http://dx.doi.org/10.18222/eae.v28i68.4584

FUNCIONAMENTO DIFERENCIAL DOS ITENS DE CIÊNCIAS DO PISA: BRASIL E JAPÃO

ANDRIELE FERREIRA MURI
TUFI MACHADO SOARES
ALICIA BONAMINO

RESUMO

*O Programa Internacional de Avaliação de Estudantes (PISA) é uma avaliação comparada, aplicada a uma amostra de estudantes de 15 anos de idade. Juntamente com vários outros países, Brasil e Japão participam desde a primeira edição, em 2000. Com o objetivo de identificar fatores capazes de explicar as diferenças de resultados encontradas no Letramento em Ciências, entre alunos brasileiros e japoneses, na edição de 2006, foi utilizada a análise de DIF (*Differential Item Functioning*), que possibilitou extrair, dos resultados dos testes, padrões de efeitos diferenciados. Para identificar os itens que apresentaram funcionamento diferencial entre Brasil e Japão, empregou-se o modelo bayesiano integrado que, além de confirmar a ocorrência, também pode explicar o DIF. Encontramos DIF em todas as covariáveis elegidas, embora nem sempre esse comportamento diferencial tenha privilegiado um dos dois países. Há competências que discriminam mais os alunos brasileiros e áreas de aplicação dos itens ora mais fáceis para o Brasil, ora para o Japão.*

PALAVRAS-CHAVE DIFFERENTIAL ITEM FUNCTIONING – DIFF • PISA • LETRAMENTO CIENTÍFICO • CIÊNCIAS.

FUNCIONAMENTO DIFERENCIAL DE LOS ÍTEMS DE CIENCIAS DEL PISA: BRASIL Y JAPÓN

RESUMEN

El Programa Internacional de Evaluación de Estudiantes (PISA) es una evaluación comparada, aplicada a una muestra de estudiantes de 15 años de edad. Junto con varios otros países, Brasil y Japón participan desde la primera edición, en el 2000. Con el objetivo de identificar factores capaces de explicar las diferencias de resultados encontradas en el Letramento en Ciencias entre alumnos brasileños y japoneses en la edición de 2006, se utilizó el análisis de DIF (Differential Item Functioning), que hizo posible que se extrajera de los resultados de las pruebas patrones de efectos diferenciados. Para identificar los ítems que presentaron funcionamiento diferencial entre Brasil y Japón se utilizó el modelo bayesiano integrado que, además de confirmar la ocurrencia, también puede explicar el DIF. Encontramos DIF en todas las covariables elegidas, aunque no siempre este comportamiento diferencial privilegie uno de los dos países. Hay competencias que discriminan más a los alumnos brasileños y a las áreas de aplicación de los ítems, que a veces son más fáciles para Brasil y otras para Japón.

PALABRAS CLAVE DIFFERENTIAL ITEM FUNCTIONING – DIFF • PISA • LETRAMENTO CIENTÍFICO • CIENCIAS.

DIFFERENTIAL FUNCTIONING OF PISA SCIENCE ITEMS IN BRAZIL AND JAPAN

ABSTRACT

The Programme for International Student Assessment (PISA) is an international comparative assessment program applied to samples of 15-year-old students. Together with other countries, Brazil and Japan have participated in this program since its first edition in 2000. DIF (Differential Item Functioning) analysis was used to identify factors that could explain performance differences in scientific literacy between Brazilian and Japanese students, in the 2006 edition. Based on the test results, this analysis showed patterns of differentiated effects. To identify the items that showed differential functioning between Brazil and Japan, we used the Bayesian integrated model. In addition to confirming this occurrence, this model may also explain the DIF. DIF was found in all covariates selected. However, differential functioning did not always favor either of the two countries. There are competencies that discriminate more against the Brazilian students and areas of application of the items that sometimes were easier for Brazilian and sometimes for Japanese students.

KEYWORDS DIFFERENTIAL ITEM FUNCTIONING – DIFF • PISA • SCIENTIFIC LITERACY • SCIENCES.

INTRODUÇÃO

O PISA – *Programme for International Student Assessment* – que em português foi traduzido como Programa Internacional de Avaliação de Estudantes, é um programa internacional de avaliação comparada, aplicado a uma amostra de estudantes de 15 anos de idade. Para o PISA, essa é a idade em que se pressupõe o término da escolaridade básica obrigatória na maioria dos países. Esse programa é desenvolvido e coordenado internacionalmente pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE) e, no Brasil, pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). As avaliações do PISA acontecem a cada três anos e abrangem três áreas do conhecimento – Leitura, Matemática e Ciências – havendo, a cada edição do Programa, maior ênfase em uma dessas áreas.

Desde o primeiro ciclo de avaliação, realizado em 2000, em função do desempenho insatisfatório dos alunos brasileiros, a divulgação dos resultados tem como foco as conclusões enfáticas de que, em termos educacionais, o Brasil não apresenta um bom nível de proficiência nas diferentes áreas

avaliadas pelo Programa. Estudos comparativos de sistemas educacionais, no entanto, não devem se limitar apenas a medir e comparar os resultados educacionais brutos dos alunos, precisando recorrer a metodologias que possibilitem identificar os principais fatores capazes de explicar as diferenças de rendimento encontradas e analisar o modo como esses fatores interagem entre si. (FERRER, 2003).

Analisamos aqui os itens do PISA 2006, já que, nessa edição do Programa, a área de Ciências foi avaliada mais detalhadamente. Ou seja, além da escala global, foi possível verificar o desempenho dos estudantes também nas competências de "identificar questões científicas", "explicar fenômenos cientificamente" e "usar evidência científica". A ênfase sobre essa área do conhecimento foi novamente objeto do Programa em 2015, mas os resultados haviam sido recém-publicados no momento da elaboração deste artigo.

As edições do PISA de 2006 e 2015 avaliaram, portanto, com maior ênfase o denominado Letramento Científico dos alunos participantes, descrevendo-o como

> [...] a capacidade de empregar o conhecimento científico para identificar questões, adquirir novos conhecimentos, explicar fenômenos científicos e tirar conclusões baseadas em evidências sobre questões científicas. Também faz parte do conceito de Letramento Científico a compreensão das características que diferenciam a ciência como uma forma de conhecimento e investigação; a consciência de como a ciência e a tecnologia moldam nosso meio material, cultural e intelectual; e o interesse em engajar-se em questões científicas, como cidadão crítico capaz de compreender e tomar decisões sobre o mundo natural e as mudanças nele ocorridas. (BRASIL, 2008, p. 34)

Nosso objetivo foi identificar fatores capazes de explicar as diferenças de rendimentos encontradas no Letramento Científico, entre alunos brasileiros e japoneses. O Japão foi escolhido tanto em razão de uma experiência vivida num programa de treinamento de professores oferecido por esse país, entre 2007 e 2009, como em virtude da sua posição de destaque nos testes comparativos internacionais.

Considerando que os sistemas são diferentes e que as características que os distinguem têm consequências nos diversos modos de elaboração e desenvolvimento do currículo e, ainda, que os conteúdos são selecionados pelos professores e abordados com ênfases diferenciadas, procuramos identificar as características dos itens do teste, em relação a competências, áreas do conhecimento, áreas de aplicação, âmbito, tipo e idioma que, por exemplo, sinalizam a existência de ênfases curriculares diferenciadas nesses dois países. Sendo a análise de DIF (*Differential Item Functioning*) uma ferramenta estatística que possibilita extrair dos resultados dos testes esses padrões de efeitos diferenciados, a análise consistiu na aplicação de métodos para detectar e identificar os itens de Ciências que apresentaram funcionamento diferencial entre Brasil e Japão.

A necessária e relevante padronização ou uniformização das condições de aplicação dos instrumentos de medida é um dos pressupostos mais importantes da avaliação, seja no âmbito psicológico, seja no educativo (ANASTASI, 1988; PASQUALI, 2000). O estudo do DIF está intimamente ligado ao suposto da padronização das condições de aplicação dos instrumentos de medida de um teste avaliativo. "Deve-se ter claro que a presença de DIF num teste é um fator que pode tornar o processo avaliativo injusto" (ANDRIOLA, 2001).

A comparação de resultados nos testes educacionais, entendidos como o resultado dos escores que medem a proficiência dos alunos, é possível graças à utilização da Teoria da Resposta ao Item (TRI). Tal comparabilidade decorre do fato de a TRI utilizar modelos estatísticos em que a dificuldade dos itens é parametrizada na mesma escala de proficiência das habilidades cognitivas dos alunos. Além disso, é necessário empregar itens comuns aos diferentes testes e esses itens devem apresentar o mesmo funcionamento nos diversos grupos de alunos para que uma boa comparabilidade seja alcançada. Em avaliações de larga escala como o PISA, essa comparabilidade é muito mais crítica, tendo em vista que nem todos os itens mostram o mesmo funcionamento.

