Análise do funcionamento diferencial dos itens do Exame Nacional do Estudante (ENADE) de psicologia de 2006

Ricardo Primi[1] – Universidade São Francisco, Itatiba, Brasil
Lucas Francisco de Carvalho – Universidade São Francisco, Itatiba, Brasil
Fabiano Koich Miguel – Universidade Estadual de Londrina, Londrina, Brasil
Marjorie Cristina Rocha da Silva – Faculdades Integradas Einstein de Limeira, Limeira, Brasil

Resumo
Parte do Sistema Nacional de Avaliação das Instituições de Educação Superior considera o desempenho dos estudantes por meio do ENADE. Neste artigo efetuou-se uma análise dos itens da prova do ENADE de psicologia aplicada em 2006 tentando-se detectar itens com funcionamento diferencial (DIF), isto é, itens com problema de equivalência ao medir ingressantes e concluintes e estudantes de instituições públicas e privada. Analisou-se uma amostra de 26.613 estudantes ingressantes e concluintes representativa de todos os cursos do país. Empregou-se a análise de Rasch e regressão logística para se detectar o DIF. Onze itens dos 30 que compunham a prova apresentaram DIF. Dois tipos de DIF ocorreram, um tipo em itens com baixa discriminação e outro em itens com alta discriminação. O subgrupo mais relevante tende a favorecer alunos de instituições públicas. Discute-se também a questão da discriminação elevada como indicativo de DIF.
Palavras-chave: Teoria de Resposta ao Item, ENADE, Regressão logística, Modelo de Rasch, Validade.

Differential item functioning of the national student exam for psychology (ENADE) 2006

Abstract
Part of the National Assessment of Institutions of Higher Education considers student performance through ENADE. In this article we performed differential item function analysis of the ENADE that took place in 2006 trying to detect items with problems in measurement equivalence in the assessment of freshman and senior students and from public and private institutions. We analyzed a sample of 26,613 freshmen and seniors representative of all the courses in the country. We used the Rasch analysis and logistic regression to detect DIF. Eleven of the 30 items composing the test showed DIF. Two types of DIF were observed, one occurring in less discriminating items and the other in more discriminating items. The most relevant subgroup of items tends to favor students from public institutions. We also discuss the issue of discrimination parameter being an indicator of DIF.
Keywords: Item Response Theory, ENADE, Logistic regression, Rasch model, Validity.

Sistemas de avaliação do Ensino Superior

Os sistemas de avaliação em larga escala, desenvolvidos por entidades públicas, têm um papel fundamental para a sociedade. Tais sistemas buscam levantar informações sobre a eficiência e qualidade das organizações que provem bens públicos essenciais à população, tais como saúde, educação e segurança. Essas informações são fundamentais para a gestão dos recursos públicos, pois permitem a visualização de virtudes e falhas do sistema para que ações interventivas e regulatórias sejam criadas com objetivo amplo de melhorar a qualidade do sistema (Primi, 2006).

Especialmente no que diz respeito ao sistema educacional, nos últimos dez anos, observa-se que o Ministério da Educação (MEC) tem colocado a avaliação como um dos alvos importantes de suas políticas. A partir da década de 1990, vários sistemas de avaliação, do ensino fundamental (Sistema de Avaliação do Ensino Básico – SAEB), médio (Exame Nacional do Ensino Médio – ENEM) e superior (Exame Nacional de Cursos – ENC e, recentemente, o Sistema Nacional de Avaliação do Ensino Superior – SINAES) foram criados pelo MEC (mais informações em www.inep.gov.br).

No que se refere ao Ensino Superior, esse processo teve início em 1995, com a criação do Exame Nacional de Cursos (ENC), chamado popularmente de Provão, que era aplicado em todos os estudantes concluintes de campos de conhecimento predefinidos (Verhine, Dantas & Soares, 2006). O ENC tinha como objetivo classificar os diferentes cursos de graduação de acordo com o desempenho dos seus respectivos alunos, assim como detectar eventuais falhas na sua formação, de acordo com os princípios expressos nas diretrizes curriculares. Contudo, em seus fundamentos apresentava vários problemas metodológicos discutidos

[1] Endereço para correspondência:
Universidade São Francisco – Laboratório de Avaliação Psicológica e Educacional (LabAPE)
Mestrado e Doutorado em Psicologia
Rua Alexandre Rodrigues Barbosa, 45 – 13251-900 – Itatiba – São Paulo
E-mail: rprimi@mac.com

em outros trabalhos, podendo-se citar, por exemplo, o fato de mensurar somente os concluintes e suas interpretações serem referenciadas à norma (Landeira-Fernandez & Primi, 2002; Primi, Landeira-Fernandez & Ziviani, 2003).

Com a mudança de governo o sistema de avaliação passou por uma discussão e revisão, de forma a surgir, em 2003, o Sistema Nacional de Avaliação da Educação Superior (SINAES), que inclui uma abordagem diferente para o exame de cursos, denominado Exame Nacional de Desempenho dos Estudantes (ENADE), na tentativa de superar as limitações do ENC (Limana & Brito, 2006; Polidori, Marinho-Araujo, & Barreyro, 2006). Trata-se de um exame composto por questões referentes à formação geral e específica, elaborado com o objetivo de aferir as habilidades acadêmicas e as competências profissionais desenvolvidas pelos estudantes ingressantes e concluintes das Instituições de Educação Superior (IES), bem como colher informações relativas às características socioeconômicas dos estudantes selecionados, por intermédio de procedimentos de amostragem.

Diferente do Provão, que avaliava somente o desempenho dos concluintes, o ENADE busca comparar o desempenho entre ingressantes e concluintes de um determinado curso de ensino superior, pretendendo avaliar o que este agrega aos estudantes durante sua formação acadêmica (Verhine, Dantas & Soares, 2006). Na literatura científica, a medida da qualidade das instituições ou da eficácia escolar é baseada em medidas de valor agregado entendido como uma medida do progresso ou mudança ocorrida em um determinado tempo em que o aluno foi exposto a um ambiente educacional. O progresso médio dos alunos de uma instituição comparado ao progresso médio total é parte da definição do efeito escola (Brandão, 2000; Ferrão, 2003; Jesus & Laros, 2004; Raudenbush, 2004a, 2004b; Rubin, Stuart, & Zanutto, 2004; Soares, Ribeiro, & Castro, 2001).

Por um lado, as medidas de valor agregado são baseadas em medidas longitudinais equalizadas de forma a possibilitar medir o crescimento em nível individual. Por outro, no desenho do ENADE as medidas são transversais, isto é, compostas de amostras de estudantes, ingressantes e concluintes no mesmo ano. Assim, assumindo-se que o nível de desempenho dos concluintes era, na época em que ingressaram, similar ao desempenho dos alunos ingressantes no ano corrente, então a diferença das médias entre ingressantes-concluintes, ainda que obtidas no mesmo ano, são tidas como um indicador de mudança de um aluno médio. Portanto, no ENADE se tem uma comparação de ingressantes e concluintes no nível do curso e não do aluno.

Como não possui dados longitudinais dos estudantes, o ENADE tenta produzir uma aproximação de uma medida da qualidade de curso no cálculo do Indicador de Diferença entre os Desempenhos Observados e Esperados (IDD). O IDD tem o propósito de trazer às instituições informações comparativas dos desempenhos de seus estudantes concluintes em relação aos resultados obtidos, em média, pelas demais instituições cujos perfis de seus estudantes ingressantes são semelhantes. Para isso, calcula-se um valor esperado para o desempenho médio do concluinte a partir do desempenho médio do ingressante, a proporção de estudantes cujos pais têm nível superior de escolaridade, e a razão entre o número de estudantes concluintes e o de ingressantes. Essas variáveis são selecionadas a fim de prover um melhor ajuste para as estimativas realizadas com base no perfil do ingressante. Assim, pressupõe-se que o IDD é uma boa aproximação do que seria considerado efeito do curso (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP, 2006).

