições das respectivas competências relacionadas a cada estágio na escala, é possível, ainda, identificar a eficácia, na aquisição de determinadas competências e habilidades.

A escala de Matemática é constituída por níveis, arbitrariamente definidos, com as respectivas interpretações pedagógicas do que representa estar situado em cada nível, englobando as seguintes subáreas da Matemática: Espaço e Forma; Grandezas e Medidas; Números e Operações, e Tratamento da Informação. As mesmas que compõem os blocos de conteúdos descritos nos Parâmetros Curriculares Nacionais de Matemática e, com pequena diferença na nomenclatura, as mesmas subáreas avaliadas no PISA.

De acordo com os dados do SAEB-2005[3], em Matemática, 70% dos estudantes da 3ª série do Ensino Médio obtiveram pontuação inferior a 300 pontos, valor correspondente à média mínima satisfatória para alunos da 8ª série. Para aqueles, ao final da escolarização básica, mesmo que consigam completar seus estudos, isso pouco representará em termos de oportunidades sociais. Seja no sentido de melhor colocação no mercado de trabalho, seja no usufruto de uma cidadania plena, com participação consciente na política e na cultura do país, pois a inclusão educacional que lhes foi oferecida não veio acompanhada de qualidade de aprendizado.

Diante dos inegáveis indicadores de baixo rendimento em Matemática nos diferentes níveis de escolarização, têm sido recorrentes as críticas ao ensino-aprendizagem dessa disciplina, na literatura especializada. O problema é antigo, mas só recentemente têm-se números para comparar, exemplificar e ter idéia de sua dimensão. Recorrente é também a preocupação de pedagogos, psicólogos e professores, na busca por uma solução para esse problema, e para identificarem as causas que contribuem para que esse ciclo gerador do fracasso escolar nessa área do conhecimento se perpetue. Uma das conclusões a que se chega com essas avaliações em larga escala é a de que a Educação Básica no Brasil apresenta como

[3] www.inep.gov.br

uma de suas principais características o descompasso entre o currículo oficialmente proposto, o currículo ensinado e o currículo aprendido. Esses resultados obtidos confirmam a pouca efetividade do currículo proposto ou indicado, revelando que o mesmo não está sendo ensinado (e, conseqüentemente, aprendido) de forma satisfatória, pois é pequeno o número de alunos que apresentam um desempenho próximo ao que seria desejável, em relação à proposta curricular.

Segundo Brito (1990), entre professores, tende a haver concordância sobre o que seria um domínio básico em Matemática, a ser progressivamente atingido ao longo das séries escolares: realização das quatro operações aritméticas fundamentais; cálculo e uso de medidas; razões, proporções e porcentagens; resolução de problemas, realização de estimativas e apreciação de resultados; conhecimento de Geometria e Álgebra; uso de conceitos elementares de probabilidade e estatística. De fato, para a NCSM[4] – *National Council of Supervisors of Mathematics* - esses aspectos, aliados a uma particular ênfase à compreensão de conceitos e princípios matemáticos e à sua utilização na solução de problemas do cotidiano, e à capacidade de raciocinar com clareza, de fundamentar e comunicar idéias matemáticas, integram as "habilidades matemáticas básicas para o século XXI" (Lorenzato, Vila, 1993). No entanto, não obstante esse consenso quanto ao que deve ser ensinado, o desempenho matemático de nossos alunos ainda é muito ruim.

Após um exaustivo levantamento bibliográfico de dissertações, teses e artigos, na área de Educação Matemática, compreendendo o período de 1983-1994, Hoff (1996) aponta as principais críticas dirigidas ao modelo pedagógico predominante no ensino-aprendizagem da Matemática. Aparecem, em muitos desses estudos, alguns dos fatores associados ao fracasso da aprendizagem dessa disciplina na escola. Para a autora, esses fatores *se articulam como partes de um mesmo quadro, cujo ponto inicial se localiza na concepção de Matemática prevalecente. Essa concepção se desdobra numa prática de ensino presente desde as séries iniciais até o 3º grau e se reitera nos cursos de formação de professores, realimentando o status quo do processo ensino/aprendizagem* (Hoff, p. 76).

[4] NCSM é uma organização internacional de pesquisadores que colaboram para que se alcance a excelência e eqüidade em Educação Matemática, em todos os níveis.

Prevalece, ainda, entre muitos professores, a concepção de Matemática como uma ciência pronta, perfeita e irrefutável. Um saber neutro, desde sempre existente, e não um produto cultural.

> "(...) é aquela que não duvida. Aceita. É aquela que não argumenta. Impõe. É aquela que não põe problemas. Apenas os resolve. É aquela que não tem processo e nem produtores. Apenas produtos. É aquela que não tem história. Surgiu pronta do nada e predestina-se ao nada e a ninguém. É aquela que não induz à curiosidade. Conforma-se com tudo igual. É vítima do hábito. É aquela que renunciou à capacidade de pensar e pensar-se. Que renunciou à condição de ciência". (Miguel, 1994. pp. 53-60)

Assim, para esses professores, aprender Matemática é uma questão de habilidade, de talento inato, e as causas do fracasso no processo ensino-aprendizagem extrapolam a sala de aula.

Do ponto de vista do aluno, na lógica dessa concepção de Matemática, cabe-lhe essencialmente a memorização. Assumir o papel passivo de ouvinte e o esforço em praticar e reproduzir soluções. Não há espaços para dúvidas e erros, para recriar relações estabelecidas por outros ou, ainda, para a elaboração de conhecimento. O que se tem, na ponta da linha, é a aquisição isolada e fragmentada de determinados conteúdos, e não a efetiva compreensão destes. Assim, é impossível para os alunos a transferência do conhecimento matemático para o âmbito de outras disciplinas e, mesmo, para que aquele passe a se constituir em ferramentas matemáticas úteis na interpretação e soluções de problemas reais.

Os problemas do sistema educacional brasileiro são muitos, e não se restringem à dimensão do pedagógico. Há os problemas, ainda, com o fluxo educacional e a grande quantidade de alunos que estão fora da série adequada à sua idade. Entendemos, no entanto, que a melhoria da educação brasileira, como um todo, e a solução desses problemas, em particular, passam, primordialmente, por um processo ensino-aprendizagem mais eficiente.

2.3

O Funcionamento Diferencial do Item (DIF)

Estudos visando identificar itens que sejam favoráveis a um determinado grupo, em detrimento de outros, ganham destaque na campo da psicometria moderna, pois ajudam a assegurar que os testes sejam tão imparciais quanto é possível fazê-los. Nesse sentido, Soares (2005) destaca que a preocupação com o

comportamento diferencial do item antecede ou, ainda, extrapola o contexto da TRI, onde a ausência do DIF é requisito para uma boa equalização entre resultados de grupos diferentes de alunos.

2.3.1– Contexto Histórico

Historicamente, a preocupação com o DIF (Differential Item Functioning) está fortemente associada ao desejo de se construirem questões de teste que não fossem afetadas por características étnico-culturais dos grupos submetidos aos testes de avaliação educacional (*cf.* Cole, 1993) – muito ligada, portanto, às campanhas em prol da melhoria dos direitos civis dos cidadãos comuns, nos anos de 1960, nos Estados Unidos da América. Esses anos foram marcados por uma enorme preocupação com a igualdade de oportunidades, pelas críticas aos sistemas educacionais discriminadores, pelo desenvolvimento de um conceito popular e legal de ações afirmativas e pela consciência racial/étnica. Diferenças educacionais, resultantes de sistemas educacionais com muita iniqüidade, bem como diferentes taxas de empregabilidade em bons empregos, passaram a ser vistas como vestígios de uma velha ordem segregadora. Assim, escores de testes, refletindo essas diferenças, foram considerados, da mesma forma, discriminadores e passou-se a usar o termo *viés*, ao referenciá-los.