A TRI é composta de um conjunto de modelos matemáticos que representam, grosso modo, a probabilidade de determinada resposta a um item ser escolhida em função

dos parâmetros que o caracterizam e do nível do respondente quanto ao traço latente[1] que está sendo medido. De acordo com Soares (2005), para possibilitar a comparabilidade dos resultados, é essencial que o modelo utilizado na avaliação garanta o pressuposto de que o item apresente o mesmo funcionamento para os diversos grupos populacionais que estão sendo avaliados. Para uma boa comparação entre resultados de grupos tão diferentes, como é o caso de alunos brasileiros e japoneses, é imprescindível, por exemplo, uma atenção especial à construção dos itens, a fim de que estes não apresentem o funcionamento diferencial.

1 A competência cognitiva dos alunos e, neste estudo, a proficiência em Ciências no PISA.

O FUNCIONAMENTO DIFERENCIAL DO ITEM – DIF

Um item de múltipla escolha, ou dicotômico, apresenta DIF quando alunos que possuem a mesma habilidade cognitiva não têm igual probabilidade de acertá-lo. Assim, na estimação das proficiências, o ideal é evitar a utilização de itens com DIF elevado, isto é, que favoreçam demasiadamente um determinado grupo de alunos. Todavia, o DIF, quando moderado e restrito a poucos itens, interfere minimamente na estimação da proficiência e sua análise pode ser uma ferramenta de diagnóstico do sistema educacional bastante útil no que se refere às diferenças curriculares, socioculturais e, no caso de estudos internacionais, como o PISA, à diversidade de realidades educacionais e à disparidade de resultado entre países. Por esse motivo principal, justifica-se a escolha do emprego dessa metodologia para conduzir um estudo comparativo entre dois países de realidades socioeconômicas e culturais tão distintas, como é o caso de Brasil e Japão.

Estudos visando a identificar itens que sejam favoráveis a determinado grupo, em detrimento de outros, ganham destaque no campo da psicometria, pois ajudam a assegurar que os testes sejam tão imparciais quanto possível (AGUIAR, 2008). Nesse sentido, Soares, Genovez e Galvão (2005) destacam que a preocupação com o funcionamento diferencial do item antecede ou, ainda, extrapola o contexto da TRI, na qual a ausência do DIF é requisito para uma boa equalização entre resultados de grupos diferentes de alunos.

ESTUDOS SOBRE DIF

Historicamente, a preocupação com o DIF está fortemente associada ao desejo de se construírem questões de teste que não sejam afetadas por características étnico-culturais dos grupos submetidos aos testes de avaliação educacional (COLE, 1993). A partir de achados em estudos sobre o viés[2] de itens e testes realizados em 1951 por pesquisadores da Universidade de Chicago, que haviam encontrado variações nos itens em aspectos peculiares, tais como conteúdo e formato, surgem os primeiros dados a respeito dos problemas técnicos presentes em determinados itens utilizados na avaliação da aprendizagem (HAMBLETON; SWAMINATHAN; ROGERS, 1991). Segundo Aguiar (2008), um desses problemas técnicos diz respeito ao uso indevido da linguagem escrita; muitas vezes, termos empregados nos testes são mais familiares a determinado grupo em detrimento de outro. O'Neil e McPeek (1993), Schmitt e Bleistein (1987), Berberoglu (1995) e Gierl *et al.* (2003) mostraram que as diferenças entre os grupos podem também estar relacionadas às características étnicas, de sexo, de nível socioeconômico, entre outras.

2 Um item é enviesado se sujeitos de habilidades iguais, mas de culturas diferentes, não têm a mesma probabilidade de acertar o item (LINN; DRASGOW, 1993).

Soares, Genovez e Galvão (2005) apresentam uma análise do comportamento diferencial dos itens de geografia aplicados aos alunos da 4ª série no Programa de Avaliação da Rede Pública de Educação Básica, o Proeb-2001, nas diferentes regiões do estado de Minas Gerais. Os resultados sugerem que itens relacionados a questões ambientais são mais fáceis para os alunos da região metropolitana de Belo Horizonte do que para aqueles do interior do estado. Por outro lado, os itens que avaliam a relação entre o espaço urbano e o espaço rural se mostram mais fáceis para os alunos do interior.

Barroso e Franco (2008) realizaram uma análise comparativa entre países participantes do PISA 2000, utilizando a TRI e a identificação de questões que apresentavam DIF. O objetivo dos autores era verificar se o desempenho dos estudantes brasileiros teria ou não características diferentes de alunos de outros países, e se essas características poderiam revelar diferentes ênfases curriculares no ensino de Ciências, apesar de o foco da edição investigada ter sido linguagem. Os resultados obtidos indicaram a existência de itens

com DIF, mas não permitiram a explicação desse comportamento com base nos parâmetros escolhidos associados às ênfases curriculares. Isso se deveu, segundo os autores, ao pequeno número de itens disponíveis em 2000, apenas 34, o que apontava para a necessidade de técnicas estatísticas mais elaboradas e a utilização dos dados do PISA 2006 para avançar nesse objetivo.

Aguiar (2008), a partir dos dados do PISA 2003, comparou as diferenças nas ênfases curriculares em Matemática, no Brasil e em Portugal. Os resultados do estudo mostraram que alguns itens de Matemática apresentam funcionamento diferencial entre alunos brasileiros e portugueses. Para o autor, os aspectos que explicam tal ocorrência estão relacionados com ênfases diferenciadas não apenas em determinados conteúdos da Matemática, mas também a processos cognitivos e ao formato do item.

Gamerman, Soares e Gonçalvez (2010) realizaram uma análise bayesiana na Teoria da Resposta ao Item aplicada ao PISA 2003 e identificaram uma série de indicadores que diferenciam os sistemas educativos dos países de língua inglesa participantes do Programa (Grã-Bretanha, Canadá, Austrália, Irlanda, Estados Unidos e Nova Zelândia). Esses indicadores, segundo os autores, podem ajudar a compreender a natureza e as possíveis origens da diferença entre esses países e mostrar um possível caminho para a incorporação de práticas que favorecem o aprendizado nesses sistemas de ensino.

Segundo Aguiar (2008), as análises sugerem que, em vez de entendermos o item do teste como a única causa do funcionamento diferencial, devemos considerar, também, questões de equidade educacional em nossas escolas e em nossa sociedade. O adequado entendimento dos resultados de DIF passa, necessariamente, pelo reconhecimento dessas desigualdades socioeducacionais. Aliados à gama de evidências empíricas produzidas pelos trabalhos de análise de DIF, muitos métodos estatísticos foram desenvolvidos no intuito de dar maior suporte a esse tipo de abordagem.

MÉTODOS DE DETECÇÃO DE ITENS COM DIF

Existem vários procedimentos formais para se estudar o funcionamento diferencial dos itens. Grosso modo, esses procedimentos podem ser divididos em dois grupos:

- os clássicos, que dependem, direta ou indiretamente, de uma estimativa prévia da proficiência, como, por exemplo, o método de Mantel-Haenszel (HOLLAND; THAYER, 1988) e o método de regressão logística (SWAMINATHAN; ROGERS, 1990);
- os baseados nos modelos da TRI, que utilizam os parâmetros dessa teoria e, apesar de não precisarem de uma proficiência já conhecida, demandam um critério alternativo para a equalização dos indivíduos *a priori*, tais como um subconjunto de itens que não possuam DIF, genericamente chamados de itens âncora. Exemplos bastante conhecidos desses métodos são o IRT-D^2 (THISSEN, 2001), o IRT-LR (THISSEN; STEINBERG; WAINER, 1993), e o método usado no BILOGMG (ZIMOWSKI et al., 1996).

Outros métodos ainda podem ser encontrados em Clauser e Mazor (1998) e em Andriola (2001). O método de Mantel-Haenszel é o mais utilizado para a análise de DIF, inclusive no *Educational Testing Service* (ETS), nos exames do *National Assessment for Educational Progress* (NAEP); e aqui no Brasil, na análise do Sistema de Avaliação da Educação Básica (Saeb).