Questões metodológicas e psicométricas sobre os sistemas de avaliação

Embora o ENADE tenha avançado bastante em relação ao Provão, ele também está sujeito a muitos questionamentos do ponto de vista metodológico (Primi & cols.a, 2009; Verhine, Dantas & Soares, 2006). Alguns deles estão relacionados com a limitação do desenho transversal e as suposições em que se fundamenta que, se não ocorrerem, podem colocar em dúvida as interpretações que são feitas. Outros questionamentos são mais amplos, não só circunscritos ao ENADE, mas aos estudos de valor agregado, especialmente sobre a dificuldade em isolar os efeitos das instituições na aprendizagem do aluno. Tais estudos são, em essência, tentativas de avaliar efeitos causais dos cursos nos alunos, ou seja, avaliar o grau de conhecimento adquirido nessa trajetória por meio do desempenho nos exames propostos. Como afirmam Rubin, Stuart e Zanutto (2004) "o objetivo da literatura sobre valor agregado parece ser estimar os "efeitos causais" de professores ou escolas; isto é, determinar quanto um professor (ou escola) "adiciona valor" aos escores dos testes de seus estudantes. Está implicado que os efeitos sendo estimados são efeitos causais: o efeito nos estudantes deles estarem em uma escola A (ou com o professor T)" (p. 2).

A aprendizagem, que está na base do que é avaliado nesses sistemas, é multideterminada e está condicionada às influências de características extra-escolares prévias (Ferrão, 2003; Formiga, 2004; Soares, 2004) como influências sociais, econômicas e culturais, no nível cognitivo (conhecimento e raciocínio) que os alunos já têm ao ingressar nas escolas e, também, extra-escolares concomitantes, isto é, aos ambientes socioculturais em que os alunos estão inseridos ao longo do curso universitário e que também exercem influência em sua aprendizagem. O desafio metodológico desses estudos, portanto, está em separar a influência do curso de outras variáveis confundidoras, certamente importantes, mas irrelevantes na construção de um indicador de avaliação da qualidade dos cursos (Primi & cols.b, 2009). O padrão ouro de estudos para inferências mais seguras das diferenças de desempenho, como refletindo o efeito da escola ou curso sobre os alunos, é o controle experimental envolvendo alocação aleatória e controle de variáveis independentes, entre outros. Mas evidentemente isso não é possível em situações como a do ENADE. Assim, a questão passa a ser quais recursos metodológicos de delineamento e análise podem ajudar a superar essas limitações permitindo investigar relações de causa-efeito dadas as limitações de delineamento (Campbel, 1969).

Outro aspecto questionável do ENADE tem relação com a prova propriamente dita, isto é, seu conteúdo. Especialmente, questões relativas a sua validade em avaliar as competências profissionais e habilidades acadêmicas conforme planejado e quais os parâmetros referenciais para sua interpretação (Primi, Hutz & Silva, no prelo; Primi & cols., 2009). Tais procedimentos analíticos de validade de construto de provas são uma prática comum em psicologia, embora talvez não tão difundidos na área educacional. Todas as inferências realizadas posteriormente, como o valor agregado e a qualidade dos cursos, são baseadas em uma prova, de forma que a questão crucial se relaciona diretamente com a qualidade dos instrumentos.

Funcionamento diferencial do item

Diante dessas questões complexas que circundam os sistemas de avaliação em larga escala, um projeto foi proposto com o objetivo de analisar a validade do ENADE, realizando estudos que investigam problemas básicos desses exames, como os que foram levantados nos parágrafos anteriores, decorrentes das limitações metodológicas (Primi, 2006). Neste artigo, em especial, se apresenta parte dos estudos, ligados a esse projeto, focalizando-se questões sobre a validade das provas, tomando os dados do ENADE de psicologia. Dentre vários aspectos de validade, um dos pontos centrais tem relação com o funcionamento diferencial do item (*Differential Item Functioning*, DIF). Como foi apontado, o ENADE avalia ingressantes e concluintes com a mesma prova e, em nível de curso, constrói um indicador a partir das diferenças de média, em cada curso, entre ingressantes e concluintes (diferença entre um valor esperado e um observado médio). Ora, para se construir esse índice é necessário que a prova avalie da mesma forma os dois grupos. Como o ENADE pretende avaliar conhecimentos, competências profissionais e habilidades acadêmicas desenvolvidas ao longo do curso, pode-se questionar se esta prova é igualmente adequada para avaliar alunos no início do curso e no final.

Geralmente, provas (e testes) são compostas por itens que operacionalizam medidas em diferentes níveis de complexidade de uma dimensão latente. Quando agregados, permitem estimar a habilidade de, uma pessoa nessa dimensão. A análise psicométrica dos itens verifica a sua eficácia em medir a suposta dimensão latente. Sob a ótica da Teoria de Resposta ao Item (TRI), os itens são caracterizados por parâmetros e índices de ajuste, os quais permitem verificar sua qualidade. Os principais parâmetros são a dificuldade "*b*", isto é, o nível de complexidade do construto que o item representa e a discriminação "*a*" que está associada à correlação do escore no item com o construto medido pelo teste. A análise do DIF verifica em que medida esses parâmetros permanecem os mesmos em diferentes grupos de pessoas, garantindo que o item funciona da mesma maneira e, portanto, mede o construto da mesma forma.

Propriedade da invariância dos parâmetros na TRI

Para entender o conceito de DIF, é preciso ter em mente uma importante decorrência da TRI, a invariância dos parâmetros. É mais fácil visualizar essa propriedade se considerarmos o parâmetro de dificuldade. Ao estimar a dificuldade, diferentemente da Teoria Clássica dos Testes (TCT), que estima um único índice de acerto para todo o grupo, a TRI se baseia na Curva Característica do Item (CCI), que pode ser compreendida como índices de acerto condicionados aos níveis de habilidades dos sujeitos (teta ou θ), isto é, probabilidades de acerto para subgrupos de pessoas com o mesmo nível de habilidade. Na TRI (especialmente nos modelos de 1 e 2 parâmetros), a dificuldade corresponde à habilidade referente a 50% de chance de se acertar o item.

Na Figura 1 são apresentadas duas CCI's calibradas pelo *software* WINSTEPS implementando o modelo de 1 parâmetro da TRI, o modelo de Rasch.

Para tanto, foi utilizado um item de um teste para avaliação do raciocínio analógico. Nos gráficos, as curvas modeladas das CCI's (linhas mais grossas nas figuras) representam a probabilidade de acerto (eixo Y) em razão da habilidade (eixo X). Pode-se observar que o aumento da habilidade associa-se ao aumento da probabilidade de acerto. Além disso, há também um conjunto de pontos ligados por uma linha mais fina, representando a curva observada.

A CCI A foi calculada a partir de um grupo de 260 pessoas, que foi subdividido em sete grupos com níveis de habilidade crescentes. A curva observada representa os índices de acerto dentro de cada grupo (os pontos na figura). A CCI A modelada (linha mais grossa) tenta representar o mais fielmente possível a curva observada. Como se verifica, o modelo representa os dados reais sem muitas discrepâncias. A figura da direita representa a CCI B, do mesmo item, mas agora estimada por outro grupo de sujeitos, diferentes dos primeiros, composto por 237 pessoas.

Os dois grupos têm distribuições muito diferentes de habilidades, o primeiro contendo mais sujeitos com baixa habilidade (a proporção de acertos global desse segundo, ou índice de dificuldade na TCT é ID=35,2%); e o segundo, alta habilidade (ID=77,8%), conforme demonstrado nas distribuições de habilidade dos dois grupos na parte inferior da Figura 1. Examinando-se a curva observada percebe-se que, no caso da CCI A, está mais para a esquerda da escala de habilidade, enquanto na CCI B, à direita. Apesar das diferenças nas duas amostras, o índice de dificuldade estimado pelo modelo de Rasch, isto é, o theta correspondente à probabilidade de 50% de acerto, é praticamente o mesmo nas duas amostras, -0,68 e -0,79, respectivamente. Embora diferentes, os valores estão dentro da margem de erro da estimação, que é de 0,12.

