Escala de Avaliação da Fadiga: funcionamento diferencial dos itens em regiões brasileiras[1]

Thiago Medeiros Cavalcanti, Romulo Lustosa Pimenteira de Melo
Universidade Federal da Paraíba, João Pessoa-PB, Brasil
Emerson Diogenes De Medeiros
Universidade Federal do Piauí, Teresina-PI, Brasil
Layrtthon Carlos De Oliveira Santos, Valdiney Veloso Gouveia[2]
Universidade Federal da Paraíba, João Pessoa-PB, Brasil

RESUMO

Este estudo teve como objetivo conhecer, por meio da Teoria de Resposta ao Item, os parâmetros psicométricos e o Funcionamento Diferencial dos Itens (DIF) da Escala de Avaliação da Fadiga (EAF). Participaram 7.008 médicos com média de idade de 47,3 anos (*DP*=11,33), a maioria do sexo masculino (62,4%). Os parâmetros dos itens foram estimados por meio do modelo de Resposta Gradual de Samejima, e o DIF por meio da Razão de Verossimilhança, da estatística Pseudo-R2 e das diferenças entre os pesos das regressões (Δβs). Os resultados mostraram boa variação dos limiares de resposta (*j*) e uma boa discriminação (*a*). Quanto ao DIF, mesmo considerando os critérios mais rigorosos indicados na literatura, os resultados apontaram para um efeito não significativo da região na resposta aos itens. Concluindo, mostrou-se que a EAF provavelmente não apresenta DIF, corroborando sua utilização no contexto brasileiro.

Palavras-chave: fadiga; teoria de resposta ao item; escala.

ABSTRACT – Fatigue Assessment Scale: Diferential item functioning in Brazilian regions

This study aimed to understand, through the Item Response Theory, the psychometric parameters and Differential Item Functioning (DIF) of the Fatigue Assessment Scale (FAS). There were 7,008 physicians participating, with a mean age of 47.3 years (*SD*=11.33), the majority being male (62.4%). We estimated the items' parameters through the Samejima Graded Response Model and the DIF by likelihood ratio, the pseudo-R2 statistics and the differences between the weights of the regressions (Δβs). Results showed good variation of the response thresholds (*j*) and good discrimination (*a*). Concerning the DIF, even considering the strictest criteria set forth in the literature, the findings showed a non-significant effect of the region in response to the items. In conclusion, probably the FAS does not have DIF, supporting its use in the Brazilian context.

Keywords: fatigue; item response theory; scale.

RESUMEN – Escala de Evaluación de la Fatiga: funcionamiento diferencial de los ítems en regiones brasileñas

Este estudio tuvo como objetivo conocer, por medio de la Teoría de Respuesta al Ítem, los parámetros psicométricos y el Funcionamiento Diferencial de los Ítems FDI) la Escala de Evaluación de la Fatiga (EEF). Participaron 7.008 médicos con edad media de 47.3 años (*DT*=11.33), la mayoría de sexo masculino (62.4%). Los parámetros de los ítems han sido estimados por medio del modelo de Respuesta Gradual de Samejima, y el DIF por medio de la Razón de Verosimilitud, de la estadística Pseudo-R2 y de las diferencias entre los pesos de las regresiones (Δβs). Los resultados mostraron buena variación de los umbrales de respuesta (*j*) y una buena discriminación (*a*). Con respecto al DIF, los hallazgos indicaron un efecto no significativo de la región en la respuesta a los ítems. Concluyendo, la EAF probablemente no presenta DIF, corroborando su utilización en el contexto brasileño.

Palabras clave: Fatiga; Teoría de Respuesta al Ítem; Escala.

A fadiga pode se apresentar como exaustão, cansaço ou letargia, isto é, estado em que acontece a diminuição do nível de consciência, caracterizando a sonolência e apatia por parte do indivíduo. De forma geral, configura-se em duas categorias: 1. física, compreendendo a inabilidade de manter funcionamento normal de suas atividades cotidianas; é notável em exercícios nos quais os indivíduos sentem uma queimação excessiva dos músculos; e

[1] O presente estudo contou com apoio do CNPq por meio de bolsa de Produtividade em Pesquisa concedida ao autor Valdiney Gouveia. Os autores agradecem a esta instituição.