De acordo com Soares, Gonçalves e Gamerman (2009), a detecção dos itens com DIF é um passo importante na análise de DIF, mas uma análise completa também requer alguns outros passos. Isso inclui uma satisfatória classificação do DIF encontrado, a identificação dos fatores a ele associados e, possivelmente, uma análise confirmatória das hipóteses. Schmitt, Holland e Dorans (1993) sugerem que estudos especialmente planejados devem ser utilizados para confirmar as hipóteses formuladas a partir do estudo dos fatores de DIF. Nesse contexto, é natural a construção de modelos de regressão que associam covariáveis, relacionadas com certas características dos itens, à magnitude do DIF. As covariáveis representariam os fatores de DIF de tal maneira que os resultados da análise de regressão podem confirmar ou não as hipóteses formuladas.

Os métodos de análise de DIF englobam várias etapas: detecção, explicação, confirmação; e neles, mesmo a detecção deve ser executada também em múltiplas etapas, como, por exemplo, detecção, purificação, nova detecção, confirmação.

Em uma nova proposta, Soares, Gonçalves e Gamerman (2009) descrevem um modelo bayesiano integrado para detecção e análise de DIF que elimina a necessidade de utilização dessas etapas separadas. Para o modelo proposto por esses autores, se existir um subconjunto de itens âncora, isto é, itens sem DIF, que é conhecido *a priori*, admite-se que os parâmetros dos demais itens possam variar entre os grupos de indivíduos, cabendo ao modelo indicar a probabilidade de eles apresentarem DIF. Assim, sempre que houver itens âncoras *a priori*, está garantida a correta identificação do DIF pelo modelo. Contudo, ele também pode ser usado quando não se conhece *a priori* um grupo de itens que não tenham DIF. Nesse último caso, é preciso que haja informação suficiente que possa ser expressa em uma probabilidade *a priori* sobre a não existência de DIF em alguns itens e/ou informação *a priori* sobre as distribuições de proficiências dos grupos focais e de referência. Por exemplo, pode-se admitir *a priori* que as proficiências dos alunos japoneses são mais elevadas do que as dos brasileiros.

Como resultado, não é necessário fixar um conjunto de itens que não apresentem DIF *a priori*, como itens âncoras para identificar o modelo. Estudos simulados, realizados pelos autores (SOARES; GONÇALVES; GAMERMAN, 2009), mostraram uma boa recuperação dos parâmetros gerados em várias situações, sendo que um exemplo real demonstrou a viabilidade da utilização do modelo em situações práticas com resultados satisfatórios e consistentes. Por essas, entre outras vantagens, no presente estudo utilizamos esse modelo integrado que elimina etapas, dado que uma análise de regressão associada aos parâmetros do DIF é introduzida no modelo de tal forma que, além de confirmar a ocorrência, também possibilita, simultaneamente, explicar o DIF.

METODOLOGIA

Ao dar início a este estudo, objetivando identificar os itens que apresentaram DIF, utilizamos o modelo integrado proposto por Soares, Gonçalves e Gamerman (2009). O modelo, como descrito na seção anterior, é integrado no sentido de permitir a detecção e explicação do DIF simultaneamente, ou seja, numa só etapa de inferência. Assim, ele utiliza apenas o pressuposto de que um subconjunto no total de itens analisados não possui DIF, sem que seja necessário os identificar, sendo capaz de calcular a probabilidade de cada item possuir DIF, assim como os parâmetros para cada item em cada grupo e a diferença entre eles. Adicionalmente, o algoritmo calcula as proficiências, médias e desvio padrão, de cada grupo.

O modelo foi implementado no *solver* OpenBUGS®,[3] que permite a realização de inferência bayesiana utilizando amostragem de Gibbs.[4] Os valores das variáveis indicadoras são estimados diretamente no modelo, indicando quais itens apresentam DIF e quais não. Soares, Gonçalves e Gamerman (2009), a partir de dois estudos simulados – um para mostrar as vantagens do modelo integrado sobre aqueles que fixam itens âncoras *a priori*, e outro que compara o modelo integrado aos métodos mais utilizados na detecção de DIF para diferentes configurações de DIF – e de uma análise do Programa Nova Escola, demonstram a eficiência do referido método.

Apesar de o PISA ser corrigido utilizando-se o modelo de Rasch, que permite identificar DIF apenas no parâmetro de dificuldade (parâmetro *bj*), para as análises de DIF do presente estudo ajustamos o modelo da TRI de três parâmetros (3PL). O ajuste desse modelo contempla uma maior flexibilidade das formas da Curva Característica do Item (CCI) que especifica a relação matemática entre a proficiência e a probabilidade de acerto de um item.

O modelo 3PL resulta da incorporação do parâmetro *c*, que representa a probabilidade de acerto ao acaso, ao modelo de dois parâmetros que leva em conta, além do parâmetro de dificuldade, o parâmetro de discriminação do item (parâmetro *aj*). O acerto casual pode representar, inclusive, a influência de um "chute" nos testes de múltipla escolha,

3 BUGS é um pacote de *software* para a realização de inferência bayesiana utilizando amostragem de Gibbs. O usuário especifica um modelo estatístico, de complexidade arbitrária, simplesmente dizendo as relações entre as variáveis relacionadas. O *software* inclui um "sistema especialista", que determina um regime adequado MCMC (cadeia de Markov Monte Carlo), com base no amostrador de Gibbs para analisar o modelo especificado.

4 Amostragem de Gibbs é um algoritmo iterativo para gerar uma sequência de amostras a partir de uma distribuição posterior conjunta por amostragem repetida a partir da distribuição condicional plena. Sob condições apropriadas, pode ser demonstrado que a sequência aleatória que representa os desenhos aleatórios sucessivos constituem uma cadeia de Markov que converge para uma distribuição estacionária igual à distribuição posterior conjunta. Para mais detalhes, ver Gamerman e Lopes (2006).

relacionando, inclusive, a uma resposta dada devido a outro traço que não exatamente aquele avaliado no teste ou ainda à resposta aleatória.

Métodos tradicionais para análise de DIF são baseados em habilidades pré-calculadas para a análise DIF. No entanto, como apontam Soares, Gonçalves e Gamerman (2009), a habilidade assim pré-calculada está contaminada justamente pelo possível DIF existente. Embora essa contaminação possa ser pequena e não interferir no resultado da análise, os autores sugerem no método proposto que proficiência e detecção do DIF sejam realizadas simultaneamente. Num cenário em que haja muito DIF em um teste, principalmente, naquele em que a presença de DIF que favorece substancialmente um grupo em detrimento de outro, a estimação simultânea do DIF apresenta considerável superioridade em relação aos métodos tradicionais. Por esse motivo, preferimos aqui reestimar as proficiências no processo de detecção do DIF.

No estudo, não consideramos a possibilidade de DIF no parâmetro *c*. Apesar de ser possível, a aplicabilidade desse caso é substancialmente limitada tanto pela sabida dificuldade de estimação desse parâmetro como por restrições práticas.

Isso posto, além dos parâmetros dos modelos, é importante que os seguintes conceitos sejam definidos:

- da_j representa a diferença entre o parâmetro a_j do grupo focal menos tal parâmetro no grupo de referência, ele indica o quanto o item *j* discrimina mais no grupo focal em relação ao de referência;
- db_j representa a diferença entre o parâmetro b_j do grupo focal menos tal parâmetro no grupo de referência, ele indica o quanto o item *j* se apresenta mais difícil no grupo focal em relação ao de referência;
- Za_j mede a probabilidade de ocorrência de DIF no parâmetro a_j no item *j*;
- Zb_j mede a probabilidade de ocorrência de DIF no parâmetro b_j no item *j*.

Note-se que, como definido aqui, o parâmetro da_j difere do parâmetro original de DIF utilizado em Soares, Gonçalves e Gamerman (2009). De fato, na notação desses autores o parâmetro de DIF é introduzido de forma multiplicativa, tal

que a discriminação do item no grupo focal g é dada por $e^{-d^{u}_{j2}} a_{j1}$, onde a_{j1} é a discriminação do item no grupo de referência. É fácil verificar que a relação entre os dois parâmetros, isto é, o parâmetro de DIF considerado aqui e o parâmetro de DIF naquele artigo é a seguinte: $daj = a_{j1} - a_{j2} = a_{j1} - e^{-d^{u}_{j2}} a_{j1}$

O modelo bayesiano proposto por Soares, Gonçalves e Gamerman (2009) permite que se compute a probabilidade *a posteriori* de o item ter DIF nos parâmetros de discriminação e dificuldade. Essas probabilidades são representadas aqui pelos termos *Zaj* e *Zbj*. Nesta pesquisa consideramos como tendo DIF na discriminação e na dificuldade os itens que apresentavam valores para *Zaj* e *Zbj* maiores que 0,6. Apesar de Soares, Gonçalves e Gamerman (2009) consideraram como tendo DIF aqueles itens que apresentavam um *Z* maior do que 0,5, o uso do valor 0,6 para a regra de classificação de DIF dá mais peso para as variáveis de regressão.