Os resultados dos processos de avaliação educacional, executados por reconhecidas instituições, tal como o *Educational Testing Service (ETS)*, foram discutidos por diversos intelectuais, como sociólogos e pedagogos. Para eles, as diferenças de rendimento, observadas entre os diversos grupos étnicos e socioeconômicos, refletiam, na realidade, disparidades nas oportunidades educacionais e discriminação contra grupos minoritários de negros, hispano-americanos, judeus e árabes (Allen, Wainer, 1989, citados por Andriola,2006). Da mesma forma, os trabalhos de Jensen (1980), sobre o viés dos testes de aptidão cognitiva aplicados nos EUA, apontaram para a mesma direção. Mostraram grandes diferenças de desempenho dos indivíduos de raça negra e hispânica quando comparados aos brancos. Esses resultados contribuíram para fomentar a polêmica em torno dos testes, que se mostravam injustos ao exigirem tarefas estranhas às culturas de algumas minorias.

Nesse contexto, psicometristas passam a ser alvos de críticas; e os testes, acusados de serem parciais. Para esses técnicos, no entanto, a neutralidade era inerente aos testes, sendo o uso que era feito deles, para o bem ou para o mal, o responsável pela eventual parcialidade. Eles acreditavam que o papel do teste não era resolver males sociais do mundo, mas reportar, com neutralidade, o que nele acontecia. Ironson (Ironson,1982, in: Angoff, 1993) menciona o exemplo de um item, o qual indagava sobre a temperatura adequada para se assar um bolo. O item, mesmo que estatisticamente parcial contra os meninos, pode ser bem apropriado para um teste de seleção de cozinheiros e padeiros.

Têm início, assim, estimulados pela discussão social, alheia, em grande parte, ao círculo psicométrico, estudos para desenvolver formas de identificar o viés, nos itens e nos testes. Esses estudos tinham por objetivo provar que os testes ou instrumentos de medida não possuíam nenhum tipo de viés (Cole,1993). Então, sob a seguinte concepção de viés: *um item é enviesado se sujeitos de habilidades iguais, mas de culturas diferentes, não têm a mesma probabilidade de acertar o item* (Angoff, 1973; Linn, Levine, Hastings & Wardrop, 1981; Shepard, Camilli & Averill, 1981; Ironson, 1982; Linn & Drasgow, 1987), muitos pesquisadores começaram a se dedicar ao estudo sistemático das diferenças entre os grupos étnicos, com o objetivo de tentarem encontrar explicações convincentes para esse fenômeno.

A partir de achados em estudos sobre o viés de itens e testes, realizados em 1951 por pesquisadores da Universidade de Chicago, que haviam encontrado variações nos itens, em aspectos bastante peculiares, tais como conteúdo e formato (Hambleton, Swaminathan, Rogers, 1991), surgem os primeiros dados a respeito dos problemas técnicos presentes em determinados itens utilizados na avaliação da aprendizagem. Um desses problemas técnicos era o uso indevido da linguagem escrita. Muitas vezes, determinados termos empregados nos testes, mais familiares aos norte-americanos brancos, proporcionavam certa vantagem a este grupo, em detrimento dos grupos minoritários, que desconheciam ou não empregavam cotidianamente os mesmos termos.

Essa discussão em torno do viés dos testes tem duas origens distintas, mas igualmente relevantes: uma preocupação de caráter técnico ou psicométrico, e outra, de caráter social. Enquanto a preocupação psicométrica restringia-se à definição do conceito de viés descrito acima, a preocupação social esteve sempre

ligada ao uso que se fazia dos resultados dos testes. Tem a ver com justiça social, está intimamente ligada ao problema do preconceito e às questões sobre desigualdades de oportunidades. Assim, é natural que haja divergências entre essas duas correntes, diante de um item que apresente comportamento diferente para grupos distintos, de mesma habilidade. Para uns, o item que for considerado enviesado, em um sentido social, deve ser retirado do teste, por estar prejudicando um dos grupos. Para outros, um bom item em um teste, em termos de qualidade e validade da medida, é exatamente aquele que consegue captar a diferença, caso ela realmente exista. Ou seja, se dois grupos, por exemplo, brancos e negros, são diferentes na proficiência por questões sociais, meio cultural inferior, ou diferenças naturais, um item deve aferir essa diferença, para que seja um item válido.

Discussões dessa natureza acabaram por gerar certa confusão ou ambigüidade em torno do termo viés, e ainda provocam controvérsias na área de avaliação até os dias atuais, embora essas divergências, entre as considerações técnicas e sociais, tenham sido maiores, nos anos de 1970. Naquela época, quando o conceito adquiriu popularidade em psicometria, o termo *"viés"* foi usado como sendo *"um tipo de invalidação que prejudica um grupo mais do que o outro"* (Shepard et al. P.318). Ou seja, a possível razão para o baixo desempenho de minorias devia-se à ênfase dada a conhecimento e habilidades próprias da cultura da classe média branca, excluindo-se as culturas das minorias. Nota-se que a primeira definição refere-se às simples observações de diferenças de desempenho, enquanto a segunda extrapola a diferença e carrega um sentido de valor, referindo-se a seus efeitos indesejáveis sobre um grupo.

Para Angoff (1993), é clara a existência de um conflito semântico: a palavra viés vinha sendo usada simultaneamente, mas quase incompreensivelmente, com dois significados inteiramente diferentes: social e técnico (estatístico). A conseqüência era a introdução de uma falta de clareza desnecessária em uma atmosfera política já confusa, na qual o viés era apontado como a causa da grande disparidade dos escores. Sugestões foram feitas para restringir o uso de viés a observações estatísticas e para definir outro termo destinado a julgamentos e avaliações de sentido social. Finalmente, a expressão "funcionamento diferencial do item" (DIF) passou a ser usada, referindo-se à

simples observação de que um item mostra propriedades estatísticas diferentes, em diferentes conjuntos de agrupamentos.

Alheios a essas considerações, os testes, todavia, continuam sendo mensageiros de más notícias: expõem diferenças de grupo as quais a sociedade preferiria acreditar que estivessem superadas, ou, que não existissem. Já no contexto técnico ou psicométrico, tem havido mudança, e o pertencimento a um grupo é característica incluída em procedimentos técnicos de avaliação, pelos psicometristas. Permanece, contudo, uma preocupação adequada para separar padrões técnicos de questões sociais.

Andriola (2002) destaca que a idéia de *grupo* é central nas diversas definições de viés e, por esse motivo, ele tem sido estudado, fundamentalmente, nas investigações acerca das diferenças relacionadas com algumas características grupais, como: classe social, idade, região, *habitat*, ou outra característica sociodemográfica relevante.

2.3.2- Definição

Na Teoria Clássica das Medidas, tem-se como unidade de análise a prova como um todo. Normalmente se trabalha com o escore dos indivíduos, que é uma medida facilmente observada, a partir da aplicação de um teste. Na Teoria de Resposta ao Item (TRI), a unidade de análise é o item, e a resposta do indivíduo a cada item é utilizada nos modelos matemáticos convenientes para estimar-se a proficiência do indivíduo, na área de conhecimento desejada.