[2] Endereço para correspondência: Universidade Federal da Paraíba, CCHLA – Departamento de Psicologia, 58051-900, João Pessoa-PB. *E-mail:* vvgouveia@gmail.com / *Site:* http://vvgouveia.net

2. mental, que surge em forma de lassidão, esgotamento e marasmo, o que reflete em dificuldades em tomar decisões e realizar avaliações (Moore & Gastão Neto, 2013). Nessa direção, Mills, Young, Pallant, e Tennant (2010) classificam a fadiga como a presença de um cansaço persistente e recorrente, que não é necessariamente causado com exclusividade por exercícios, não se eliminando imediatamente com o descanso físico, podendo ter como consequência a redução de níveis de atividades laboral, social e pessoal.

De modo geral, a fadiga é relacionada a diversos problemas de saúde, como esclerose múltipla (Alvarenga Filho, Carvalho, Dias, & Alvarenga, 2010), doença de Parkinson (Havlikova et al. 2008), lúpus, doenças autoimunes (Pereira & Duarte, 2010) e câncer (Lawrence, Kupelnick, Miller, Devine, & Lau, 2004). Como consequência, tem-se a diminuição da participação em atividades necessárias para a reabilitação dos pacientes, interferindo na qualidade de vida das pessoas que a apresentam (Sanches & Cardoso, 2012).

Fini e Cruz (2010) identificaram no Brasil a prevalência de fadiga também na população geral; seus resultados apoiaram uma incidência maior em mulheres, pessoas mais velhas e que moram sozinhas. Os mesmos autores ainda indicaram a existência de duas medidas de fadiga validadas no Brasil (DUFS e DEFS), sendo que ambas consideraram amostras de participantes com cuidados em nível primário ou pessoas com câncer. No entanto, em busca realizada no *Index Psi* usando a expressão-chave "escala de avaliação da fadiga", foram encontrados 37 resultados. Especificamente, no *PePSIC*, foram identificados três artigos, um tratando da validação da Escala de Avaliação da Fadiga para profissionais da saúde (Gouveia, et al., 2015); no *SciELO* foram observados 34 registros, porém apenas um referente a estudos psicométricos (Pereira & Duarte, 2010), mas sem especificar o contexto em que a medida foi adaptada.

A Escala de Avaliação da Fadiga (EAF) se apresenta como uma das medidas mais parcimoniosas e de fácil aplicação. Elaborada inicialmente por Michielsen, de Vries, van Heck, van de Vijver, e Sijtsma (2004) no contexto holandês, reunindo dez itens que, teoricamente, formam uma estrutura unidimensional, avaliando tanto a fadiga física quanto a psicológica, em coerência com a literatura (Moore & Gastão Neto, 2013). Essa medida já foi adaptada para outros contextos culturais, como o alemão (Hinz, Fleischer, Brähler, Wirtz, & Bosse-Henck, 2011). No Brasil, a primeira tentativa de adaptá-la foi realizada por Oliveira, Gouveia, Peixoto, e Soares (2010), que a traduziram e efetuaram sua validação semântica para logo executarem dois estudos empíricos. No primeiro, fizeram uma análise de componentes principais (CP), que revelou um único componente cuja consistência interna (alfa de Cronbach) foi de 0,80, obtendo índices de escalabilidade de 0,33 (*H*) e 0,82 (*Rho*). O segundo estudo replicou o anterior, confirmando a estrutura unidimensional; seu α foi de 0,85, apresentando os seguintes índices de escalabilidade: H=0,40 e Rho=0,87. O coeficiente de congruência entre as soluções fatoriais desses estudos foi 0,99. Portanto, comprovaram-se evidências de sua validade fatorial e consistência interna.