Como já mencionado, utilizamos aqui os dados do PISA 2006, do Brasil e do Japão, que são de domínio público. Trabalhamos apenas com itens dicotômicos, de forma que, de um total de 103 itens de Ciências, foram excluídos da análise seis itens com respostas corrigidas na forma de crédito parcial e um item não comum aos dois países, resultando num total de 96 itens selecionados para a presente análise. A amostra de alunos do Brasil foi considerada o grupo de referência e a do Japão o grupo focal. Visando a utilizar ao máximo sem, no entanto, extrapolar a capacidade de processamento do *software* OpenBUGS®, foi selecionada uma amostra com 3.500 casos, sendo 2.104 alunos brasileiros e 1.396 japoneses, a fim de obter cerca de 1.000 respostas para cada item, por amostragem aleatória simples sem reposição. No final, obteve-se uma média de 1.018 respostas para cada item, com um mínimo de 949 e um máximo de 1.070 respostas.

Os objetivos principais do trabalho foram:

- identificar os itens com DIF nos parâmetros de dificuldade e discriminação, analisando, ainda, a magnitude do DIF encontrado e verificando se ele beneficia um dos dois países estudados. Para essa etapa, o modelo de Soares, Gonçalves e Gamerman (2009) foi utilizado sem covariáveis explicativas; e

- explicar a existência e a magnitude do DIF por meio de seis covariáveis: competências, áreas do conhecimento, áreas de aplicação, âmbito, tipo e idioma. Para essa etapa foram realizadas análises separadas para cada covariável, uma vez que, conquanto o modelo de Soares, Gonçalves e Gamerman (2009) possa testar todas as covariáveis simultaneamente, questões como multicolinearidade e excesso de covariáveis decorrentes do número de categorias testadas em cada uma delas reduzem a eficácia e a sensibilidade na detecção do DIF. Por outro lado, embora o modelo desses autores permita, a implementação do programa disponível fornece apenas a significância estatística para cada categoria da covariável na explicação da magnitude do DIF dos itens com DIF, mas não comporta verificar estatisticamente se o número de itens com DIF para uma dada categoria é maior do que para as outras categorias. Assim, a fim de se mensurar estatisticamente se uma dada categoria de uma covariável apresenta mais itens com DIF do que as outras categorias, o que poderíamos denominar num certo sentido de *prevalência* de DIF, foram empregados testes estatísticos posteriores à identificação do DIF para comparar se uma dada categoria de uma covariável apresenta ou não mais itens com DIF do que as demais categorias, por exemplo. Naturalmente, os testes estatísticos diminuem seu poder nesse caso, mas não há outra opção devido à restrição nas *saídas* produzidas pelo programa. Em todos os casos, o nível de significância adotado foi 0,05.

RESULTADOS E DISCUSSÃO

Do total de 96 itens analisados, 20 apresentaram DIF no parâmetro *aj* e 50 no parâmetro *bj*. Contudo, independentemente do parâmetro avaliado, a soma de itens com DIF é composta de apenas 62 itens, uma vez que oito itens registraram DIF tanto no parâmetro *aj* como no parâmetro *bj*.

Recorrendo a um item público, liberado para divulgação pelo consórcio que administra o PISA, exemplificamos, resumidamente, os procedimentos adotados na identificação do DIF. Esclarecemos que o mesmo procedimento foi adotado com os demais itens de Ciências do PISA 2006, mas eles não serão apresentados aqui.

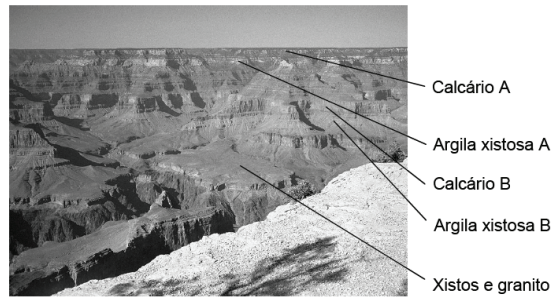
O item S426Q03 (Figura 1) é de múltipla escolha e trata o tema do "meio ambiente" no âmbito "social". A competência envolvida é a de "explicar fenômenos cientificamente", sobretudo no que diz respeito ao conhecimento de "terra e sistemas espaciais". Esse item foi elaborado pelo instituto australiano ACER originalmente em inglês.

FIGURA 1 – Unidade Grand Canyon, Questão 3. Código S426Q03

O GRAND CANYON

O Grand Canyon está localizado em um deserto nos Estados Unidos. Ele é um cânion grande e profundo formado por muitas camadas de rochas. No passado, os movimentos na crosta terrestre ergueram estas camadas. Atualmente, o Grand Canyon apresenta 1,6 km de profundidade em determinadas partes. O Rio Colorado percorre todo o fundo do cânion.

Veja a foto abaixo do Grand Canyon tirada da margem sul. Várias camadas diferentes de rochas podem ser vistas nas paredes do cânion.



QUESTÃO 3: O GRAND CANYON *S426Q03*

A temperatura no Grand Canyon varia de menos de 0 °C a mais de 40 °C. Embora ele esteja localizado em uma área desértica, as fendas das rochas, algumas vezes, contêm água. De que maneira essas mudanças de temperatura e a água contida nas fendas das rochas ajudam a acelerar a decomposição das rochas?

A A água congelada dissolve as rochas quentes.
B A água consolida as rochas entre si.
C O gelo torna lisa a superfície das rochas.
D A água congelada se expande nas fendas das rochas.

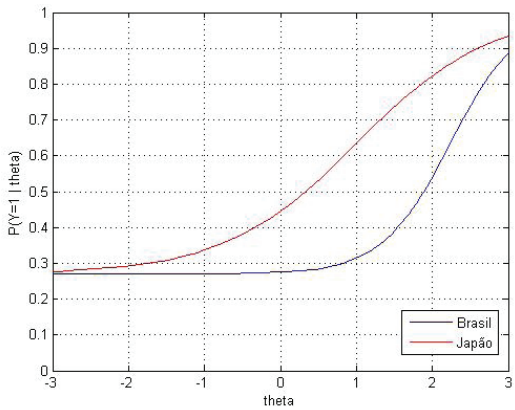
Fonte: Brasil (2008).

A resposta correta dessa questão – letra D – requer que o aluno saiba que a água congela quando a temperatura está

abaixo de zero grau, assim como conheça a propriedade da água de se expandir ao congelar, relacionando um fenômeno físico com um efeito geológico visível. Há um nítido contraste do percentual válido de respostas certas entre os alunos do Brasil (31%) e os do Japão (68%). No Relatório Nacional do Inep, os técnicos apontam como fator favorável aos estudantes da OCDE, cujo percentual de acerto foi um pouco menor do que o do Japão – 66,3% –, a maior convivência deles com as características do fenômeno de congelamento da água, devido ao clima frio. Esse fator também pode ser atribuído aos estudantes japoneses, para os quais esse item é mais fácil.

A partir da análise gráfica da Figura 2, observa-se a presença do DIF tanto na dificuldade (diferença no parâmetro *bj*) quanto na discriminação (diferença no parâmetro *aj*). De imediato percebe-se que as curvas características de Brasil e Japão são diferentes. A do Brasil é mais vertical do que a do Japão, indicando que o item discrimina mais os alunos brasileiros (*daS426Q03*= -0,5274). A probabilidade de acerto ao item, no entanto, é mais alta entre os alunos japoneses, indicando que, para quase todas as faixas de proficiências, o item é mais fácil para o Japão (*db S426Q03*=1,234).

FIGURA 2 – Curva característica do item S426Q03



Fonte: Dados do PISA 2006 (elaboração própria a partir do *software* Matlab. 2017).

Normalmente, os itens que apresentam DIF elevados e sistemáticos são identificados em pré-testes e análises estatísticas preliminares, realizados antes de serem utilizados para a produção da proficiência do aluno. São dedicados esforços e recursos substanciais para alcançar amplitude e equilíbrio culturais e linguísticos dos instrumentos da avaliação. Aplicam-se mecanismos rigorosos de garantia de qualidade na tradução, na amostragem e na coleta de dados. Não se espera, em princípio, que se encontrem itens com padrões bem definidos associados à existência de DIF. No entanto, alguns itens que exibem algum grau de comportamento diferencial, como o que foi mostrado acima, podem trazer informação adicional relevante para entender algumas das possíveis diferenças educacionais existentes entre os países analisados. Na sequência, descrevemos cada uma das características relacionadas aos itens, buscando associá-las com o sentido e a magnitude do DIF.