A grande vantagem da TRI sobre a Teoria Clássica é que a primeira nos permite fazer comparações que antes só seriam possíveis se todos os indivíduos fossem submetidos às mesmas provas ou, pelo menos, às formas paralelas de teste. Em contrapartida, como as ferramentas matemáticas utilizadas na TRI são bem mais complexas, exigindo uma análise estatística mais sofisticada, o distanciamento entre quem processa os cálculos e aqueles a quem efetivamente interessam esses resultados fica maior, dificultando a compreensão e a interpretação adequada das informações produzidas. Outra questão a ser considerada é que o aumento da utilização de técnicas derivadas da TRI, na área de Avaliações Educacionais, faz com que aumente também a circulação de textos e relatórios, os quais fazem referência à boa parte dos termos e técnicas

estatísticas utilizados. É recomendável, pois, que os textos destinados a educadores e gestores da área que não tenham, necessariamente, formação matemática sejam precedidos, sempre que possível, de uma explicação e definição dos termos e técnicas aplicadas. Seguindo esse preceito, passo a algumas considerações, para uma melhor compreensão do conceito de DIF.

As investigações para a detecção do DIF têm por base uma mesma argumentação: *a existência do DIF é um fator que influencia a validade da interpretação, que é realizada a partir da pontuação obtida pelo sujeito num item ou teste* (Andriola, 2002). Ou seja, a estimação da proficiência não pode ter sido gerada sob o impacto de itens com algum tipo de funcionamento diferencial, sob pena de não se constituir numa medida confiável.

Além dos cuidados na elaboração dos itens, a padronização das condições de aplicação dos instrumentos de medida, no que tange ao tempo destinado à resolução do teste, às instruções fornecidas, às condições do ambiente e à própria aplicação, merece especial atenção e se constitui em requisito básico para todo processo avaliativo que se quer justo.

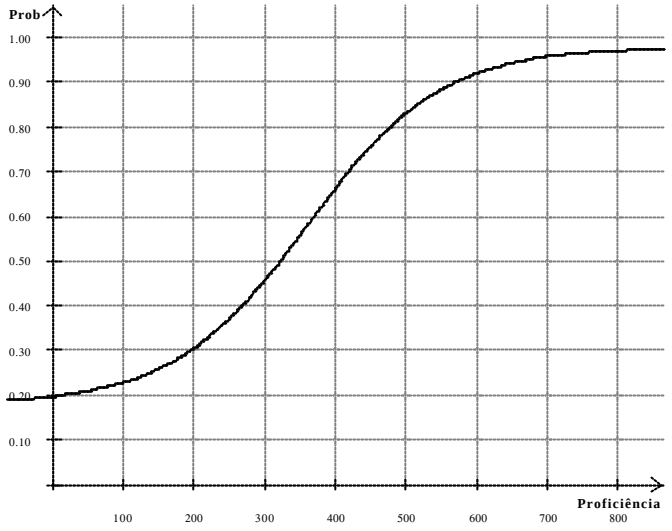
No âmbito da TRI, é possível dizer que o item não tem DIF, quando a curva característica do item (CCI) é a mesma para os grupos comparados em um mesmo nível de habilidade ou proficiência (q) medida através do item. Em linguagem matemática, podemos expressar a ausência de DIF com respeito à variável G (grupo) dado Z (nível de q) se, e somente se, $F(X \mid g,z) = F(X \mid z)$, onde:

- X é a pontuação no item (X=1 correto, X=0 errado);
- G é o valor obtido, segundo a variável G;
- Z é o valor obtido, segundo a variável Z.

Nesse contexto, os valores esperados por $E(X \mid g,q) = E(X \mid q)$ para todo g e q. No caso de itens dicotômicos, os valores esperados são as probabilidades de acerto ao item, que podem ser expressas nos seguintes termos:

$P(X = 1 \mid g,q) = P(X = 1 \mid q)$, para todo g e q. No segundo caso $P(X = 1 \mid q)$, a equação expressa, na realidade, a curva característica do item (CCI), representada na figura abaixo (Andriola,2006).

Fig. 1: Curva Característica do Item (CCI)



Geralmente, os estudos para a determinação do DIF utilizam dois grupos, denominados de *referência* (GR) e *focal* (GF). O termo *grupo*, utilizado nesta tese, refere-se a subdivisões de uma população. Especificamente, a *população* considerada são os alunos de 15 anos de idade, dos 41 países participantes que realizam os testes do PISA-2003, e os *grupos* são os países selecionados para as análises.

2.3.3 – Diferentes Tipos de DIF no Âmbito da TRI

Para uma melhor compreensão dos tipos de DIF que irei apresentar, cabe retomar alguns conceitos básicos da Teoria de Resposta ao Item. A TRI propõe a utilização de modelos que representam a probabilidade de um indivíduo responder corretamente a um item, como função da habilidade cognitiva ou proficiência do respondente. Essa função é sempre expressa de tal forma que, quanto maior for a habilidade, maior será a probabilidade de acerto, no item.

Dentre os modelos propostos pela TRI, um dos mais utilizados é o modelo logístico de três parâmetros, cuja equação é dada por:

$$P(X_{ij}=1|q_j) = c_i + (1-c_i)\frac{1}{1+e^{-Da_i(q_j-b_i)}}$$

Com i= 1,2,3,....., I itens e j = 1, 2, 3.....,n indivíduos , onde:

X_{ij} é uma variável dicotômica que assume os valores: 1, quando o indivíduo j responde corretamente ao item i; ou 0, quando o indivíduo j não responde corretamente ao item i.

q_j habilidade (traço latente) do j-ésimo indivíduo.

$P(X_{ij}=1|q_j)$ é a probabilidade de um indivíduo j com habilidade q_j, responder corretamente ao item i.

b_i é o parâmetro de dificuldade (ou de posição) do item, medido na mesma escala da habilidade.

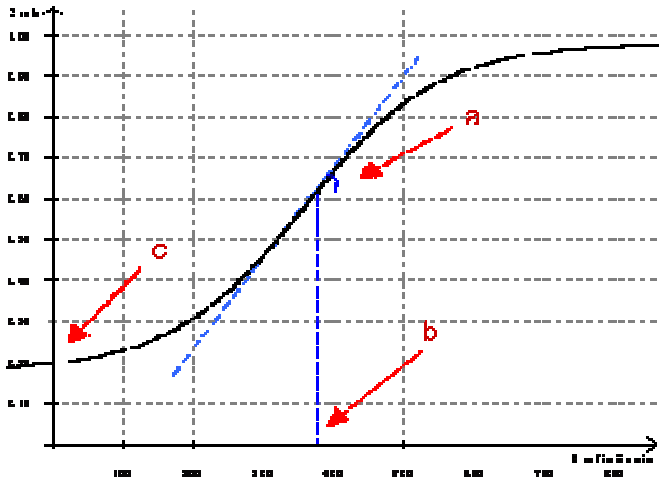
a_i é o parâmetro de discriminação (ou de inclinação) do item i. Refere-se à capacidade do item de distinguir alunos com diferentes níveis de habilidade.

c_i é o parâmetro do item que representa a probabilidade de indivíduos, com baixa habilidade, responderem corretamente ao item i (acerto casual).

D é um fator de escala, constante e igual a 1. Utiliza-se o valor 1.7, quando deseja-se que a função logística forneça resultados semelhantes ao da função ogiva normal.

Note que $P(X_{ij}=1|q_j)$ pode ser vista como a proporção de respostas corretas ao item i, dentre todos os indivíduos da população com habilidade q_j. A relação existente entre $P(X_{ij}=1|q_j)$ e os parâmetros do modelo é apresentada na figura abaixo, chamada de *Curva Característica do Item* (CCI).