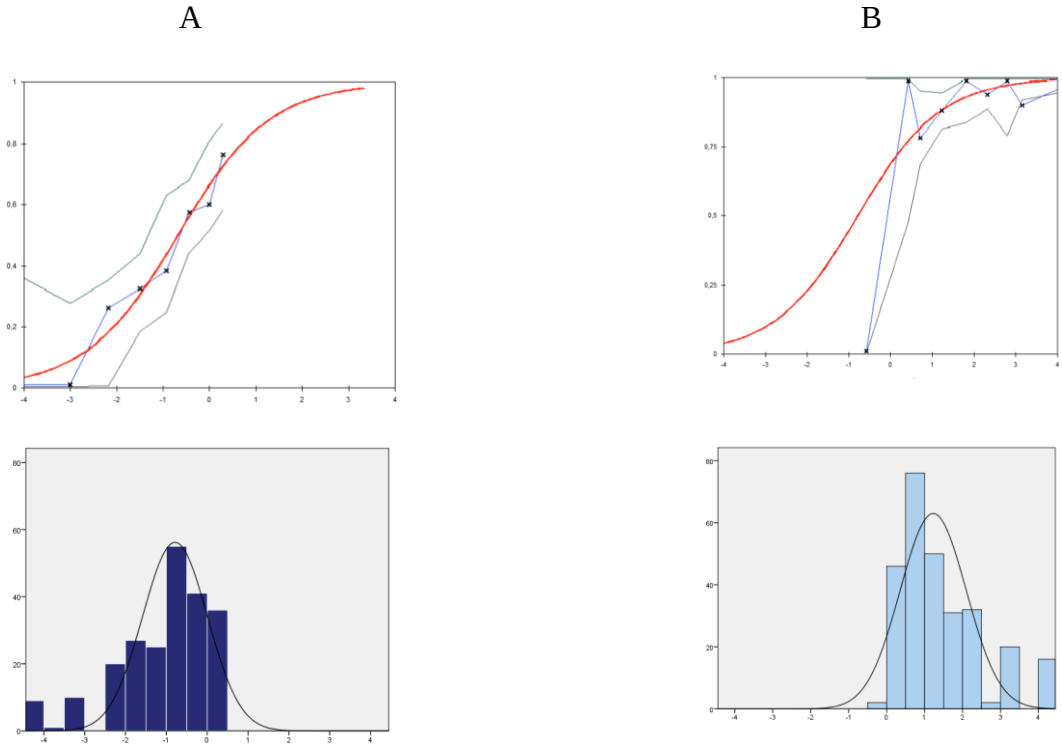

Figura 1 – Curva característica de um item estimada a partir de dois grupos com habilidades distintas

Este exemplo, então, demonstra o princípio geral da invariância dos parâmetros da TRI. Embora os grupos tenham distribuições diferentes, com o condicionamento das probabilidade de acertos às habilidades, as duas curvas ficam semelhantes permitindo-se a estimação de um parâmetro de dificuldade invariante quanto à amostra utilizada. Intuitivamente é possível entender como esse condicionamento funciona: embora os grupos sejam diferentes quanto à proporção de pessoas em cada nível de habilidade, a CCI é construída calculando-se a probabilidade de acerto para seções desse grupo com habilidades semelhantes. Como a dificuldade é definida como o nível de habilidade associado à probabilidade

de 50% de acerto, ela será sempre a mesma. Desse modo, obtêm-se curvas semelhantes, embora "construídas" a partir de setores diferentes da escala de habilidade.

DIF e suas explicações

Como os parâmetros estimados pela TRI devem ser invariantes em relação à habilidade da amostra, quando isso não ocorre, isto é, se há diferenças acentuadas (acima do que seria esperado pelos erros de estimação) nos parâmetros dos itens quando calculados a partir de grupos distintos, observa-se uma situação anômala chamada de DIF. Portanto, o DIF é definido como a observação de probabilidades de acerto substancialmente diferentes, para pessoas de um grupo em relação a outro, mas que tenham a mesma habilidade, isto é, a observação, em pessoas com a mesma habilidade, de uma chance diferenciada de acerto favorecendo um dos grupos (Masters, 1988; Smith, 2004; Sisto, 2006b; Zumbo, 1992, 2007).

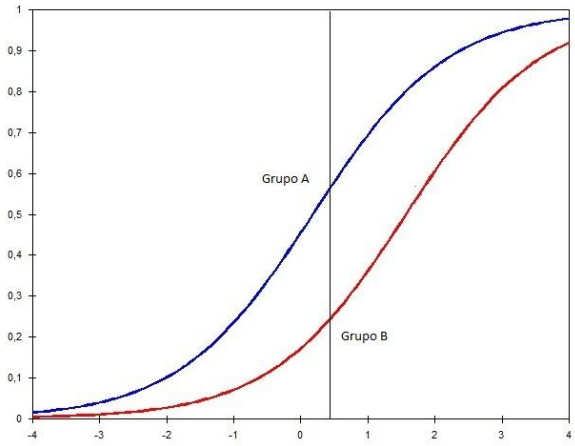

Figura 2 – Curvas características de um item representando DIF estimadas a partir de dois grupos

A Figura 2 exemplifica um item com DIF. Nela estão demonstradas as CCI's de um mesmo item calibradas em dois grupos, A e B, resultando em *b's* de 0,18 e 1,57, isto é, dificuldades com uma diferença substancial de 1,39 *logits*, magnitude superior ao erro de estimação de 0,13. Nesse caso o item é mais difícil para o grupo B. Como mostrado na figura, pessoas do grupo A, com habilidade 0,4, têm aproximadamente 2,5 (0,55/0,22) vezes mais chance de acertar o item do que as pessoas do grupo B, ainda que possuam a mesma habilidade.

Cabe ressaltar a distinção entre DIF e diferenças reais entre grupos, que na literatura é chamada de impacto (Ackerman, 1992). Podem existir diferenças reais entre grupos que indicam diferenças no construto avaliado ou podem existir diferenças causadas por DIF que não são diferenças no construto, mas sim em fatores secundários, muitas vezes não relevantes. Como exemplificado na Figura 1, os dois grupos de fato tinham diferentes distribuições de habilidade, então, a probabilidade total de acertos de cada grupo era díspar, sugerindo impacto. Mas, ao consideramos pessoas dos diferentes grupos com mesma habilidade, estas terão a mesma probabilidade de acerto, pois naquele exemplo não havia DIF. Já na Figura 2, mesmo depois de controlar o nível de habilidade, ainda existem diferenças, visto que pessoas com a mesma habilidade, que supostamente deveriam ter a mesma chance de acertar o item, diferem. Essa situação caracteriza a ocorrência do DIF. A probabilidade total de acerto dos dois grupos será diferente em razão do DIF.

Situações nas quais o DIF ocorre estão relacionadas à multidimensionalidade. Idealmente, a probabilidade de acerto deveria variar somente em função da dimensão principal, isto é, do construto mensurado no teste, que é representada no eixo horizontal X. A linha vertical (Figura 2), cruzando o eixo X no nível 0,4, indica que para um mesmo nível de theta a probabilidade de acerto varia em razão do grupo. Essa variação não pode estar associada à dimensão principal, pois os sujeitos têm o mesmo theta. Então, a variação deve estar associada a uma segunda dimensão, na qual os grupos devem diferir, o que acaba por provocar índices desiguais de acerto (Ackerman, 1992; Smith, 2004).

Por exemplo, um item medindo compreensão de texto, contendo gírias regionais de Minas Gerais,

aplicado a um grupo A, de mineiros, com um mesmo nível geral de compreensão de textos que pessoas de um grupo B, paulistas, teria uma vantagem por causa do conhecimento léxico das palavras usadas. Mesmo que se tenham pessoas com níveis iguais de compreensão de textos, que é o fator principal medido pelo teste, os mineiros teriam vantagem pela maior habilidade no segundo fator (conhecimento do léxico regional). Assim, uma segunda dimensão irrelevante, cuja distribuição é diferente nos dois grupos, é a causa do DIF.