DIF SEGUNDO AS COMPETÊNCIAS

Considerando os dois países, encontramos DIF em todas as competências avaliadas pelo PISA 2006. A competência com maior concentração de DIF no parâmetro discriminação – *a* – foi "identificar questões científicas" (28,6% dos itens dessa competência apresentam DIF em *a* comparada a 24,5% e 7,7% dos itens das competências "explicar fenômenos cientificamente" e "usar evidência científica", respectivamente). No que diz respeito à dificuldade do item – *b* –, a competência em que os itens mais se comportam de maneira diferente para alunos brasileiros e japoneses é "explicar fenômenos cientificamente" (57,1% dos itens dessa competência apresentam DIF em *b* comparados a 42,9% e 50,0% dos itens das competências "identificar questões científicas" e "usar evidência científica", respectivamente).

A competência cujos itens menos apresentam DIF no parâmetro *aj* é "usar evidência científica". Apenas 7,7% dos itens classificados nessa competência apresentaram DIF. Para testar estatisticamente a hipótese de que essa competência concentra menos itens com DIF do que as demais, foi

construída uma tabela de contingência 2x2 segundo a qual se verifica a significância da diferença entre as distribuições do número de itens com e sem DIF para a competência "usar evidência científica" e para as demais consideradas conjuntamente. O teste x^2 de Pearson para associação confirma a hipótese com um p-valor de 0,053, sugerindo uma tendência à ocorrência de itens com DIF no parâmetro *aj* menor nessa competência do que nas outras.

Na Tabela 1, apresentamos os resultados referentes à direção e à intensidade do DIF, no parâmetro de discriminação *a* (*daj*). Se houvesse valores em média positivos, eles indicariam que os itens alocados em determinada competência seriam mais discriminantes no Japão. Contudo, todos os coeficientes médios são negativos e, portanto, discriminam mais os itens para os alunos brasileiros do que para os japoneses. Esse resultado é estatisticamente significativo (p=0,059). Sendo assim, os itens da prova de Ciências do PISA 2006, segundo as competências avaliadas no Programa, são mais eficazes em diferenciar alunos brasileiros, em relação aos japoneses, com níveis distintos de proficiência. O teste *post hoc* demonstra que a diferença das médias também é significativa ao nível de 0,05: a começar pelos itens que usam evidência científica, seguidos daqueles que identificam questões científicas até chegar naqueles que explicam fenômenos cientificamente, de modo geral, os itens com DIF no parâmetro *aj*, segundo a competência, discriminam mais os alunos brasileiros.

TABELA 1 – Direção e intensidade do DIF, entre Brasil e Japão, segundo as competências no parâmetro *aj* nos itens de Ciências do PISA 2006

COMPETÊNCIA	N	MÉDIA	DESVIO PADRÃO	ERRO PADRÃO
Explicar fenômenos cientificamente	49	-0,1322	0,25601	0,03657
Identificar questões científicas	21	-0,0246	0,17434	0,03804
Usar evidência científica	26	-0,0238	0,16701	0,03275

Fonte: Dados do PISA 2006 (elaboração própria a partir dos resultados das análises de DIF).

* Pearson Chi-Square (p=0,059).

As diferenças no parâmetro *bj* não são tão expressivas quanto aquelas apresentadas no parâmetro *aj*. As faixas percentuais de presença e ausência de DIF no parâmetro *bj*, segundo a competência do item, estão distribuídas quase que uniformemente, em torno de 50%, indicando que, considerando-se a dificuldade, praticamente não há comportamento diferencial dos itens que favoreça ou prejudique algum dos dois grupos analisados. Todas as competências apresentam, portanto, a mesma prevalência de DIF (p=0,532).

DIF SEGUNDO A ÁREA DO CONHECIMENTO DO ITEM

Os conhecimentos científicos presentes na avaliação do PISA 2006 eram de dois tipos: conhecimento de Ciência; e conhecimento sobre Ciência. Os conhecimentos de Ciências relacionam-se diretamente ao conhecimento dos alunos sobre o mundo natural e foram selecionados a partir dos principais campos da Física, Química, Biologia, Ciências da Terra e do Espaço e Tecnologia. O conhecimento sobre Ciência tem mais relação com a Ciência propriamente dita. A primeira categoria, "investigação científica", centra-se em inquérito como o processo central da ciência e os vários componentes desse processo, ou seja, como os cientistas obtêm os dados. A segunda categoria, intimamente relacionada com a investigação, é "explicações científicas" e se refere mais aos resultados da investigação científica e à forma como os cientistas utilizam os dados colhidos.

Encontramos DIF em todas as áreas de conhecimento avaliadas pelo PISA 2006. Na área do conhecimento de Ciência, o descritor com maior concentração de DIF no parâmetro discriminação – *aj* – foi "terra e sistemas espaciais" (36,4% dos itens dessa área apresentaram DIF em *a* comparada às demais áreas). Nenhuma das outras áreas contempladas atingiu mais do que 25% de probabilidade de apresentarem DIF em *a*. No que diz respeito à dificuldade do item – *b* –, a concentração ocorreu em "sistemas vivos" (68,2% dos itens desse descritor apresentaram DIF). Ao contrário do que aconteceu no parâmetro *aj*, todas as áreas do conhecimento apontaram probabilidades, se não superiores, bem próximas a 50% de concentrarem DIF. Quando o conhecimento aferido foi sobre

Ciência, os DIFs se concentraram em "investigações científicas" (27,30% dos itens apresentam DIF nessa área, contra 12,50% dos itens na área "explicações científicas") no parâmetro *aj* e em "explicações científicas" (43,80% comparados a 40,90% de presença em "investigações científicas") no parâmetro *bj*.

Assim, aparentemente, os itens com DIF estão distribuídos homogeneamente. Não há concentrações tão representativas que sugiram uma incidência maior ou menor da ocorrência de DIF em determinada área de conhecimento nem no parâmetro *aj* e tampouco no parâmetro *bj*. Ainda assim, pelo fato de 68% dos itens alocados na área do conhecimento "sistemas vivos" apresentarem DIF no parâmetro *bj*, testamos a hipótese de esse descritor estar concentrando mais itens com DIF do que os demais, podendo, assim, estar favorecendo um grupo de alunos em detrimento do outro. Para tanto, construímos uma tabela 2x2 apenas considerando a correlação entre a distribuição de se ter DIF ou não, no parâmetro *bj,* para a área do conhecimento sobre Ciência "sistemas vivos" com a distribuição dos outros descritores agregados, inclusive aqueles do conhecimento sobre Ciência, conforme pode ser visto na Tabela 2. O teste x^2 de Pearson por associação, contudo, não confirmou tal hipótese, com um p-valor de 0,069, não se podendo afirmar que haja uma maior ocorrência de DIF em itens alocados na área de conhecimento "sistemas vivos", ainda que esta tenha uma quantidade representativa de itens com DIF.

TABELA 2 – DIF, entre Brasil e Japão, segundo as áreas do conhecimento no parâmetro *bj* dos itens de Ciências do PISA 2006

	APRESENTA DIF EM b		TOTAL
	NÃO	SIM	
Outros	39	35	74
	52,7%	47,3%	100,0%
Sistemas vivos	7	15	22
	31,8%	68,2%	100,0%
Total	46	50	96
	47,9%	52,1%	100,0%

Fonte: Dados do PISA 2006 (elaboração própria a partir dos resultados das análises de DIF).

* Pearson Chi-Square (p=0,069).

Da mesma forma, no parâmetro *aj,* a área de conhecimento dos itens de Ciências do PISA 2006 não privilegiou Brasil ou Japão. Apesar de todos os descritores apresentarem valores em média negativos no parâmetro *daj* (Tabela 3), o que tornaria os itens um pouco mais discriminantes no Brasil, as diferenças em relação ao Japão não são estatisticamente significativas (p=0,564).

TABELA 3 – Direção e intensidade do DIF, entre Brasil e Japão, segundo as áreas do conhecimento no parâmetro *aj* nos itens de Ciências do PISA 2006

	N	MÉDIA *daj*	DESVIO PADRÃO	ERRO PADRÃO
Terra e sistemas espaciais	11	-0,1093	0,27709	0,08355
Sistemas vivos	22	-0,0973	0,21894	0,04668
Sistemas físicos	17	-0,1219	0,28297	0,06863
Investigação científica	22	-0,0241	0,17015	0,03628
Explicações científicas	16	-0,0287	0,1908	0,0477
Sistemas tecnológicos	8	-0,1511	0,21515	0,07607
Total	96	-0,0793	0,22316	0,02278

Fonte: Dados do PISA 2006 (elaboração própria a partir dos resultados das análises de DIF).

* Pearson Chi-Square (p=0,564).