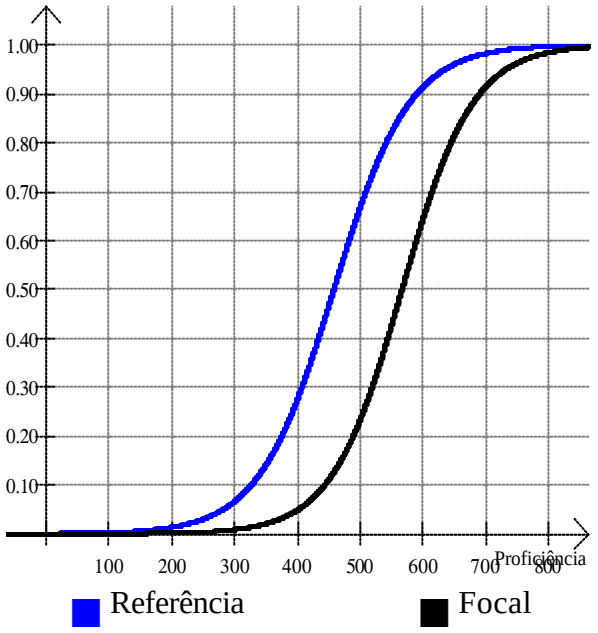
Fig. 2: Modelo Logístico de 3 parâmetros



Como, pelo pressuposto da TRI de que a probabilidade de acerto ao item é função da proficiência do aluno, essa curva tem que ser a mesma, para dois grupos de alunos que tenham a mesma proficiência. Dito de outra maneira, um item apresenta DIF, portanto, se sua CCI não é a mesma para grupos diferentes, no nosso caso, países diferentes.

Existem, basicamente, dois tipos diferentes de DIF. O primeiro é o DIF *uniforme* ou *consistente,* que ocorre quando as CCIs do item estudado para o GR e para o GF são diferentes, indicando que o item favorece uniformemente um dos grupos, em relação ao outro. Em outras palavras, as curvas não se cruzam em nenhum ponto, ao longo da proficiência (q). A figura abaixo mostra um exemplo de item apresentando DIF uniforme.

Fig. 3: Representação de um Item com DIF Uniforme

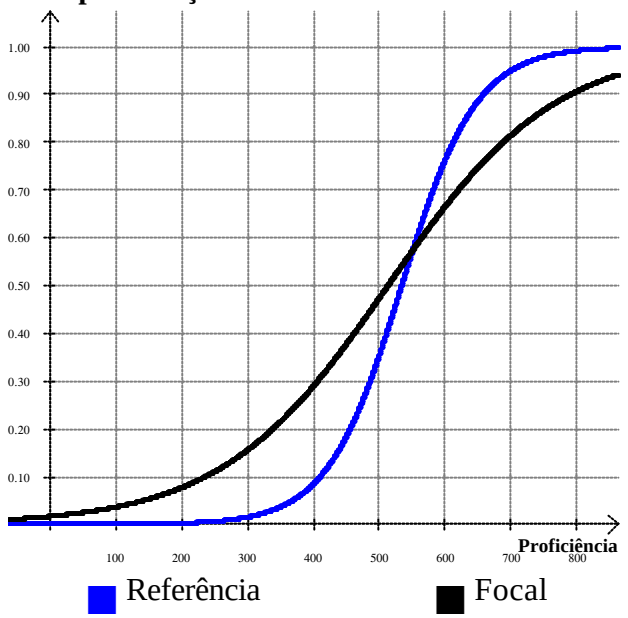


De acordo com a figura 3, observa-se que a CCI do grupo de referência está situada mais à esquerda que a CCI do grupo focal, o que indica que o item é mais fácil para o grupo de referência em todos os níveis de proficiência. Essa diferença indica que o item apresenta DIF, nesse caso, favorável ao grupo de referência. Supondo que as curvas representam dois países, cujos respectivos alunos foram submetidos a um mesmo item, poderíamos afirmar que esse item apresenta DIF no parâmetro ***b***, ou seja, apenas na dificuldade. Isto porque o parâmetro ***c*** é igual a zero para os dois grupos e a inclinação da curva, descrita pelo parâmetro ***a,*** é também a mesma para os dois grupos. De acordo com esse exemplo, alunos com proficiências iguais a 500, nos dois grupos, têm chances

diferentes de acertarem o item. O grupo focal tem 25% e o grupo de referência 65%, o que caracteriza um comportamento anômalo desse item.

O segundo tipo de DIF, denominado DIF *não-uniforme* ou *inconsistente*, ocorre quando há uma interação entre o nível de proficiência e a performance no item, de modo que a direção do DIF muda ao longo da escala de proficiência. Observa-se que as CCIs são diferentes e se cruzam em algum ponto do contínuo da proficiência, como pode ser observado na figura abaixo.

Fig. 4: Representação de um Item com DIF Não-uniforme



Fonte: Relatório Técnico do PISA 2003

Assim, de acordo com esse exemplo, para níveis de proficiências mais baixas, o item favorece o grupo focal. À medida que temos os dois grupos nivelados por proficiências mais altas, o DIF se inverte e passa a favorecer o grupo de referência. Neste segundo tipo de DIF, é inapropriado examinar-se globalmente os dados, porque tal procedimento poderia ocultar sua presença, pois a peculiar variabilidade do DIF, que se verifica em distintas zonas da variável latente (proficiência), pode cancelar, total ou parcialmente, sua detecção (Martinez Arias, 1997, citado em Andriola, 2006).

2.3.4 – Métodos de Investigação de DIF

Existem vários procedimentos formais para se estudarem itens com funcionamento diferencial. De um modo geral, aqueles podem ser divididos em dois grupos: os chamados clássicos, que necessitam de uma proficiência já

conhecida, e os métodos baseados nos modelos da TRI, que não precisam de uma proficiência já conhecida, mas que dependem de alguma hipótese que garanta a comparabilidade dos resultados de proficiência para os grupos analisados, em particular, que exista e seja conhecido, *a priori*, um subconjunto de itens que não possuam DIF. Uma revisão dos métodos clássicos pode ser encontrada, por exemplo, em Andriola, (2002) e Soares *et al*, (2005), Valle, (2002).

Um significativo desenvolvimento nessa área foi alcançado com o artigo de Lord e NovicK (1968), no qual Lord explica o modelo da Teoria de Resposta ao Item (TRI). Logo, ficou claro que esse modelo poderia ser usado, com proveito, no estudo do funcionamento diferencial do item. Como se sabe agora, a base da teoria reside na função da resposta ao item, ou seja, a curva em forma de *S* da proporção de indivíduos de mesmo nível de habilidade, que responde corretamente a um determinado item. Pressupondo que a habilidade considerada seja unidimensional e que o item meça a mesma habilidade, a curva é única sob as condições de um modelo particular: exceto para variações aleatórias, a mesma curva é encontrada, independente da natureza do grupo para o qual a função é plotada. A curva é freqüentemente definida por três parâmetros: ***a***, ***b*** e ***c***, como definidos acima. Devido à natureza única da curva de resposta ao item, sob as condições mencionadas, o fato de a curva de resposta não ser a mesma para dois grupos é a evidência de que os pressupostos não são satisfatórios para um ou ambos os grupos. Podemos, então, pensar em investigar a presença de DIF, comparando os parâmetros que determinam a CCI. Há, ainda, testes para o exame da área entre as curvas, isto é, o grau no qual as curvas não coincidem. Por este método, calcula-se a diferença entre as áreas, sob as duas curvas[5].