DIF, viés e validade

A partir disso, a presença de DIF pode potencialmente indicar violação da unidimensionalidade, condição básica para a aplicação da TRI, sugerindo que uma segunda dimensão está interferindo além daquela que se imagina que o teste está medindo. Conforme afirmam Nunes e Primi (2010)

> *um estudo DIF procura verificar se pessoas com o mesmo nível de habilidade, mas de grupos distintos, têm probabilidades de acerto diferentes ao item. Se essas pessoas possuem a mesma habilidade, não importa que grupo façam parte, deveriam ter a mesma chance de acertar o item. Se isso não ocorre há presença de DIF e isso pode afetar outros parâmetros psicométricos do teste, nomeadamente os parâmetros normativos o que potencialmente pode gerar viés favorecendo certos grupos e prejudicando outros.* (p. 121)

Portanto, o estudo de DIF está diretamente relacionado ao estudo de viés nos testes, já que, se a segunda dimensão é irrelevante e influencia os resultados do teste favorecendo certos grupos, ela pode comprometer a equidade na avaliação. Alguns grupos podem ser beneficiados em razão de construtos secundários não-relevantes. Portanto, quando se encontram diferenças entre grupos nos escores globais de um teste, é preciso, antes de tudo, verificar se essas diferenças decorrem de alguns itens com DIF. Caso isso ocorra, as diferenças podem não ser reais no construto (impacto), mas diferenças em aspectos secundários irrelevantes ao propósito principal do instrumento. Em termos de validade, Nunes e Primi (2010) classificam como um problema básico de indicação de validade da estrutura interna:

> *pois a presença de DIF indica que uma segunda dimensão tem uma importância não negligenciável e os subgrupos com características distintas têm notas diferentes nessa segunda dimensão que, se não tratados, se confundem com o resultado da primeira. Portanto, os estudos de DIF verificam as influências que uma segunda dimensão, especialmente relacionada a subgrupos compostos por variáveis socioeconômicas distintas, têm nos itens do instrumento, alterando sua dificuldade. Assim, quando somados podem produzir diferenças no escore entre esses grupos que não se relacionam de fato ao construto medido (dimensão principal). Portanto, a presença de DIF pode indicar uma alteração estrutural, especialmente nas complexidades dos itens, em relação a algum subgrupo específico da amostra para a qual se pretende usar o teste, gerando vieses na interpretação.* (p. 121)

Embora se possa compreender DIF como uma questão genérica de estrutura interna, ele é mais complexo, pois está relacionado à multidimensionalidade com diferenças de grupo em dimensões irrelevantes, que ficam confundidas na nota do teste afetando parâmetros normativos favorecendo certos grupos. Por isso, diz respeito ao grau em que o teste atinge a propriedade ideal de invariância, medindo da mesma forma diferentes grupos de sujeitos. É importante mencionar que esses estudos não revelam a causa do DIF e também se há de fato viés, isto é, se há valorização de um grupo em decorrência de algum fator irrelevante. O DIF é uma condição necessária mas não suficiente para o viés (Zumbo, 1999). Após se detectar o DIF, deve-se investigar com mais profundidade suas possíveis causas, se configuram em um viés, suas implicações e formas de corrigir sua influência.

Em síntese, por um lado, estudos de DIF são importantes na validação de instrumentos, especialmente quando se encontram diferenças entre grupos segundo suas características globais como gênero, experiências culturais, entre outras, já que tais diferenças podem não ser verdadeiras. Por outro lado, esses estudos podem mostrar aspectos mais substanciais do construto, pois indicam como os itens que operacionalizam sua medida interagem com características peculiares a certos grupos.

Já se pode notar na literatura nacional a incorporação de estudos DIF na criação de instrumentos (Andriola, 2000, 2001; Karino, Laros & Jesus, no prelo; Rueda, 2007; Sisto, 2006a; Sisto, Bartholomeu, Santos, Rueda & Suehiro, 2008). Em sistemas de avaliação em larga escala os estudos DIF assumem uma importância muito grande, dadas as implicações que tais medidas possuem para a gestão pública. Também nessa área verifica-se a existência de estudos nacionais (Aguiar, 2010; Soares, Gamerman & Goncalves, 2007).

Em face da relevância do ENADE no contexto atual, no que concerne à aferição das habilidades acadêmicas e das competências profissionais desenvolvidas pelos estudantes das IES, o

objetivo do presente estudo foi investigar o DIF tentando verificar a equivalência das questões do ENADE de 2006 do curso de psicologia para medir estudantes em diferentes momentos de sua formação (ingressante e concluinte) e de diferentes sistemas de ensino (público e privado). Ao mesmo tempo, pretendeu ilustrar o uso da regressão logística na análise DIF. Este estudo espera trazer uma contribuição metodológica para a validade do ENADE, que enfrenta o desafio de construir provas para uso em diferentes grupos de alunos (ingressantes e concluintes) de uma ampla variedade de contextos, já que o exame é aplicado nacionalmente em todos os sistemas de ensino.

Método

Participantes

Este estudo analisou os dados do banco de dados do ENADE 2006 de psicologia. O banco contém uma amostra representativa de 26.613 estudantes de psicologia ingressantes (N=12940) e concluintes (N=10673) de todas as regiões do país. Desses, 19.859 são mulheres (84,1%) e a idade variou de 16 a 78 anos (69,2% da amostra entre 16 e 26 anos) com M=26,1 e DP=8,43 tendo os ingressantes M=24,7 (DP=8,56) e concluintes 27,8 (DP=8,0). Desses, 16,3% são provenientes de instituições públicas (federais, estaduais e municipais) e 83,7 de instituições privadas.

Material

A prova do ENADE 2006 de psicologia é composta por 30 questões de componente específico e 10 questões de Formação Geral. Das questões específicas 26 são em formato de múltipla escolha e quatro em formato de respostas dissertativas. Os dados psicométricos da prova foram apresentados em Primi e cols (no prelo). Dos 30 itens da prova do componente específico, dois foram eliminados pela comissão por possuírem problemas na formulação. Os 28 itens restantes foram submetidos à análise de Rasch para calibração dos parâmetros pelo programa WINSTEPS (Linacre, 2010). Considerando os critérios técnicos discutidos na literatura (índices de ajuste, precisão, índice de separação para a medida dos estudantes e análise fatorial dos resíduos) concluiu-se que os dados indicam um bom ajuste dos valores calibrados pelo modelo com os dados observados.

Procedimentos

Analisou-se o DIF dos itens considerando-se duas variáveis de interesse: momento do curso (ingressantes ou concluintes) e tipo de administração da instituição (sistema privado ou público). Empregou-se dois procedimentos para a análise do DIF: (a) comparação dos índices de dificuldade estimados separadamente para cada grupo que é o procedimento padrão no WINSTEPS (Linacre, 2010), e (b) por meio da regressão logística (Zumbo, 1999). A análise de DIF a partir da regressão logística é feita em três passos consecutivos conforme os modelos:

Modelo 1: (Condiciona a habilidade) $$\ln\left[\frac{P_i}{(1-P_i)}\right] = b_0 + b_1 Teta$$

Modelo 2: DIF uniforme (variações da dificuldade) $$\ln\left[\frac{P_i}{(1-P_i)}\right] = b_0 + b_1 Teta + b_2 Grupo$$

Modelo 3: DIF não-uniforme (variações da discriminação) $$\ln\left[\frac{P_i}{(1-P_i)}\right] = b_0 + b_1 Teta + b_2 Grupo + b_3 (Grupo \times Teta)$$

No primeiro passo (Modelo 1) se prevê o acerto ao item a partir da habilidade dos sujeitos (Teta). No segundo passo (Modelo 2), acrescenta-se uma variável *dummy* (Grupo, 1 para concluintes e 0 para ingressantes, por exemplo), testando DIF uniforme, isto é, se, para além da habilidades dos sujeitos, o fato do estudante pertencer a um dos grupos afeta a probabilidade de acerto. No terceiro passo (Modelo 3) acrescenta-se o termo de interação originado do produto Grupo X Teta. Esse termo testa o DIF não-uniforme, isto é, se, para além da habilidade global, o fato do aluno ser ingressante ou concluinte e com determinado nível de habilidade afeta a probabilidade de acerto. Nesse último termo testa-se DIF associado

ao parâmetro de discriminação. Como se pode perceber, o primeiro passo "condiciona" a habilidade. Assim, no segundo passo é testada, além da habilidade, se a variável grupo também afeta a probabilidade de acerto. Se afetar, então há uma segunda dimensão atrelada ao grupo contribuindo para a previsão do acerto ao item.