Apesar de estudos prévios terem focado nos parâmetros psicométricos da EAF, avaliando sua validade e precisão (e.g. Gouveia et al., 2015; Oliveira et al., 2010), estes não avaliaram mais criteriosamente seus itens. Assim, considerando os avanços com a Teoria de Resposta ao Item (TRI), pareceu justificável o intento de novos estudos que possam melhor avalizá-la. Esse aspecto motivou o presente estudo, que teve como objetivo conhecer os parâmetros psicométricos e o Funcionamento Diferencial dos Itens da Escala de Avaliação da Fadiga por meio da Teoria de Resposta ao Item, considerando as cinco regiões geopolíticas do Brasil. O intuito de realizar tais análises é comprovar a pertinência de um instrumento que possa avaliar fadiga independentemente da região geográfica, favorecendo comparações mais adequadas das pessoas. A propósito, os estudos a respeito têm sido apenas no Nordeste, quer com estudantes universitários (Oliveira et al., 2010) ou profissionais da saúde (Gouveia et al., 2015). Portanto, o presente estudo expande o uso da EAF para todo o Brasil, além de apresentar mais evidências de sua qualidade psicométrica.

Funcionamento diferencial do item

Na elaboração de escalas que avaliam construtos latentes, costuma-se discutir acerca do Funcionamento Diferencial dos Itens (DIF) em função da pertença a determinado grupo social ou cultural, principalmente com os testes cognitivos, que revelavam aptidões diferentes para pessoas de etnias distintas. Nesse mesmo contexto, passou-se a perceber que o conteúdo dos testes trazia estímulos que eram estranhos à cultura de algumas populações, sugerindo que as diferenças entre os grupos não se deviam exclusivamente à aptidão, mas poderiam ser também explicadas por um viés cultural que os itens possuíam (Pasquali, 2007).

Ficou evidente, então, que inclusive pessoas com a mesma aptidão ou magnitude no traço latente apresentavam probabilidades diferentes de escolher ou acertar determinado item (Somerville, 2012). No caso, o item foi compreendido de forma diferente, em função do grupo que pertence o participante (Mukherjee, Gibbons, Kristjansson, & Crane 2013), sendo que esse item tem potencial de produzir uma polarização entre os grupos (Makransky & Glas, 2013). Portanto, em razão do DIF, a medida pode privilegiar alguns grupos em detrimento de outros, causando falsas diferenças detectadas e, em consonância, distorcendo o processo avaliativo.

As primeiras técnicas para detectar DIF parecem não ter surtido o efeito pretendido, sendo que uma das

dificuldades foi encontrar um consenso entre os parâmetros que o indicariam (Wang, Tay, & Drasgow, 2013). Atualmente, dentre as principais técnicas, estão a de Angoff, baseada na Teoria Clássica dos Testes (TCT), e aquelas que se fundamentam na Teoria de Resposta ao Item (TRI) (Pasquali, 2007).

A técnica de Angoff também é conhecida como *delta-plot* (plotagem delta), a qual consiste em avaliar o parâmetro de dificuldade dos itens de um mesmo teste, aplicado a grupos diferentes. Na prática, essa técnica é desenvolvida em quatro passos: 1. obter a dificuldade de cada item (parâmetro *p*) para cada grupo; 2. transformar esse escore de dificuldade no escore padronizado z; 3.transformar o escore z no escore delta; e 4. os escores deltas são plotados em coordenadas cartesianas. A maior crítica feita a essa técnica é que ele não considera o poder discriminativo do item e a probabilidade de o item ser acertado ao acaso (Pasquali, 2007).

Na literatura da Teoria de Resposta ao Item (TRI), os dois grupos a serem analisados são chamados de focal e referência. O primeiro é aquele em que os dados são utilizados com padrão, ou seja, verifica-se o ajuste do grupo referência em relação ao grupo focal. Um item possui DIF pela TRI quando são padronizadas as métricas da aptidão entre os grupos e depois são calculadas as Curvas Características dos Itens (CCIs) para cada grupo. Caso exista sobreposição entre as CCIs, conclui-se que o item não apresenta DIF (Pasquali, 2007; Wang et al., 2013).