Existem várias críticas a esses métodos. Uma delas é que o teste estatístico utilizado na comparação dos parâmetros é assintótico, ou seja, a distribuição da estatística do teste não é exata, o que significa, em linhas gerais, que sua distribuição exata só seria conhecida no caso de grandes amostras. No entanto, pode não ser apropriado o uso de distribuições assintóticas no caso em que os parâmetros de item (os parâmetros **a**, **b** e **c**) e as habilidades (os parâmetros ?) são estimados simultaneamente, ou seja, todos desconhecidos, como é a situação mais freqüente (Soares, 2007).

[5] As expressões exatas para o cálculo dessas diferenças podem ser consultadas em Andriola, 2001.

Outra crítica ao método é que diferenças significativas entre os parâmetros podem ser encontradas quando, na prática, não existem diferenças na faixa de variação da habilidade de mais interesse. A escala de habilidade é arbitrária, mas normalmente, é representada, tendo média 0 e desvio padrão 1. Neste caso, a amplitude de variação de interesse concentra-se de -3 a 3. Considerando os parâmetros das CCI de dois grupos: $(a_1 = 1.8\ ; b_1 = 3.5\ ; c_1 = 0.2)$ e $(a_2 = 0.5\ ; b_2 = 5.0\ ; c_2 = 0.2)$, estudos mostraram que foram encontradas diferenças estatísticas significativas entre os parâmetros das CCI dos dois grupos e, no entanto, as duas curvas diferem menos do que 0.05 na região de variação de habilidade de maior interesse (-3 a 3).

A fim de superar os problemas associados a esses métodos e também a outros derivados da TRI, foram desenvolvidos métodos alternativos que não utilizam técnicas derivadas da TRI, na detecção do DIF, ou seja, não-paramétricos. Para esses métodos, testes estatísticos são difíceis de serem obtidos e, consequentemente, a detecção de DIF pode ser baseada somente em valores de referência, os quais dependem fortemente de resultados empíricos e são, portanto, muito subjetivos.

Dentre os mais conhecidos, estão o procedimento de detecção de DIF através da regressão logística, proposto por Swaminathan e Rogers (1990) e o Método de Mantel-Haenszel (Holland e Thayer, 1988).

Uma crítica ao método de Mantel-Haenszel é que ele não é sensível ao DIF não-uniforme. Este problema motivou a busca por técnicas de detecção do DIF, que superassem essa limitação, como é o caso da regressão logística. No entanto, o método de Mantel-Haenszel ainda é o mais utilizado para a análise do DIF, inclusive pelo *Educational Testing Service*, nos exames do *National Assessment for Educational Progress* (NAEP); e aqui no Brasil, na análise do SAEB (Valle, 2002).

Ambos os métodos foram utilizados nesta tese, razão pela qual passo a explicar mais detalhadamente esses procedimentos.

2.3.4.1 – Regressão Logística

O modelo para predizer a probabilidade de ocorrência de uma resposta correta a um item, mais conhecido como método da regressão logística, tem a seguinte formulação matemática:

$$P(u=1)=\frac{e^{z}}{1+e^{z}}$$

Onde:

- u é a resposta ao item estudado, sendo $z=t_0+t_1q+t_2g+t_3(qg)$

Para:

- t_0 ponto de intersecção da reta de regressão com o eixo das abscissas;
- t_1 inclinação da reta de regressão;
- t_2 diferença entre o rendimento dos grupos no item em foco;
- t_3 parâmetro indicador da possível interação entre q e g;
- q habilidade ou variável latente, medida pelo item;
- g grupo (de referência ou focal) ao qual pertencem os sujeitos;

Para explicar o DIF nos grupos de interesse (de referência e focal), deveremos especificar distintas equações. Assim, um item terá DIF uniforme ou consistente, se $t_2 \neq 0$ e $t_3 = 0$; e terá DIF não-uniforme ou inconsistente, se $t_3 \neq 0$ (seja ou não $t_2 = 0$).

2.3.4.2 – Mantel-Haenszel

Este método foi desenvolvido por N. Mantel e W. Haenszel no ano de 1959, e aplicado ao estudo do DIF, por P. W. Holland e D. T. Thayer em 1988 (Angoff,1993; Dorans&Holland, 1993). Consiste, basicamente, na comparação das freqüências observadas e esperadas de acertos e erros nos diferentes grupos, de acordo com os distintos níveis de habilidades escolhidos pelo investigador. Uma tabela 2 x 2 x K é construída com base na performance (acerto ou erro) dos membros de cada grupo (focal ou referência) e o escore total, no teste. Este procedimento emparelha os grupos em uma medida de desempenho que, nas aplicações usuais, é o escore total no teste. Assim, para cada um dos k níveis da

variável de emparelhamento, o procedimento de MH constrói uma tabela 2 x 2, como a apresentada a seguir.

Tabela 1: Freqüências Observadas de Respostas a um item Hipotético

Grupos	Acertos (1)	Erros (0)	Total
Referência	a_k	b_k	n_{Rk}
Focal	c_k	d_k	n_{Fk}
Total	M_{1k}	M_{0k}	T_k

Baseados nesta lógica, N. Mantel e W. Haenszel propuseram a seguinte fórmula para a comparação das freqüências, que corresponde à razão de chances, estimada para comparar os dois grupos em termos da sua chance de responder ao item corretamente, condicionada ao escore total no teste. O procedimento MH utiliza a hipótese de que essa razão de chances α é constante em todos os k níveis da variável de emparelhamento.

$$\hat{\alpha}_{MH} = \frac{\sum_{k=1}^{k} \frac{a_k d_k}{T_k}}{\sum_{k=1}^{k} \frac{b_k c_k}{T_k}}$$

Onde:

- a_k é a freqüência observada das respostas corretas do grupo de referência, nos distintos níveis de pontuação escolhidos;

- b_k é a freqüência observada das respostas incorretas do grupo de referência, nos distintos níveis de pontuação escolhidos;

- c_k é a freqüência observada das respostas corretas do grupo focal nos distintos níveis de pontuação escolhidos;

- d_k é a freqüência observada das respostas incorretas do grupo focal nos distintos níveis de pontuação escolhidos;

- T_k é o total de erros e acertos, de cada grupo, nos níveis de pontuação escolhidos;

Assim, para um dado escore k, n_{Rk} e n_{Fk} são, respectivamente, o número de alunos nos grupos de referência, e focal, M_{1k} é o número de alunos que acertou o item, M_{0k} é o número de alunos que errou o item.

Em geral, utiliza-se uma transformação dessa estatística, dada por:

$$\hat{\Delta}_{MH} = -2.35 \log_e (\hat{\alpha}_{MH})$$

Por convenção, essa transformação é definida de modo que tal medida assume valores negativos, quando o item é mais difícil (condicionalmente aos valores da variável de emparelhamento) para o grupo focal.

Para uma melhor interpretação dos valores de $\hat{\Delta}_{MH}$ que serão apresentados nos exemplos, devemos lembrar que $\hat{\alpha}_{MH}$ é uma estatística cujo numerador está relacionado aos acertos do grupo de referência e erros do grupo focal, e cujo denominador ao complementar dessas quantidades, ou seja, aos erros do grupo de referência e acertos do grupo focal. Dessa maneira, essa estatística fornece uma medida do desempenho de grupo de referência com relação ao grupo focal. Valores de $\hat{\alpha}_{MH}$ iguais a 1 significam que os dois grupos tiveram o mesmo desempenho no item; valores de $\hat{\alpha}_{MH} > 1$ indicam um melhor desempenho do grupo de referência, em relação ao grupo focal; e, por fim, valores de $\hat{\alpha}_{MH}$ entre 0 e 1 indicam um melhor desempenho do grupo focal. Portanto, pela definição matemática, teremos um valor negativo para $\hat{\Delta}_{MH}$, quando o desempenho do grupo de referência no item é melhor do que o do grupo focal; e $\hat{\Delta}_{MH}$ positivo no caso contrário.