Para o teste da significância do DIF, considerou-se o valor do contraste, isto é, a diferença entre os índices de dificuldade dividida pelo erro padrão conjunto das estimativas, que resulta em um teste *t* ao qual já estão atrelados os valores da significância (t >=1,96 é significativo a p<0,05). Na regressão logística, para cada passo, considerou-se o Qui-Quadrado, testando se os parâmetros incluídos diferem significativamente de zero (para gl=1, χ^2>=6,64 é significativo a 0,01).

Como acontece na presente pesquisa, em amostras muito grandes, diferenças muito pequenas, negligenciáveis do ponto de vista prático (alteração que o item poderá gerar nas medidas dos sujeitos), são estatisticamente significativas. Assim, além da significância, os critérios incluem a magnitude mínima a partir da qual se pode julgar que o DIF tem o potencial de alterar as medidas. No manual do WINSTEPS se considera um contraste mínimo de 0,5 e significativo como indicativo de DIF (Linacre, 2010). Na regressão logística, Zumbo (1999) propõe que o Qui-Quadrado seja significativo pelo menos a 0,01 e o valor do R^2 seja pelo menos 0,13. Aliás, uma das vantagens da regressão logística é a obtenção dessa estatística, que pode ser interpretada como uma medida tradicional de magnitude do efeito (Shimizu & Zumbo, 2005; Zumbo, 1999).

Resultados e discussão

A questão central investigada nas análises era se a dificuldade dos itens, estimada pelo modelo de um parâmetro da TRI, seria diferente em subgrupos de alunos ingressantes e concluintes (variável momento) ou quando provenientes de IES públicas e privadas (categoria administrativa). Na Tabela 1 apresentamos os resultados considerando o momento do curso e na Tabela 2 o tipo de instituição. Essas tabelas fornecem, na primeira coluna, o nome do item, sendo que CE indica componente específico e o número, colocado na sequência indica o número do item na prova oficial (a numeração começa em 11, pois os itens de 1 a 10 são da prova de Formação Geral). Em seguida, aparece um código associado ao conteúdo (mais informações em Primi & cols., no prelo). A segunda e terceira colunas apresentam os resultados do WINSTEPS, sendo o DIF o contraste $b_{\text{ingressante}} - b_{\text{concluinte}}$, na Tabela 1, e $b_{\text{IES pública}} - b_{\text{IES privada}}$ na Tabela 2, isto é, quando positivo, indica que o item foi mais fácil para estudantes igualmente hábeis concluintes ou de IES privadas, e quando negativo, o inverso. Em seguida é apresentado o teste *t*. Nas colunas restantes são apresentados os resultados da regressão logística. Nas colunas quatro a seis estão colocados os Qui-Quadrados relativos, respectivamente, aos modelos 1, 2 e 3 (ver a seção Procedimentos em Método). Deve-se ressaltar que os Qui-Quadrados apresentados para os modelos 2 e 3 referem-se somente aos parâmetros adicionados testando-se o DIF uniforme e não-uniforme, respectivamente, e não o valor total do modelo. Nas últimas colunas, de sete a nove, são apresentados o R^2 de Nagelkerke. Novamente, para os modelos 2 e 3 (M2 e M3) apresenta-se somente o valor adicionado ao Modelo 1, que a inclusão de cada parâmetro incrementa (grupo em M2 e grupo X teta em M3).

Tabela 1 – Resultados da análise DIF para ingressantes e concluintes (Continua)

Item	DIF	*t*	χ^2 *M1*	χ^2 *M2*	χ^2 *M3*	R^2 *M1*	R^2 *M2*	R^2 *M3*
ce11_FndBas_his	-0,36	-12,50	1319,9	86,5	0,1	0,076	0,005	0,000
ce12_FndBas_prof	0,27	5,54	5466,0	37,9	44,6	0,362	0,003	0,002
ce13_FndBas_sist	-0,48	-16,10	512,8	53,0	0,2	0,030	0,003	0,000
ce14_FndBas_sig	-0,22	-7,34	883,6	2,6	49,7	0,053	0,000	0,003
ce15_FndBas_sensocom	**-0,71**	**-23,40**	**509,5**	**242,0**	**0,5**	**0,031**	**0,014**	**0,000**
ce16_Mtdmed_Bas_psicom	**-0,80**	**-27,00**	**171,4**	**209,5**	**3,2**	**0,010**	**0,013**	**0,000**
ce17_Mtdmed_Bas_dados	-0,37	-13,10	1504,8	112,9	1,5	0,085	0,006	0,000
ce18_Mtdmed_Bas_corr	-0,31	-10,70	2345,2	113,8	4,9	0,128	0,006	0,000
ce19_Procbas_Bas_mem	-0,10	-3,26	3912,1	82,9	6,3	0,210	0,004	0,000
ce20_Procbas_Bas_desado	0,12	2,92	5560,2	69,0	11,3	0,334	0,003	0,001

Tabela 1 – Resultados da análise DIF para ingressantes e concluintes (Continuação)

Item	DIF	*t*	χ^2 *M1*	χ^2 *M2*	χ^2 *M3*	R^2 *M1*	R^2 *M2*	R^2 *M3*
ce21_Procbas_Bas_desinf	-0,12	-3,83	3909,5	101,5	3,8	0,216	0,006	0,000
ce22_Procbas_Bas_psicopat	**1,25**	**39,25**	**5594,2**	**1118,8**	**119,7**	**0,283**	**0,049**	**0,005**
ce23_Procbas_Bas_saudoenc	**0,70**	**23,77**	**4747,5**	**301,6**	**56,7**	**0,243**	**0,014**	**0,002**
ce24_Procbas_Bas_inclusao	-0,37	-12,60	794,8	36,2	15,8	0,047	0,002	0,001
ce25_Procbas_Bas_represoc	0,28	9,07	5986,8	4,4	11,5	0,308	0,000	0,001
ce26_Procbas_Bas_psican	-0,20	-7,24	3053,5	102,6	1,9	0,162	0,005	0,000
ce27_Procbas_Bas_aprendiz	-0,46	-12,90	393,1	42,6	19,4	0,028	0,003	0,001
ce28_Intrfc_Bas_gestao	-0,16	-5,71	1826,8	14,5	1,7	0,101	0,000	0,001
ce29_Intrfc_Bas_locultura	0,17	5,84	3889,8	0,2	8,1	0,204	0,000	0,001
ce30_Intrfc_Bas_neuroc	**-0,50**	**-15,30**	**426,4**	**57,1**	**8,0**	**0,028**	**0,003**	**0,001**
ce31_Intrfc_Bas_intelgen	-0,06	-1,99	4539,9	109,8	3,3	0,234	0,005	0,000
ce32_PratcEsc	0,17	5,59	1384,6	79,8	17,9	0,082	0,004	0,001
ce33_PratcTrab	0,20	6,38	4180,8	0,4	0,2	0,222	0,000	0,000
ce34_PratcSaud	-0,13	-4,22	1223,0	0,0	60,7	0,072	0,000	0,003
ce35_PratcDiag	-0,13	-4,72	1861,1	4,8	10,1	0,101	0,000	0,001
ce36_PratcGrp	0,31	9,83	5609,7	0,0	16,7	0,291	0,000	0,001
ce37d_Pratc_Etic	0,05	5,86	4595,4	49,8	0,8	0,336	0,004	0,000
ce38d_Mtdmed_Bas	0,03	3,08	2016,1	5,4	4,9	0,348	0,001	0,001
ce39d_Pratc_Esc_Tarb	0,14	13,84	1932,6	2,1	6,5	0,235	0,000	0,001
ce40d_Pratc_Clinc	0,00	0,00	4887,9	9,0	7,1	0,342	0,000	0,001

Como pode ser notado, na Tabela 1, cinco itens (destacados em negrito) atingiram o critério mínimo para DIF, em relação aos ingressantes e concluintes, segundo os limites sugeridos por Linacre (2010). Embora a regressão logística tenha demonstrado maior magnitude para esses mesmos itens, indicando uma convergência dos métodos de análise, nenhum deles atingiu o critério mínimo sugerido por Zumbo (1999) de 0,13.