No parâmetro *bj*, inicialmente, apenas os itens alocados nos conhecimentos de "terra e sistemas espaciais" e "sistemas tecnológicos" estariam favorecendo um pouco os alunos japoneses. Todos os demais, por apresentarem valores em média negativos, seriam mais fáceis para os alunos brasileiros (Tabela 4). Contudo, essas diferenças encontradas no parâmetro *bj*, assim como aquelas observadas no parâmetro *aj*, entre Brasil e Japão, não são estatisticamente significativas (p=0,470) e, portanto, não se pode afirmar que este ou aquele descritor esteja favorecendo um grupo em detrimento do outro.

TABELA 4 – Direção e intensidade do DIF, entre Brasil e Japão, segundo as áreas do conhecimento no parâmetro *bj* nos itens de Ciências do PISA 2006

	N	MÉDIA *dbj*	DESVIO PADRÃO	ERRO PADRÃO
Terra e sistemas espaciais	11	0,0887	0,63643	0,19189
Sistemas vivos	22	-0,3131	0,81572	0,17391
Sistemas físicos	17	-0,0564	0,47605	0,11546
Investigação científica	22	-0,0726	0,55554	0,11844
Explicações científicas	16	-0,0466	0,33442	0,0836
Sistemas tecnológicos	8	0,0222	0,44956	0,15894
Total	96	-0,0941	0,58781	0,05999

Fonte: Dados do PISA 2006 (elaboração própria a partir dos resultados das análises de DIF).

* Pearson Chi-Square (p=0,470).

DIF SEGUNDO A ÁREA DE APLICAÇÃO DO ITEM

Além das competências e das áreas do conhecimento, outra característica pública dos itens de Ciências do PISA 2006 é sua área de aplicação, que está centrada em seu emprego em relação a contextos pessoais, sociais e globais, tais como: saúde, recursos naturais, meio ambiente, fenômenos naturais e limites da ciência e da tecnologia.

Tanto no parâmetro *aj* quanto no *bj* há ocorrência de itens com DIF em todas as áreas avaliadas. Contudo, no parâmetro *aj*, os itens com DIF estão mais concentrados nas áreas de "meio ambiente" e "limites da ciência e da tecnologia", 33,3% dos itens em ambos os casos. Já no parâmetro *bj,* as áreas que apresentam mais itens com DIF são "fenômenos naturais" (76,9%) e "saúde" (72%).

Para testar a hipótese de que "meio ambiente" e "limites da ciência e da tecnologia" estariam concentrando itens com DIF no parâmetro *aj*, recodificamos a variável "área de aplicação" em "área de aplicação.rec", ou seja, numa nova variável em que foi agregado o conjunto de áreas de aplicação diferentes de "meio ambiente" e "limites da ciência e da tecnologia". Consideramos, portanto, apenas a correlação entre a distribuição de se ter DIF ou não, no parâmetro *aj* para essas duas áreas, com a distribuição das demais áreas agregadas, conforme pode ser visto na Tabela 5. O teste x^2 de Pearson por associação confirma a hipótese com um p-valor de 0,045, sugerindo uma inclinação maior à ocorrência de DIF, no

parâmetro *aj*, em itens alocados nessas áreas de "meio ambiente" e "limites da ciência e da tecnologia" do que nas demais.

TABELA 5 – DIF, entre Brasil e Japão, segundo as áreas de aplicação no parâmetro *aj* nos itens de Ciências do PISA 2006

	APRESENTA DIF EM a		TOTAL
	NÃO	SIM	
Meio ambiente	10	5	15
	66,7%	33,3%	100,0%
Limites da ciência e da tecnologia	16	8	24
	66,7%	33,3%	100,0%
Outras	50	7	57
	87,7%	12,3%	100,0%
Total	76	20	96
	79,2%	20,8%	100,0%

Fonte: Dados do PISA 2006 (elaboração própria a partir dos resultados das análises de DIF).

Apesar de concentrarem os itens com DIF no parâmetro *aj*, as áreas "meio ambiente" e "limites da ciência e da tecnologia" não fazem distinção entre brasileiros e japoneses, ou seja, do ponto de vista estatístico, não se pode afirmar que os itens alocados nessas duas ou nas demais áreas de aplicação de Ciências do PISA 2006 discriminem mais os estudantes brasileiros do que os japoneses e vice-versa (p=0,801). Decidimos filtrar a categoria "outras", por esta apresentar apenas dois itens, mas ainda assim não se encontrou um p-valor que permitisse sustentar a hipótese de que itens de qualquer das áreas discriminassem mais no Brasil ou no Japão (p=0,071).

Os resultados da ocorrência de DIF no parâmetro *bj,* segundo as áreas de aplicação, apontam uma concentração de itens com DIF nas áreas "fenômenos naturais" e "saúde". Diferentemente do que se observou no parâmetro *aj*, no que diz respeito à dificuldade (parâmetro *bj*), os itens de fato se comportam de maneira diferente para alunos brasileiros e japoneses. O teste x^2 de Pearson por associação confirma o DIF segundo a área de aplicação no parâmetro *bj* com um p-valor de 0,020. Filtramos a categoria "outras", novamente por esta apresentar apenas dois itens, mas ainda assim encontrou-se um p-valor que sustenta a hipótese de que as

áreas de aplicação avaliadas pelo PISA concentram itens com DIF considerando Brasil e Japão (p=0,071).

A fim de verificar a direção e a intensidade do DIF, em *bj* (*dbj*), construímos a Tabela 6 considerando a correlação entre a distribuição da direção e da intensidade do DIF das áreas de aplicação, excluindo a categoria "outras", com apenas dois itens. Três delas, "meio ambiente", "saúde" e "recursos naturais", tendem a apresentar valores em média negativos, enquanto as demais – "limites da ciência e da tecnologia" e "fenômenos naturais" – apresentam, em média, valores positivos. Como vimos, os valores em média positivos indicam que os itens alocados em determinada competência seriam mais fáceis para o Japão. Ao contrário, aqueles negativos seriam mais fáceis para os alunos brasileiros em relação aos japoneses. Esse resultado é estatisticamente significativo (p=0,050).

TABELA 6 – Direção e intensidade do DIF, entre Brasil e Japão, segundo as áreas de aplicação no parâmetro *bj* nos itens de Ciências do PISA 2006

	N	MÉDIA	DESVIO PADRÃO	ERRO PADRÃO
Meio ambiente	15	-0,0805	0,57094	0,14742
Limites da ciência e da tecnologia	24	0,0797	0,52475	0,10711
Fenômenos naturais	13	0,2019	0,7074	0,1962
Saúde	25	-0,2607	0,63613	0,12723
Recursos naturais	17	-0,2964	0,40617	0,09851
Total	94	-0,0875	0,5912	0,06098

Fonte: Dados do PISA 2006 (elaboração própria a partir dos resultados das análises de DIF).

* Pearson Chi-Square (p=0,050).

DIF SEGUNDO O ÂMBITO OU CONTEXTO DO ITEM

No PISA 2006, as situações da vida real que demandam do aluno posicionamento ou conhecimentos podem corresponder a três âmbitos ou círculos concêntricos de abrangência da questão: pessoal, social e/ou global. O contexto que mais apresentou DIF no parâmetro *aj* foi o "pessoal". Contudo, nenhum dos contextos registra mais de 25% de DIF no parâmetro *aj*. Já no parâmetro *bj*, observa-se maior predominância de DIF no contexto "global" (68,8%), seguido pelo "pessoal" (61,5%) e, por fim, mais de 40% no contexto "social".

De fato, os dois primeiros contextos descritos anteriormente, "global" e "pessoal", tendem a concentrar mais DIF no parâmetro de dificuldade do item (*b*) do que o contexto "social" (p=0,035). Essa hipótese foi testada a partir de uma tabela de contingência 2x2 que agregou os contextos "global" e "pessoal", associando-os ao contexto "social" responsável pela maior quantidade de itens no teste, com 54 no total, contra 42 dos outros dois contextos juntos (Tabela 7). Isso quer dizer que, mesmo em menor número no teste e juntos, os itens alocados nos contextos "global" e "pessoal" concentram mais comportamento diferencial do que aqueles construídos no âmbito "social".

TABELA 7 – DIF, entre Brasil e Japão, segundo o contexto no parâmetro *bj* nos itens de Ciências do PISA 2006

		APRESENTA DIF EM b		TOTAL
		NÃO	SIM	
Contexto do item	Outro	15	27	42
		35,7%	64,3%	100,0%
	Social	31	23	54
		57,4%	42,6%	100,0%
Total		46	50	96
		47,9%	52,1%	100,0%

Fonte: Dados do PISA 2006 (elaboração própria a partir dos resultados das análises de DIF).