Alem do sinal de $\hat{\Delta}_{MH}$, também existe uma relação direta entre sua magnitude e a diferença entre o desempenho dos dois grupos, de maneira que, quanto maiores essas diferenças, maior o valor de $\hat{\Delta}_{MH}$. O quadro abaixo mostra a classificação considerada nesta tese para os valores encontrados, com relação aos itens que apresentaram DIF:

Tabela 2: Classificação da Magnitude do Funcionamento Diferencial

Valores da estatística de Alfa (delta) deMantel Haenzel	Magnitude do DIF
$\lvert\hat{\Delta}_{MH}\rvert \leq 0{,}5$	DIF insignificante
$0{,}5 < \lvert\hat{\Delta}_{MH}\rvert \leq 1{,}0$	DIF pequeno
$1{,}0 < \lvert\hat{\Delta}_{MH}\rvert \leq 1{,}5$	DIF intermediário
$\lvert\hat{\Delta}_{MH}\rvert > 1{,}5$	DIF alto

2.3.4.3 – Abordagem Integrada

Esse tipo de abordagem procura detectar, dimensionar e explicar o DIF, simultaneamente, evitando-se utilizar processos em estágios múltiplos e fragmentados.

A abordagem que utilizo neste estudo foi proposta por Soares *et al*, (2007), que apresenta uma modelagem integrada para o problema, a qual permite detectar, estimar os parâmetros de DIF e explicar o DIF encontrado, se não, em uma única análise, no máximo, em duas etapas. Considere a extensão do modelo de três parâmetros da TRI, que incorpora a possibilidade de DIF[6]. Para uma abordagem integrada, que considere simultaneamente o problema de detecção, estimação e explicação do DIF, seja $Z_{ig}^{h}, (h=a,b),$ uma variável indicadora tal que $Z_{ig}^{h}=1,$ se $i \in I_{d}^{h,g}$, e, $Z_{ig}^{h}=0$ caso contrário. Isto é, Z_{ig}^{h} é uma variável que indica se o item é um item sem DIF no grupo g, ou não; se ele pode ter DIF. Todos os trabalhos anteriores consideraram situações onde Z_{ig}^{h} é conhecida, *a priori* (ver a discussão sobre os itens âncoras em Soares,2007). Essa novidade foi introduzida em Soares, Gonçalves e Gamerman, 2007. Para explicar o DIF, complementa-se o modelo, admitindo-se, ainda, que:

$$d_{ig}^{h} = (g_{0g}^{h} + \sum_{k=1}^{K^h} g_{kg}^{h}\, W_{ik}^{h} + h_{ig}^{h})\, Z_{ig}^{h}$$

Onde g_{kg}^{h} são parâmetros fixos do modelo de explicação do DIF, W_{ik}^{h} são as variáveis explicativas, associadas aos itens; e h_{ig}^{h} é o fator específico para o DIF,

[6] Para maiores detalhes consultar Soares (2007).

apresentado pelo item em cada grupo. Assume-se, também, a hipótese de normalidade para o fator específico, isto é, que $h_{ig}^{h} \sim N(0, \left(t_{g}^{h}\right)^{2})$. Assim, se $Z_{ig}^{h}=1$, então $d_{ig}^{h} \sim N\left((g_{0g}^{h} + \sum_{k=1}^{K^h} g_{kg}^{h} \; W_{ik}^{h}) , \left(t_{g}^{h}\right)^{2} \right)$.

Note-se que a estrutura de regressão é imposta para todos os itens exceto os itens âncoras. Para estes deve-se, a princípio, assegurar que $d_{ig}^{h} = 0$. De fato, uma solução menos extrema e que facilita alguns aspectos de modelagem e outros aspectos computacionais consiste em admitir que, quando $Z_{ig}^{h}=0$, $d_{ig}^{h} \sim N(0, s^2 \left(t_{g}^{h}\right)^{2})$. Assim, se o valor de s for pequeno, o mesmo ocorrerá com o valor de d_{ig}^{h}, de tal forma que, para fins práticos, pode-se garantir que $d_{ig}^{h} \approx 0$. A partir dessa consideração, tem-se, então, que:

$$d_{ig}^{h} \sim N\left((g_{0g}^{h} + \sum_{k=1}^{K^h} g_{kg}^{h} \; W_{ik}^{h}) \; Z_{ig}^{h} \; , \left[s^2 \right]^{1-Z_{ig}^{h}} t_{g}^{h} \right)$$

Como o objetivo do trabalho é apresentar uma análise Bayesiana para o problema do DIF, o modelo se completa com a especificação das distribuições *a priori* para os parâmetros. As prioris adotadas para os parâmetros estruturais foram: $a_i \sim LN(0,2)$, $b_i \sim N(0,1)$ e $c_i \sim beta(5,17)$. Estas prioris são habitualmente empregadas, como por exemplo, são *defaults* no *software Bilog-mg*, e naturais, tendo-se em vista as características dos parâmetros. Para os parâmetros correspondentes ao modelo de explicação do DIF, admite-se que $\mathbf{d}_{g}^{h} \mid \mathbf{W}^{h}, \gamma_{g}^{h}, t_{g}^{h}\,\mathbf{I} \sim N(\mathbf{W}^{h}\,\gamma_{g}^{h}, t_{g}^{h}\,\mathbf{I})$, com a priori $g_{g}^{h} \sim N(g_{0}^{h}, S_{0}^{h})$. Admite-se, ainda, que $\mu_g \mid \sigma_g \sim N(0, \sigma_g)$, com $\sigma_g^2 \sim GI(\alpha_g, \beta_g)$, onde *GI* representa a distribuição Gama Inversa.

Cabe ressaltar que esta abordagem integrada, pelo menos em teoria, deve ser mais precisa que a abordagem tradicional, isso porque a proficiência produzida é naturalmente purificada, em função do DIF (Soares,2007).

2.3.5 – Alguns Estudos de DIF

A questão da justeza de itens e testes padronizados tem acompanhado as atividades dos psicometristas, desde muito tempo, apesar de haver crescido consideravelmente a partir dos anos 1970. A evidência de DIF tem sido usada para identificar itens com *viés* e as pesquisas sobre o tema é uma resposta às preocupações de todos os envolvidos com os procedimentos de estimativa das medidas de proficiência. O´Neill e McPeek (1993) investigaram a relação entre o DIF e as formas do item e do teste, buscando identificar características que possam ocasionar diferenças significativas em resultados de testes. O estudo por eles realizado baseia-se nas análises *post hoc* de testes existentes, reunidos antes dos procedimentos Mantel-Haenszel (M-H) terem sido estabelecidos entre os especialistas.

Inicialmente, são apresentados os resultados de estudos nos testes de leitura (*verbal tests)*. Examinando os testes que incluíam textos sobre uma variedade de tópicos, tais como humanidades, ciências sociais, ciência biológica, etc., concluíram que, quando mulheres são comparadas a homens, elas têm desempenho tipicamente pior que eles em compreensão de leitura nos itens relacionados a conteúdo científico. Em alguns testes do GRE – General Test - itens baseados em textos científicos são mais difíceis para mulheres que para homens; e itens baseados em ciências sociais e humanidades são, geralmente, mais fáceis para mulheres que para homens (Scheuneman & Gerritz, 1990 ; Wild & McPeek, 1986). A pesquisa de textos de leitura do SAT revelou que o conteúdo relacionado a aspectos técnicos da ciência (o oposto acontece quanto à história ou à filosofia) é mais difícil para mulheres que para homens (Lawrence & Curley,1989; Lawrence, Curley & McHale,1988). A diferença é encontrada a despeito de toda a informação necessária estar presente no item.