Dos cinco itens, três favoreceram os ingressantes e dois favoreceram os concluintes. Na Figura 3 são apresentadas as CCI's dos itens CE16 e CE22 com maior magnitude de DIF representando esses dois padrões opostos. É interessante notar que os itens com DIF favorecendo os ingressantes foram considerados de má qualidade (Primi & cols., no prelo) e geralmente apresentam correlação item total baixa. Dois deles (CE16 e CE30) foram questionados, sugerindo-se a anulação, sendo que um foi efetivamente anulado (CE16). Uma hipótese explicativa é que como os concluintes têm um conhecimento e vocabulário mais precisos e técnicos do que os ingressantes, talvez tenham tido mais dificuldade em entender esses itens e, por isso, erram mais pois interpretam alternativas diferentes das consideradas corretas no gabarito oficial como possíveis. No caso da questão CE30, por exemplo, além da B, a alternativa D também é escolhida por pessoas com capacidade mais alta. No item CE16 não era possível distinguir claramente a alternativa certa das demais.

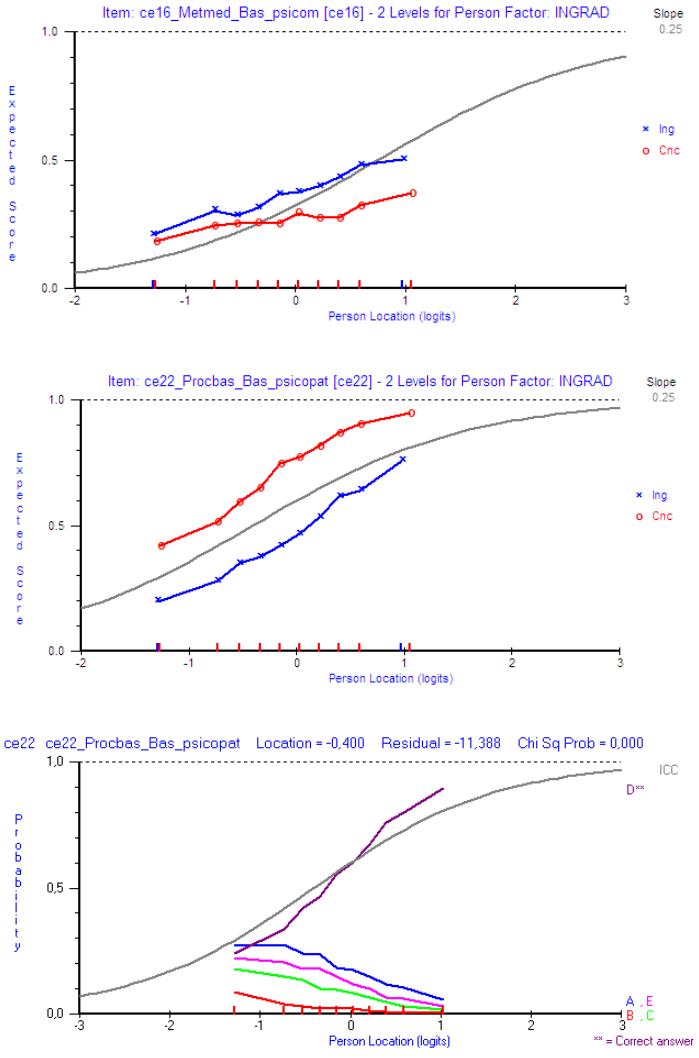

Figura 3 - Dois itens extremos com efeitos opostos CE16 mais fácil para ingressantes (superior) e CE22 mais fácil para concluintes (meio) e curva original CE22 (inferior)

As duas questões que favoreceram concluintes têm propriedades psicométricas mais adequadas e tratam de conhecimentos específicos de psicopatologia e do problema da medicalização da educação. É interessante notar que questões similares, quanto a conteúdos específicos, foram identificadas como mais válidas na análise dos itens do Exame Nacional de Cursos de Psicologia realizada por Landeira-Fernandez e Primi (2002). Essas duas questões também contribuem para um terceiro subfator (incluindo itens de processos básicos, psicopatologia, aprendizagem e fundamentos e métodos de análise de dados em investigações científicas) identificado na análise fatorial dos itens dessa prova feita por Primi e cols. (no prelo). Essa convergência de achados com diferentes métodos é coerente com a explicação de que o DIF está associado ao problema da uni ou multidimensonalidade da prova (Ackerman, 1992; Nunes & Primi, 2010).

Outro fato interessante de se notar é que as curvas empíricas desses itens tendem a ser mais discriminativas que o modelo de Rasch prevê (ver CCI inferior da Figura 3). Isso ilustra o argumento de Masters (1988), expresso no título de seu artigo "quando mais se torna pior", de que itens com alta discriminação podem, no fundo, ser sinal de DIF. Isso ocorre quando a segunda dimensão está correlacionada com a primeira. Assim sua influência se soma à da primeira dimensão, fazendo o item parecer mais discriminativo. Nesses itens sujeitos com alta habilidade têm ainda uma vantagem a mais em razão de uma segunda dimensão mas que pode ser irrelevante (Masters, 1993).

Tabela 2 – Resultados da análise DIF para IES públicas e privadas

Item	Dif	*T*	χ^2 *M1*	χ^2 *M2*	χ^2 *M3*	R^2 *M1*	R^2 *M2*	R^2 *M3*
ce11_FndBas_his	-0,18	-4,03	1295,9	8,469	69,198	0,077	0,000	0,004
ce12_FndBas_prof	-0,43	-4,52	5391	75,364	264,52	0,366	0,004	0,016
ce13_FndBas_sist	**0,71**	**14,17**	**506,07**	**197,77**	**65,716**	**0,031**	**0,012**	**0,004**
ce14_FndBas_sig	-0,12	-2,59	858,74	6,885	114,76	0,053	0,000	0,007
ce15_FndBas_sensocom	0,10	2,11	517,87	5,064	156,03	0,032	0,000	0,010
ce16_Mtdmed_Bas_psicom	0,39	8,11	169,4	65,527	175,68	0,01	0,004	0,011
ce17_Mtdmed_Bas_dados	-0,38	-8,33	1480,4	41,8	119,7	0,086	0,002	0,007
ce18_Mtdmed_Bas_corr	-0,11	-2,15	2320,4	12,419	270,21	0,13	0,000	0,015
ce19_Procbas_Bas_mem	-0,45	-8,04	3866,9	1,688	146,3	0,213	0,000	0,007
ce20_Procbas_Bas_desado	**-0,57**	**-6,92**	**5500**	**15,04**	**193,53**	**0,338**	**0,001**	**0,011**
ce21_Procbas_Bas_desinf	-0,36	-5,94	3836,2	7,036	207,09	0,218	0,000	0,011
ce22_Procbas_Bas_psicopat	-0,46	-8,96	5493,3	11,402	14,028	0,285	0,000	0,001
ce23_Procbas_Bas_saudoenc	**-0,55**	**-11,10**	**4641,6**	**40,776**	**26,903**	**0,244**	**0,002**	**0,001**
ce24_Procbas_Bas_inclusao	-0,45	-9,90	788,64	75,006	163,37	0,048	0,004	0,010
ce25_Procbas_Bas_represoc	**-1,03**	**-15,50**	**5866,2**	**98,378**	**116,87**	**0,31**	**0,004**	**0,006**
ce26_Procbas_Bas_psican	-0,46	-9,33	2997,4	22,439	115,92	0,163	0,001	0,006
ce27_Procbas_Bas_aprendiz	-0,35	-6,87	402,82	53,354	118,47	0,029	0,004	0,009
ce28_Intrfc_Bas_gestao	-0,20	-4,31	1799,7	3,559	109,97	0,102	0,000	0,006
ce29_Intrfc_Bas_locultura	**-0,93**	**-18,90**	**3831,9**	**238,84**	**44,659**	**0,206**	**0,012**	**0,002**
ce30_Intrfc_Bas_neuroc	0,32	6,16	432,57	30,1	51,293	0,029	0,002	0,003
ce31_Intrfc_Bas_intelgen	**-1,06**	**-18,80**	**4454,7**	**206,62**	**136,44**	**0,235**	**0,010**	**0,006**
ce32_PratcEsc	0,21	4,42	1369,1	26,27	5,316	0,083	0,001	0,001
ce33_PratcTrab	-0,12	-2,22	4114,7	16,776	86,727	0,224	0,001	0,004
ce34_PratcSaud	0,05	1,18	1215,4	4,145	73,997	0,073	0,000	0,004
ce35_PratcDiag	0,37	8,24	1861,6	135,88	95,987	0,104	0,007	0,005
ce36_PratcGrp	**-0,53**	**-9,12**	**5497**	**4,872**	**93,19**	**0,293**	**0,000**	**0,004**
ce37d_Pratc_Etic	0,08	6,18	4448,8	8,477	0,195	0,334	0,001	0,000
ce38d_Mtdmed_Bas	0,00	0,00	1980,3	2,003	0,005	0,349	0,001	0,000
ce39d_Pratc_Esc_Tarb	0,07	4,80	1898,8	3,313	3,467	0,235	0,001	0,000
ce40d_Pratc_Clinc	0,22	18,20	4722,7	40,358	3,952	0,34	0,003	0,000