Não se encontraram diferenças estatisticamente significativas entre as médias dos coeficientes de *daj* e *dbj*, de intensidade do DIF na dificuldade e/ou discriminação dos itens. Isso significa que o DIF aparentemente está distribuído de maneira uniforme entre os itens dos diferentes contextos, e não privilegia nem prejudica nenhum dos dois países. Apesar de a maioria das médias dos coeficientes ter sido negativa e, assim, sugerir uma maior discriminação e/ou facilidade dos itens para o Brasil, os p-valores encontrados não foram estatisticamente significativos (p-valor de 0,927 para a diferença no parâmetro *aj* (*daj)* e de 0,350 para o parâmetro *bj* (*dbj*)).

DIF SEGUNDO O TIPO DE ITEM

Os tipos de itens empregados no teste de Ciência do PISA 2006 foram de múltipla escolha e resposta construída. Os itens de múltipla escolha eram, no entanto, padronizados com quatro alternativas de respostas, a partir das quais os alunos eram obrigados a selecionar a melhor; ou complexos, apresentando várias declarações para cada um, entre as quais os estudantes deviam escolher uma das várias possíveis respostas (sim / não, verdadeiro / falso, correto / incorreto, etc.). Os itens de resposta construída também foram classificados de forma diferenciada pelo PISA. Nos itens de resposta construída fechada, era necessário que os alunos construíssem uma resposta numérica dentro de restrições muito limitadas, ou apenas uma palavra ou uma curta frase como resposta. Os itens de resposta construída aberta exigiam respostas mais completas ou extensas, que frequentemente abarcavam alguma explicação ou justificativa.

Todos os tipos de item apresentaram DIF. No entanto, a ocorrência de comportamento diferencial no parâmetro *aj* (50%) não foi tão expressiva quanto no parâmetro *bj* (70,4%). O DIF encontrado na dificuldade e na discriminação dos itens não se mostrou associado ao tipo de item, de tal forma que esse aspecto não torna um item mais ou menos discriminante, mais fácil ou mais difícil, para os alunos de nenhum dos dois países. Em outras palavras, não se encontraram diferenças significativas entre as médias de intensidade do DIF na dificuldade do item para os diferentes tipos de itens (p-valor de 0,516) e tampouco entre as médias de intensidade do DIF na discriminação do item para os diferentes tipos de itens de Ciências do PISA 2006 (p-valor de 0,107). Isso quer dizer que, embora existam itens com DIF, não há indícios suficientes de que o DIF esteja privilegiando um grupo em detrimento do outro, facilitando ou discriminando mais, por exemplo, o desempenho dos alunos brasileiros e/ou dos japoneses.

DIF SEGUNDO O IDIOMA DO ITEM

Os itens de Ciências do PISA 2006 foram originalmente escritos em dez idiomas distintos. No entanto, mais de 30% do

total foi escrito originalmente em inglês e 11 itens em inglês apresentam DIF no parâmetro *aj* e 12 no parâmetro *bj*. Contudo, itens com DIF em *a* não necessariamente apresentam DIF em *b* e vice-versa (p=0,283). Os itens escritos, originalmente, na língua inglesa mostram mais DIF no parâmetro *aj* do que os itens escritos nos demais idiomas (correlação testada estatisticamente pelo teste qui-quadrado para associação entre variáveis, tendo-se encontrado um p-valor de 0,007). Isso pode ser observado na Tabela 8 de contigência 2x2 apresentada a seguir.

No entanto, comparando-se a intensidade e direção do DIF, por meio dos coeficientes *daj* calculados para cada item, não se encontram diferenças significativas entre as suas médias (-0,2938 para os itens escritos em outros idiomas e -0,0865 para aqueles em inglês), segundo um teste t para diferenças entre médias para o qual se encontrou um p-valor de 0,439.

Os DIFs encontrados no parâmetro *bj* dos itens também não se mostraram associados ao idioma (p-valor de 0,477), de tal forma que o idioma não torna mais fácil ou mais difícil um item para os alunos de nenhum dos dois países. Da mesma forma, não se encontraram diferenças significativas entre as médias de intensidade do DIF na dificuldade do item para os diferentes idiomas (p-valor de 0,283). Isso quer dizer que, embora existam itens com DIF segundo o idioma, não há indícios suficientes de que o DIF esteja privilegiando um grupo em detrimento do outro, facilitando, por exemplo, o desempenho dos alunos brasileiros e/ou dos japoneses.

TABELA 8 – DIF, entre Brasil e Japão, segundo o idioma no parâmetro *aj* nos itens de Ciências do PISA 2006

	APRESENTA DIF EM a		TOTAL
	NÃO	SIM	
Outros idiomas	58	9	67
	86,6%	13,4%	100,0%
Inglês	18	11	29
	62,1%	37,9%	100,0%
Total	76	20	96
	79,2%	20,8%	100,0%

Fonte: Dados do PISA 2006 (elaboração própria a partir dos resultados das análises de DIF).

* Pearson Chi-Square (p=0,007).

CONCLUSÃO

As análises comparativas realizadas demonstram que, não obstante os cuidados que cercam a elaboração e seleção de itens dessa avaliação internacional de grande porte, há significativa presença de DIF nos itens de Ciências do PISA 2006, quando se comparam o Brasil e o Japão. Cabe lembrar que mesmo os itens diagnosticados com DIF nem sempre são capazes de comprometer o processo avaliativo ao privilegiar um grupo em detrimento do outro.

Neste estudo, no total de 96 itens analisados, foram identificados 62 com DIF, oito dos quais com DIF tanto no parâmetro *aj* quanto no *bj*. As conclusões sobre as características do DIF a que chegamos, após os resultados estimados pelo modelo bayesiano integrado, podem ser assim expressas:

- considerando o número de itens, a prova de Ciências foi mais fácil para o Japão – 28 dos 50 itens com DIF no parâmetro *bj* foram mais fáceis para os alunos japoneses. Contudo, o DIF encontrado nesses itens não afeta significativamente os resultados gerais do teste, tendo em vista que ele está localizado em uma parte dos itens e que alguns são mais fáceis para o Japão e outros para o Brasil. Por outro lado, sob o mesmo critério de número de itens, a prova discrimina mais os alunos brasileiros. Dos 20 itens com DIF no parâmetro *aj*, 14 discriminam mais a "população" de alunos do Brasil;
- no que diz respeito às competências:
  - itens que mobilizam a competência "usar evidência científica" tendem a apresentar menos DIF no parâmetro *aj* quando comparada às demais competências;
  - os itens de Ciências apresentam mais DIF nas discriminações favoráveis ao Brasil. Isso significa que a maior parte dos itens que apresentam DIF no parâmetro de discriminação é favorável ao Brasil;
  - as diferenças no parâmetro *bj* estão distribuídas quase que uniformemente entre os grupos de itens das diferentes competências e indicam

que, considerando-se a dificuldade do item, a proporção de itens que favorecem um grupo e outro se distribui igualmente entre os diferentes grupos de itens formados pelas diferentes competências. Assim, não há uma competência em que o DIF se concentre ou que apresente menos DIF do que o achado no teste como um todo. Isso aponta para o fato de que todas as competências apresentam a mesma prevalência de DIF;

- de acordo com a área de conhecimento do item não se pode afirmar que os itens dos testes de Ciência privilegiem o Brasil ou o Japão. Apesar de terem sido identificados itens com DIF em todos os descritores avaliados, não há evidência estatística que mostre que um ou outro descritor concentre maior prevalência de DIF e tampouco que esses comportamentos diferenciais tomem um sentido único de privilegiar um dos países;

• quanto à área de aplicação dos itens:
  - há uma tendência de maior ocorrência de DIF no parâmetro *aj* em itens alocados nas áreas de "meio ambiente" e "limites da ciência e da tecnologia" do que nas demais áreas avaliadas pelo PISA em 2006. No entanto, essa tendência não tem um sentido definido, ou seja, o DIF observado ora indica maior discriminação nos alunos do Japão ora nos do Brasil;
  - itens em três das áreas avaliadas – "meio ambiente", "saúde" e "recursos naturais" – tendem a apresentar valores de *dbj* negativos e, assim, são mais fáceis para os alunos brasileiros, enquanto os das demais áreas – "limites da ciência e da tecnologia" e "riscos" – tendem a apresentar valores de *dbj* positivos e mostram-se, portanto, mais fáceis para os alunos japoneses do que para os brasileiros;

• itens que medem habilidades nos contextos "global" e "pessoal" tendem a concentrar mais DIF no parâmetro de dificuldade do que aqueles relacionados ao contexto "social";

- os DIF encontrados na dificuldade e na discriminação dos itens não se mostraram associados ao formato do item, de tal maneira que esse aspecto não torna um item mais ou menos discriminante, ou mais fácil ou mais difícil, para os alunos de um dos dois países. Assim, embora existam itens com DIF, não há indícios suficientes de que o DIF esteja privilegiando um grupo em detrimento do outro, facilitando ou discriminando mais, por exemplo, o desempenho dos alunos brasileiros ou dos japoneses;
- embora os itens sejam elaborados em diferentes idiomas e, posteriormente, traduzidos para o idioma de cada país avaliado, o vocabulário e os termos utilizados não se constituem, *a priori*, num obstáculo à resolução do item tanto para os alunos japoneses como para os brasileiros que seja traduzido na análise do DIF.