Questões de compreensão em leitura, que se referem ao universo masculino e feminino (opostas àquelas que se referem apenas a homens ou que não mencionam pessoas) mostram melhor desempenho de mulheres que de homens. Este resultado foi encontrado em análises do GMAT, SAT e NTE[7], mas o efeito não foi significativo no GRE (Carlton & Harris, 1989b ; O´Neill, McPeek

[7] GMAT - Graduate Management Admission Test . GRE Test Preparation Practice Exercises . SAT - is a registered trademark of the College Entrance Examination Board . NTE – National Teacher Examination.

& Wild). O resultado pode ser análogo, quando se trata do desempenho superior de examinandos de minorias, comparados a examinandos brancos, em textos orientados para questões das minorias: um tema que seja de maior interesse para um grupo particular é mais fácil para este grupo que para o grupo de referência.

Considerando o conteúdo de leitura do item, as causas particulares de DIF para homens e mulheres geralmente se mantêm evasivas. Poucos itens que apresentam uma quantidade significativa de funcionamento diferencial podem ser explicados com facilidade. Palavras pertencentes ao repertório tipicamente masculino, como caça ou *hockey* no gelo, por exemplo, freqüentemente apresentam maior dificuldade para mulheres que para homens. Do mesmo modo, palavras relacionadas ao estereótipo do universo feminino, como costura ou bordado, freqüentemente, são mais fáceis para mulheres que para homens. Apesar dos exemplos de conteúdo estereotipado é possível encontrar alguns poucos contra-exemplos de tal generalização. Por exemplo, um item cujo conteúdo relaciona-se à guerra apresentou desempenho praticamente idêntico de homens e mulheres com mesma habilidade em linguagem.

Para a maioria de itens que apresenta uma quantidade substantiva de DIF, é possível pensar que este resulte do acúmulo de efeitos de várias características individuais dos itens. Essas características, não identificadas separadamente, podem ter apenas pequenos efeitos, em geral, não aparentes. Além disso, as características do item que causam DIF só o fazem, quando certas combinações delas estão presentes. Em decorrência, não é difícil entender o insucesso na elaboração de hipóteses sobre as causas de DIF, com relação à maioria de itens que apresenta DIF extremo.

Diferentemente dos resultados de mulheres, com relação a itens de conteúdo científico, a pesquisa de DIF, relacionado à raça e conteúdo de textos, apresenta um quadro misto. No GRE, negros geralmente têm desempenho pior, em questões associadas a brancos. No SAT e NTE, não houve relação entre DIF e conteúdo de textos para examinandos negros. Os resultados do GRE podem estar relacionados a características particulares de formatos dos testes estudados, ou dos examinandos do GRE.

O estudo mostra um claro desempenho superior de negros e hispânicos, comparados a brancos, em itens baseados em textos relevantes, relacionados a interesses e preocupações dessas minorias. O resultado aparece no GRE, GMAT,

NTE e SAT (Carlton & Harris, 1989b ; McPeek & Wild, 1986; Medley & Quirk,1974 ; Wild & McPeek,1986). No SAT, negros têm desempenho acima dos brancos, em questões que se referem aos afro-americanos (Carlton & Harris, 1989a) e hispânicos têm melhor desempenho que brancos, em questões que se referem a mulheres de origem mexicana, e em questões que contêm referências a um matemático negro (Schmitt,1986). Tudo indica que, em assuntos de especial interesse de minorias, o grupo em questão tem melhor desempenho que o grupo de referência, que no caso destes estudos são os estudantes brancos. Possivelmente, as explicações para o desempenho superior das minorias nestes itens estejam no maior interesse étnico, no conhecimento e na autoconfiança, com relação ao assunto tratado.

A revisão prossegue, analisando as características de conteúdo de itens de Matemática associados ao DIF, para mulheres. O conteúdo matemático também se associa à dificuldade diferencial no desempenho das mulheres que, em Álgebra, saem-se melhor que os homens, o que se evidencia em escores de testes dessa área. Os homens têm melhor desempenho em itens de Geometria e de resolução de problemas. Esse resultado, encontrado em vários testes de seleção (admissão) ETS (Carlton & Harris, 1989b ; O´Neill, Wild &McPeek,1989) e em exames ACT (Doolittle,1989), pode refletir a natureza e a quantidade de trabalhos feitos pelas mulheres em seus cursos, as atitudes que elas têm com relação à Matemática, ou suas experiências em atividades extracurriculares.

Por conseguinte, itens que utilizam símbolos são mais fáceis para mulheres que para homens, porque o conteúdo da Álgebra, freqüentemente, requer símbolos para representarem quantidades matemáticas (McPeek & Wild,1987 ; O´Neill, Wild & McPeek , 1989).

Outra diferença no desempenho em itens numéricos observa-se, com relação ao modo de apresentação de um problema – os itens podem ser problemas contextualizados (*word problems*), limitados em termos de uma situação de fato, ou são problemas relacionados a uma apresentação puramente matemática, como os que envolvem fórmula, equação ou teoria. Em geral, homens têm melhor desempenho do que as mulheres em problemas contextualizados e mulheres superam os homens, em itens de Matemática pura. Essa diferença de performance em problemas contextualizados (*word problems*) foi encontrada em itens do GRE e GMAT (O´Neill, Wild & McPeek , 1989), itens SAT (Carlton & Harris,1989b)

e itens ACT (Doolittle & Cleary,1987). Embora a explicação de tais resultados não seja óbvia, uma área de exploração é sugerida por um item característico, estudado no teste quantitativo SAT, qual seja, a que diz respeito à relação entre item e currículo, identificando itens mais ou menos similares a livros-texto e deveres de casa. Para essa variável, o grupo focal (mulheres ou minorias) tem desempenho que supera o do grupo de referência (homens ou brancos), quando o item é muito semelhante a problemas encontrados em livros-texto, mas o contrário não é verdadeiro. Nesses testes, antes de estabelecer a correspondência entre homens e mulheres, visando à análise do DIF, o escore quantitativo médio das mulheres é menor que o dos homens, e seu escore médio de linguagem é, aproximadamente, igual ao deles. Devido à correlação fortemente positiva entre escores de linguagem e numéricos, os grupos de mulheres, em cada nível do escore, geralmente têm escores de linguagem mais altos do que os dos homens, que apresentam o mesmo escore, em itens numéricos. Isso faz com que seja difícil acreditar, consistentemente, que as mulheres têm desempenho pior que os homens, em problemas numéricos contextualizados (w*ord problems*). Esses tipos de itens que tratam de problemas da vida cotidiana não se assemelham aos problemas de livros-texto, e podem envolver *insights* e soluções inovadoras. É necessário que se pesquise mais, para se determinar que aspectos desse(s) tipo(s) de problema(s) são especialmente difíceis para as mulheres.

No que diz respeito às características de conteúdo de itens numéricos associados ao DIF para minorias, um resultado encontrado foi de que negros e hispânicos tendem a apresentar melhores resultados em itens de Álgebra que examinandos brancos, em teste de seleção (admissão), embora isso não aconteça no SAT (O´Neill, Wild & McPeek).

Em alguns estudos de DIF, características de formato de itens de linguagem e itens numéricos também foram analisadas. Nesses casos, mais do que o conteúdo do item, também, foram consideradas as características relacionadas a como o conteúdo foi apresentado. Dentre os resultados, destaco dois: o formato das respostas e o uso de recursos visuais (gráficos, imagens, etc.)