Como pode ser notado, na Tabela 2, sete itens (destacados em negrito) atingiram o critério mínimo para DIF em relação às instituições públicas e privadas, e mais uma vez, somente atingindo os critérios sugeridos por Linacre (2010). Desses itens seis favorecem estudantes das instituições públicas, sendo mais difíceis para estudantes de IES privadas, mesmo considerando alunos com nível de habilidade semelhante. Desses itens, cinco se referem à influência da cultura e do contexto na formulação de explicações para fenômenos psicológicos (desenvolvimento adolescente, medicalização de problemas sociais na escola, a influência da cultura na definição de transtornos mentais e a importância do contexto nas práticas grupais), e um deles se relaciona à base genética do comportamento.

É interessante notar que todos esses seis itens são mais discriminativos, caso fosse presentemente utilizado o modelo de dois parâmetros. Portanto, mais uma vez, se observa que a dimensão secundária está associada à primeira e por isso produz maior discriminação. Isso ocorre porque os alunos de universidades públicas tiveram desempenho mais alto no ENADE. Os itens com DIF favorecem esses estudantes, por isso, há a correlação entre a dimensão secundária e a primária. O único item que favorece estudantes de instituições privadas, referindo-se à abordagem histórico crítica, teve baixa discriminação, indicando a existência de alternativas confusas.

Considerações finais

Este estudo procurou verificar a equivalência de questões do ENADE de 2006 do curso de psicologia para medir estudantes em diferentes momentos de sua formação e sistemas de ensino. A partir disso, pode-se concluir que a prova é equivalente para esses grupos testados? Os dados mostraram que um número reduzido de itens apresentou DIF. Também se verificaram dois tipos de DIF que conduzem a respostas diferentes.

Um grupo menor de itens com DIF apresentou problemas de formulação identificados pela análise de seu conteúdo, feita pela comissão da área de psicologia, e também pela baixa discriminação. Esses itens tendem a favorecer os grupos com menor habilidade. Uma possível explicação, do fator secundário associado a esses itens, é o processamento mais superficial e simplificado, com interpretação dos conceitos baseada na acepção mais comum, não-técnica. Ingressantes teriam esse tipo de processamento, pois ainda não possuem conhecimento mais aprofundado. Já os concluintes, como têm maior conhecimento, tenderiam a ter um processamento relativamente mais complexo e a se confundir mais, pois, como os itens estão mal formulados, a alternativa correta não está claramente diferenciada das outras, existindo, portanto, ambiguidades que poderiam conduzir ao erro. Por um lado, como esses itens favorecem pessoas com baixa habilidade, eles introduzem ruído na avaliação e deveriam ser eliminados da prova. Em outras oportunidades já se havia notado esse problema nas provas de psicologia (Pasquali, 2002; Primi, Landeira-Fernandez & Ziviani, 2003; Ziviani & Primi, 2002). Por outro, como eles são em número pequeno, o efeito que têm na prova tende a ser baixo.

Outro grupo de itens mais numeroso apresentou DIF favorecendo grupos com habilidade mais alta e apresentam propriedades psicométricas mais adequadas. O conteúdo desses itens foi avaliado como adequado e medindo conhecimentos mais específicos. Analisou-se que o fator secundário está correlacionado com o fator principal medido pela prova, isto é, está ajudando a diferenciar os sujeitos na mesma direção da primeira dimensão. Conforme argumentado em Landeira-Fernanez e Primi (2002), nesse tipo de prova, itens válidos devem ser capazes de diferenciar concluintes de ingressantes, mas em menor grau ingressantes entre si, já que pretendem avaliar o que se desenvolve ao longo do curso. Assim, esse segundo tipo de DIF consiste em um fator secundário que diferencia concluintes de ingressantes, portanto, produzindo informação válida. Para esses casos, pode-se questionar até que ponto a presença de DIF produz um viés desfavorável para validade das inferências realizadas a partir das pontuações na prova.

Uma explicação do fator secundário envolvido se relaciona com a oportunidade de aprendizagem, por parte dos concluintes, do conteúdo específico dos itens, em relação aos ingressantes. Entretanto, há uma maior quantidade de itens com esse tipo de DIF, isto é, favorecendo alunos da rede pública, o que sugere uma oportunidade de aprendizagem diferenciada também para esses estudantes. Assim, o julgamento da relevância do viés apontado pelo DIF passa pela análise do conteúdo desses itens para determinar se tratam de componente curricular importante para todos os estudantes ou a uma ênfase mais específica. No primeiro caso, não teríamos viés, mas no segundo sim, pois a prova poderia beneficiar uma abordagem curricular de alguns cursos.

Um ponto paralelo, mas relevante, se relaciona a um certo paradoxo nesses dados. Se a dimensão secundária se refere a um conhecimento mais específico e válido, o que a primeira dimensão estaria medindo? Apesar do presente estudo não possibilitar a resposta a essa pergunta, ele sugere que, quanto mais a prova for válida, maior a quantidade de DIF e, consequentemente, mais difícil a criação de uma medida comparável entre ingressantes e concluintes.

A presença de DIF significa que, mesmo considerando pessoas com o mesmo nível de habilidade a partir do escore global na prova, certos subgrupos terão uma vantagem adicional nos itens com DIF. Uma implicação disso é que a previsão do modelo pela TRI, para esses itens, estará mais sujeita a erros. Nesse sentido, o DIF afeta a adequação do modelo em representar os dados. Contudo, há alternativas metodológicas para se tentar lidar com esse problema, tais como a resolução do DIF. Mas a principal questão tem a ver com a validade de conteúdo, já que essa análise irá fornecer informações para julgar se o DIF implica ou não viés. Em caso afirmativo, seria ainda necessário aplicar procedimentos de resolução do DIF e recalcular a habilidade dos sujeitos para verificar se os resultados ficariam substancialmente diferentes.

Em resposta à questão da equivalência da prova, pode-se concluir que há um número pequeno de itens com potencial para produzir viés. Esse estaria mais associado aos sistemas de ensino do que ao momento de formação. Ainda assim, o julgamento do viés precisa ser feito com base em uma discussão mais ampla do conteúdo dos itens com DIF e da análise da influência que esses itens têm na medida global obtida a partir das pontuações dos estudantes ao responderem a prova.

Referências

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Aguiar, G. S. (2010). O funcionamento diferencial do item (DIF) como estratégia para captar ênfases curriculares diferenciadas em matemática. *Estudos em Avaliação Educacional*, *21*(45), 169-190.

Andriola, W. B. (2000). funcionamento diferencial dos itens (DIF): estudo com analogias para medir o raciocínio verbal. *Psicologia. Reflexão e Crítica*, *13*(3), 473-481.

Andriola, W. B. (2001). Descrição dos principais métodos para detectar o funcionamento diferencial dos itens (DIF). *Psicologia. Reflexão e Crítica*, *14*(3), 643-652.

Brandão, Z. (2000). Fluxos escolares e efeitos agregados pelas escolas. *Em aberto, 17*(71), 41-48.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, *24*(4), 409-429.

Ferrão, M. E. (2003). *Introdução aos modelos de regressão multinível em educação.* Campinas: Komedi.

Formiga, N. S. (2004). O tipo de orientação cultural e sua influência sobre os indicadores do rendimento escolar. *Psicologia: Teoria e Prática, 6*(1), 13-29.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP (2006). Resumo Técnico ENADE 2005. Ministério da Educação. Brasília (DF). Obtido em 22 de novembro de 2010 do World Wide Web: http://www.inep.gov.br/download/enade/2005/Resumo_Tecnico_ENADE_2005.pdf>.