Os modelos mais tradicionais da TRI pressupõem que os itens apresentem o mesmo funcionamento em diferentes grupos. Uma boa e justa comparação entre resultados de grupos diferentes de alunos requer, portanto, que os itens que compõem o teste não apresentem comportamento diferencial excessivo, pois, do contrário, isso significaria que um grupo em particular estaria sendo privilegiado em detrimento de outro. Diante desse pressuposto, usualmente busca-se produzir itens de teste que não apresentem DIF, ainda que essa seja uma tarefa muito difícil quando as populações avaliadas são tão distintas como é o caso de alunos de diferentes países. No entanto, parece que o teste do PISA tem sido produzido com qualidade o suficiente para a boa comparabilidade dos resultados entre os alunos do Brasil e do Japão.

REFERÊNCIAS

AGUIAR, Glauco. *Estudo comparativo entre Brasil e Portugal, sobre diferenças nas ênfases curriculares de Matemática, a partir da análise do Funcionamento Diferencial do Item (DIF) do PISA 2003*. 2008. 246f. Tese (Doutorado em Educação) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

ANASTASI, Anne. *Psychological testing*. New York: MacMillan. 1988.

ANDRIOLA, Wagner. Descrição dos principais métodos para detectar o funcionamento diferencial dos itens (DIF). *Psicologia:* Reflexão e Crítica, Rio Grande do Sul, v. 14, n. 3, p. 643-652, 2001.

BARROSO, Marta; FRANCO, Creso. Avaliações educacionais: o PISA e o ensino de ciências. In: ENCONTRO DE PESQUISA EM ENSINO DE FÍSICA, 11., 2008, Curitiba. *Anais...* Curitiba, 2008. Disponível em: <http://www.if.ufrj.br/~marta/artigosetal/2008-epef11-PISA.pdf>. Acesso em: 27 jul. 2014.

BERBEROGLU, Giray. Differential item functioning (DIF) analysis of computation, word problem and geometry questions across gender and SES groups. *Studies in Educational Evaluation*, Great Britain, v. 21, n. 4, p. 439-456, 1995.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Resultados nacionais – PISA 2006*: Programa Internacional de Avaliação de Estudantes (PISA). Brasília, DF: Inep, 2008.

CLAUSER, Brian; MAZOR, Kathleen. Using statistical procedures to identify differential item functioning test items. *Educational Measurement:* Issues and Practice, Philadelphia, v. 17, n. 1, p. 31-44, 1998.

COLE, Nancy. History and development of DIF. In: HOLLAND, Paul W.; WAINER, Howard (Ed.). *Differential item functioning*. Hilldsdale, NJ: Lawrence Erlbaum, 1993.

FERRER, Alejandro Tiana. Que variáveis explicam os melhores resultados nos estudos internacionais? In: AZEVEDO, Joaquim. *Avaliação dos resultados escolares*. Porto: ASA, 2003.

GAMERMAN, Dani; LOPES, Hedibert. *Markov Chain Monte Carlo*: Stochastic simulation for Bayesian inference. New York: Chapman & Hall / CRC, 2006.

GAMERMAN, Dani; SOARES, Tufi; GONÇALVES, Flávio. Bayesian analysis in item response theory applied to a large-scale educational assessment. In: O'HAGAN, Anthony; WEST, Mike. *The Oxford handbook of applied Bayesian analysis*. New York: Oxford University, 2010. p. 624-652.

GIERL, Mark; BISANZ, Jeffrey; BISANZ, Gay; BOUGHTON, Keith. Identifying content and cognitive skills that produce gender differences in mathematics: a demonstration of the DIF analysis framework. *Journal of Educational Measurement*, Philadelphia, v. 40, n. 4, p. 281-306, 2003.

HAMBLETON, Ronald; SWAMINATHAN, H.; ROGERS, Jane. *Fundamentals of Item Response Theory.* Newburry Parks: Sage, 1991.

HOLLAND, Paul; THAYER, Dorothy. Differential item performance and the Mantel–Haenszel procedure. In: HOLLAND, Paul W.; WAINER, Howard (Ed.). *Differential item functioning*. Hilldsdale, NJ: Lawrence Erlbaum, 1988. p. 129-145.

LINN, Robert; DRASGOW, Fritz. Implications of the golden rule settiemernt for test construction. In: HOLLAND, Paul W.; WAINER, Howard (Ed.). *Differential item functioning*. Hilldsdale, NJ: Lawrence Erlbaum, 1993.

O'NEIL, Kathleen; McPEEK, Miles. Item and test characteristics that are associated with differential item functioning. In: HOLLAND, Paul W.; WAINER, Howard (Ed.). *Differential item functioning*. Hilldsdale, NJ: Lawrence Erlbaum Associates, 1993. p. 255-276.

PASQUALI, Luiz. *Psicometria*: teoria dos testes psicológicos. Brasília, DF: Prática, 2000.

SCHMITT, Alicia P.; BLEISTEIN, Carole A. *Factors affecting differential item functioning for black examinees on scholastic aptitude test analogy items (ETS RR-87-23).* Princeton, NJ: Educational Testing Service, 1987.

SCHMITT, Alicia; HOLLAND, Paul; DORANS, Neil. Evaluating hypotheses about differential item functioning. In: HOLLAND, Paul W.; WAINER, Howard (Ed.). *Differential Item Functioning*. Hilldsdale, NJ: Lawrence Erlbaum, 1993. p. 281-316.

SOARES, Tufi. Utilização da Teoria de Resposta ao Item na produção de indicadores sócio-econômicos. *Pesquisa Operacional*, Rio de Janeiro, v. 25, n. 1, p. 83-112, jan./abr. 2005.

SOARES, Tufi; GENOVEZ, Silene; GALVÃO, Ailton. Análise do Comportamento Diferencial dos Itens de Geografia: estudo da 4ª série avaliada no Proeb/Simave, 2001. *Estudos em Avaliação Educacional*, São Paulo, v. 16, n. 32, p. 81-110, jul./dez. 2005.

SOARES, Tufi; GONÇALVES, Flávio; GAMERMAN, Dani. Na integrated Bayesian model for DIF analysis. *Journal of Educational ad Behavioral Statistics*, Washington, v. 34, n. 3, p. 348-377, Sep. 2009.

SWAMINATHAN, Hariharam; ROGERS, Jane. Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, Philadelphia, v. 27, p. 361-370, 1990.

THISSEN, David. *IRTLRDIF v.2.0.b*: software for the computation of the statistics involved in item response theory Likelihood-Ratio Tests for differential item functioning. 2001.

THISSEN, David; STEINBERG, Lynne; WAINER, Howard. Detection of differential item functioning using the parameters of item response models. In: HOLLAND, Paul W.; WAINER, Howard (Ed.). *Differential Item Functioning*. Hilldsdale, NJ: Lawrence Erlbaum, 1993. p. 67-114.

ZIMOWSKI, Michele F.; MURAKI, Eiji; MISLEVY, Robert J.; BOCK, R. Darrell. *BILOG-MG*: Multiple Group IRT Analysis and test maintenance for binary items. [Computer software]. Chicago: Scientific Software International, 1996.

ANDRIELE FERREIRA MURI
Doutoranda do Programa de Pós-Graduação em Educação da Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Rio de Janeiro, Brasil
andrielemuri@yahoo.com.br

TUFI MACHADO SOARES
Professor do Programa de Pós-Graduação em Educação da Universidade Federal de Juiz de Fora (UFJF). Coordenador da Unidade de Pesquisa do Centro de Políticas Públicas e Avaliação da Educação (CAEd) da UFJF, Juiz de Fora, Minas Gerais, Brasil
tufi@caed.ufjf.br

ALICIA BONAMINO
Professora do Programa de Pós-Graduação em Educação da Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Rio de Janeiro, Brasil
alicia@puc-rio.br

Recebido em: MARÇO 2017
Aprovado para publicação em: AGOSTO 2017