Com relação ao primeiro, um conjunto de respostas de itens pode ser apresentado em uma lista vertical ou em uma série de palavras (horizontal). No formato vertical, cada opção de resposta fica numa linha separada, já no formato horizontal, as opções de resposta estão uma ao lado da outra, sem mudança de

linha indicativa da nova opção. A decisão sobre o formato a ser usado para um item específico é tomada pelos elaboradores de teste, especializados em diagramação de página. Há alguma evidência de que, em itens de conteúdo análogo, examinandos brancos podem se sair melhor que negros, hispânicos e asiáticos, quando se usa o formato horizontal. Como outros itens de linguagem do SAT e itens numéricos não evidenciaram essa diferença, é preciso que outras pesquisas, sobre este aspecto, sejam realizadas.

A segunda característica de formato identifica itens de Matemática que usam material visual, como gráficos, quadros ou diagramas. Em tipos de item quantitativo, como os de interpretação de dados do GRE, todos os estímulos podem ser apresentados em gráficos ou tabelas; em outros tipos, como os itens de Matemática simples (SAT) ou de resolução de problemas (GRE), itens com material visual geralmente são acompanhados de algum material escrito. Em ambos os tipos, examinandos negros têm desempenho pior que os brancos. Examinandos hispânicos têm desempenho pior que os brancos, em itens de interpretação de dados, mas não em itens SAT, com figuras ou gráficos. Em análises de desempenho de negros e brancos, consistente com os resultados de examinandos negros, há uma forte correlação da dificuldade com o DIF, no que diz respeito a itens de interpretação de dados (O´Neill, Wild & McPeek). Resultados similares, indicando pior desempenho de negros que de brancos em itens com material visual, foram observados, também, em itens de ciências sociais que usam mapas e quadros.

Nas análises de DIF relacionados a gênero, há uma correlação moderada entre o DIF e a dificuldade de itens numéricos que usam material visual como estímulo, em itens do GRE, mas não em outros testes, como GMAT. Mulheres tendem a ter um desempenho melhor do que os homens, em itens difíceis, e, ao contrário, pior desempenho em itens mais fáceis. Na medida em que o material visual se relaciona a testes para avaliar êxito em Matemática, pesquisadores como Benbow e Stanley (1980) formularam uma hipótese, segundo a qual, a diferença de gênero em habilidade espacial pode contribuir para o diferencial do desempenho. Contudo, Linn e Hyde (1989) afirmam que as metas-análise sobre diferenças de gênero não fornecem evidência para tal hipótese.

Snetzler e Qualls (2000) examinaram a incidência de DIF em três subtestes do Iowa Tests of Basic Skills. Na pesquisa *"Exame do DIF em bateria*

de resultados padronizados de estudantes com proficiência limitada em inglês", estudantes do Alaska, classificados como "nativos" e "brancos", foram avaliados em dois momentos: 4º e 5º graus ou 6º e 7º graus. A análise da proficiência em linguagem foi orientada, exclusivamente, para os nativos, classificados, segundo as três categorias aceitas para a proficiência em linguagem: limitada, bilíngüe, proficiente. Embora um decréscimo consistente na mudança do efeito-tamanho tenha sido observado ao longo do tempo, na comparação entre bilíngües e proficientes, a incidência do DIF indicou tendência a ser esporádica, em relação ao grupo favorecido. Reconhecendo a ligação inseparável entre o background da linguagem e étnico-cultural, uma análise adicional do DIF, comparando "brancos" e "nativos" de igual nível de proficiência, foi levada a termo, visando ao entendimento da comparação da linguagem. Diferenças do efeito-tamanho, favorecendo os "brancos", foram maiores que aquelas observadas nas comparações de linguagem.

Berberoglu (1995), em estudo intitulado *"Análise do DIF em Questões de Computação, Problemas Contextualizados e Geometria"*, para grupos categorizados por gênero e NSE, discute que a avaliação de resultados em Matemática, considerando gênero e NSE, tem sido preocupação de muitos pesquisadores. Com relação a desempenho específico e áreas de conteúdo, parece ser significativa a diferença entre gêneros. Verifica-se, por exemplo, a superioridade masculina em problemas que lidam com medidas, probabilidade e componentes espaciais; a superioridade feminina aparece em itens de computação, Álgebra e análise de relações simbólicas. Geometria e raciocínio matemático são mais difíceis para mulheres, enquanto elas têm mais facilidade em algoritmos e computação. No entanto, Ellington (1990) relata não haver diferença significativa em qualquer área de conteúdo nos dados do II Estudo Internacional de Matemática, realizado em 24 países. Parece que diferenças de gênero e desempenho em conteúdos diversos é um debate em aberto e as fontes ou razões para os estudos apresentados pela literatura da área padecem da falta de clareza em suas explanações.

Por outro lado, efeitos de NSE em resultados de Matemática foram consistentemente verificados, em vários projetos de pesquisa. A maioria dos estudos de gênero e NSE depende de técnicas nas quais as diferenças médias nos escores do teste ou do item são comparadas entre grupos, sem levar em conta os

níveis de habilidade. Assim, os resultados podem ser reflexos de diferenças de habilidade, em geral. A TRI pode oferecer resposta para essa dificuldade, uma vez que possibilita a interpretação por item, e não para o teste, como um todo. Neste estudo, as CCIs são comparadas entre gêneros e NSE, evidenciando se um dos grupos tem vantagem, em resolução de questões de Matemática, em conteúdos de computação, problemas contextualizados e Geometria. Os dados são provenientes do subteste da UEE, aplicado na Turquia. Além disso, grupos de gênero foram comparados dentro de grupos de NSE, visando a determinar se os índices DIF, obtidos por grupos de homens e mulheres, são independentes do NSE. Mais especificamente, o estudo examinou as direções e a magnitude do DIF, em itens de computação, problemas contextualizados e Geometria, para grupos de homens e de mulheres e para grupos de alto NSE e de baixo NSE. Além disso, determinou se os resultados do DIF observados por gênero são independentes do NSE, para cada tipo de item, comparando as CCIs de: grupos de homens e de mulheres, com elevado NSE e grupos de homens e mulheres com baixo NSE.

No Brasil, Soares (2005), a partir dos dados do Programa de Avaliação da Rede Pública de Educação Básica – Proeb - analisa o funcionamento diferencial - DIF - dos itens de geografia, aplicados aos alunos da 4ª série do Ensino Fundamental, nas diferentes regiões do Estado de Minas Gerais. O estudo apontou diferenças de competência, especialmente, com relação a itens que tratavam das diferenças entre o espaço urbano e rural, que se mostraram desfavoráveis para os alunos da região Metropolitana do Estado. Itens associados ao meio ambiente, também apresentaram DIF, sendo mais difíceis para os alunos do interior do que os da região metropolitana.

Esses resultados ratificam a relevância das análises de funcionamento diferencial do item, na identificação de diferenças em testes de habilidades. Essas análises nos sugerem que, ao invés de entendermos o item de teste como a única causa do desempenho diferencial, devemos considerar, também, questões de eqüidade educacional em nossas escolas e em nossa sociedade. Sabemos que nem todas as escolas oferecem as mesmas oportunidades a seus alunos. Inúmeras pesquisas na área de educação apontam para as diferenças entre e intra-escolar. Há diferenças de clientela, de gestão, de equipamentos e de cursos oferecidos, entre outras. O perfeito entendimento dos resultados de DIF passa, necessariamente, pelo reconhecimento dessas iniqüidades educacionais.