Jesus, G. R. & Laros, J. A. (2004). Eficácia escolar: regressão multinível com dados de avaliação em larga escala. *Avaliação Psicológica, 3*(2), 93-106.

Karino, C. A., Laros, J. A. & Jesus, G. R. (no prelo). Funcionamento diferencial dos itens do Teste Não-Verbal de Inteligência SON-R 2,5-7. *Psicologia: Teoria e Pesquisa*.

Landeira-Fernanez, J. & Primi, R. (2002). Comparação do desempenho entre calouros e formandos no Provão de Psicologia 2000. *Psicologia: Reflexão e Crítica, 15*(1), 219-234.

Limana, A. & Brito, M. R. F. (2006). O modelo de avaliação dinâmica e o desenvolvimento de competências: algumas considerações a respeito do ENADE. Em D. Ristoff, A. Limana M. R. F. Brito (Orgs.). *ENADE: perspectiva de avaliação dinâmica e análise de mudanças.* (pp. 17-44). Brasília: INEP/DAES/MEC.

Linacre, J. M. (2010). *Winsteps® (Version 3.70.0)* [Computer Software]. Beaverton, Oregon: Winsteps.com. Obtido em 1 Janeiro de 2010 do World Wide Web: http://www.winsteps.com/.

Masters, G. N. (1988). Item discrimination: when more is worse. *Journal of Educational Measurement*, *25*(1), 15-29.

Masters, G. N. (1993). Undesirable item discrimination. *Rasch Measurement Transactions, 7*(2), 289.

Nunes, C. H. S. S. & Primi, R. (2010). Aspectos técnicos e conceituais da ficha de avaliação dos testes psicológicos. Em Conselho Federal de Psicologia – CFP (Org.). *Avaliação psicológica: diretrizes na regulamentação da profissão* (pp. 101-128). Brasília: CFP.

Pasquali, L. (2002). Provão (ENC) de Psicologia 2000 e 2001: Análise dos parâmetros psicométricos. Em R. Primi (Org.). *Temas em avaliação psicológica* (pp. 152-178). Campinas: Instituto Brasileiro de Avaliação Psicológica.

Polidori, M. M., Marinho-Araujo, C. M. & Barreyro, G. B. (2006). *SINAES: perspectivas e desafios na avaliação da educação superior brasileira. Ensaio: avaliação de políticas públicas em educação*. Rio de Janeiro: Fundação Cesgranrio, 425-436.

Primi, R. (2006). *A validade do ENADE para avaliação da qualidade dos cursos de instituições de ensino superior*. Projeto de Pesquisa. Itatiba: Universidade São Francisco, LabAPE.

Primi, R., Hutz. C. S. & Silva, M. C. R. (no prelo). *A prova do ENADE de Psicologia 2006: concepção, construção e análise da prova*.

Primi, R., Landeira-Fernandez, J. & Ziviani, C. (2003). O provão de psicologia: objetivos, problemas, conseqüências e sugestões. *Psicologia: Teoria e Pesquisa, 19*(2), 109-116.

Primi, R., Nunes, C. H. S. S., Silva, M. C. R., Carvalho, L. F., Miguel, F. K. & Vendramini, C. M. M. (2009a). Aplicação da Teoria de Resposta ao Item na Interpretação das Notas do ENADE de Psicologia. *Revista de Educação AEC*, *38*, 115-124.

Primi, R., Silva, M. C. R., Bartholomeu, D., Vendramini, C. M. M., Nunes, C. H. S. S. & Mata, A. S. (2009b). Questões metodológicas referentes ao Exame Nacional de Desempenho dos Estudantes (ENADE). *Revista de Educação AEC*, *38*, 125-134.

Raudenbush, S. W. (2004a). *Schooling, statistics, and poverty: can we measure school improvement?* New Jersey: Educational Testing Service.

Raudenbush, S. W. (2004b). What are value-added models of estimating and what does this imply for

statistical practice. *Journal of Educational and Behavioral Statistics, 29*(1), 121-129.

Rubin, D. B., Stuart, E. A. & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29,* 103-116.

Rueda, F. J. M. (2007). O funcionamento diferencial do item no Teste Pictórico de Memória. *Avaliação Psicológica*, *6*(2), 229-237.

Shimizu, Y. & Zumbo, B. D. (2005). A logistic regression for differential item functioning primer. *Japan Language Testing Association Journal*, *7*, 110-124.

Sisto, F. F. (2006a). O funcionamento diferencial dos itens. *Psico-USF*, *11*(1), 35-43.

Sisto, F. F. (2006b). Estudo do funcionamento diferencial de itens para avaliar o reconhecimento de palavras. *Avaliação Psicológica*, *5*(1), 1-10.

Sisto, F. F., Bartholomeu, D., Santos, A. A. A., Rueda, F. J. M. & Suehiro, A. C. B (2008). Funcionamento diferencial de itens para avaliar a agressividade de universitários. *Psicologia. Reflexão e Crítica*, *21*(3), 349-356.

Smith, R. M. (2004). Detecting item bias with the Rasch model. Em: E. C. Smith Jr. & R. M. Smith (Orgs.). *Introduction to rasch measurement* (pp. 391-418). Maple Grove, Minesota: JAM Press.

Soares, J. F. (2004). O efeito da escola no desempenho cognitivo de seus alunos. *Revista Electrónica Iberoamericana sobre Calidad, Eficácia y Cambio em Educación, 2*(2), 83-104.

Soares, J. F., Ribeiro, L. & Castro, C. M. (2001). Valor agregado de instituições de ensino superior em Minas Gerais para os cursos de Direito, Administração e Engenharia Civil. *Dados, 44*(2) 363-396.

Soares, T. M., Gamerman, D. & Gonçalves, F. B. (2007). Análise bayesiana do funcionamento diferencial do item. *Pesquisa Operacional* 27(2) 271-291.

Verhine, R. E., Dantas, L. M. V. & Soares, J. F. (2006). Do provão ao ENADE: uma análise comparativa dos exames nacionais utilizados no ensino superior brasileiro. *Ensaio. Avaliação e Políticas Públicas em Educação*, *14*(52), 291-310.

Ziviani, C. & Primi, R. (2002). Teoria de resposta ao item e o modelo Rasch de mensuração: uma análise do provão de Psicologia. Em R. Primi (Org.). *Temas em avaliação psicológica* (pp. 131-151). Campinas: Instituto Brasileiro de Avaliação Psicológica.

Zumbo, B. D. (1992). Statistics and research design in the behavioral sciences – a review. *Educational and Psychological Measurement, 52*, 787-794.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and Likert-Type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three generations of dif analyses: considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223-233.

Recebido em agosto de 2010
Reformulado em setembro de 2010
Aprovado em novembro de 2010

Sobre os autores:

Ricardo Primi é psicólogo pela PUCCampinas, doutor em Psicologia Escolar e do Desenvolvimento Humano pela Universidade de São Paulo com parte desenvolvida na Yale University (EUA) sob orientação de Robert J. Sternberg. Em 2009 foi professor visitante na Univeristy of Toledo (Ohio, EUA) com bolsa Fulbright/CAPES É professor associado do Programa de Pós-Graduação em Psicologia da Universidade São Francisco.

Fabiano Koich Miguel é psicólogo pela Universidade Presbiteriana Mackenzie, mestrado e doutorado em Avaliação Psicológica na Universidade São Francisco. É professor adjunto da Universidade Estadual de Londrina.

Lucas Francisco de Carvalho é psicólogo pela Universidade Presbiteriana Mackenzie, mestre e doutorando em Psicologia com ênfase em Avaliação Psicológica pela Universidade São Francisco.

Marjorie Cristina Rocha da Silva é psicóloga, mestre e doutoranda em Psicologia com ênfase em Avaliação Psicológica pela Universidade São Francisco e docente das Faculdades Integradas Einstein de Limeira (FIEL).

Agências financiadora: Esse trabalho foi produzido a partir do financiamento do Edital Observatório da Educação CAPES/INEP e do CNPq.