Uma informação importante fornecida pela TRI é estabelecer se o DIF é uniforme ou não. O item apresenta DIF uniforme quando não existe interação entre o atributo medido e a pertença a um grupo, ou seja, a probabilidade de acertar o item é a mesma para todos os níveis de aptidão, independentemente do grupo do qual o participante foi considerado. No caso do DIF não uniforme, existe uma interação entre indivíduos classificados em um mesmo nível de habilidade no atributo medido, e o fato de pertencerem a grupos distintos (e.g., homens e mulheres), indicando que a probabilidade de acertar o item muda em função de pertencer a grupos diferentes (Sisto, 2006).

Em resumo, o pesquisador, ao realizar a DIF, independentemente da técnica utilizada, está interessado em conhecer se os itens de determinado instrumento são invariantes em dois ou mais grupos. Uma vez que em um mesmo país podem existir diversos contextos com populações distintas, a DIF possibilita avaliar estimativas de calibração da medida. Esse procedimento é fundamental, pois permite esclarecer dúvidas quanto aos vieses do instrumento (Embretson & Reise, 2000). Dessa maneira, compreende-se que a DIF acontece quando os valores dos parâmetros psicométricos dos itens apresentam variações com a mudança de padrões populacionais, prejudicando a comparação dos resultados (Hauck-Filho & Teixeira, 2013).

Método

Participantes

Participaram deste estudo 7.008 médicos, em atividade profissional, divididos entre as cinco regiões do Brasil (Sudeste=2.548, Nordeste=2.107, Sul=1.464, Centro-Oeste=458 e Norte=431). Nas cinco regiões, a maioria dos participantes foi do sexo masculino (Sudeste=62,6%, Nordeste=59,0%, Sul=69,1%, Centro-Oeste=66,8% e Norte=59,4%); a média de idade foi de 47,3 anos (*DP*=11,33), variando pouco em razão das regiões Sudeste=47,0 (*DP*=11,30), Nordeste=47,7 (DP=11,32), Sul=47,5 (*DP*=11,62), Centro-Oeste=45,8 (*DP*=10,43) e Norte=48,0 (*DP*=10,93). Tratou-se de uma amostra aleatória, participando aqueles que, quando solicitados, decidiram voluntariamente fazer parte do estudo.

Instrumentos

Os participantes responderam a Escala de Avaliação da Fadiga (EAF; Michielsen et al., 2004) e três perguntas demográficas (idade, sexo e região do país em que reside). A EAF é composta por 10 itens (e.g., "Sinto-me incomodado devido à fadiga", "Fico cansado muito rapidamente"), respondidos em escala de cinco pontos, variando de 1 (Nunca) a 5 (Sempre), devendo o participante indicar como tem se sentido geralmente nos últimos 30 dias. No Brasil, essa escala foi adaptada por Oliveira et al. (2010), os quais oferecem evidências de sua validade fatorial e consistência interna. No caso, ela apresentou uma estrutura unidimensional com confiabilidade aceitável (α>0,70).

Procedimento

A coleta de dados foi realizada de duas maneiras: 1. por meio dos correios; e 2. entrega pessoal de questionários. No primeiro caso, os médicos foram localizados por meio de seu respectivo conselho profissional. Em seguida, um envelope foi encaminhado ao endereço cadastrado de cada profissional, que continha: (a) o questionário, (b) as informações sobre a importância, a finalidade do estudo e a maneira de proceder para respondê-lo; (c) uma solicitação de preenchimento e procedimento para devolução dos questionários; e (d) um envelope selado e sobrescrito para devolução.

No caso dos médicos que responderam o questionário pessoalmente, o contato foi estabelecido por meio do endereço que constava no registro em seu conselho profissional. Na etapa posterior, foi marcado um dia em que seria feita a devolução do questionário preenchido, que era colocado em envelope lacrado e depositados em uma urna sem qualquer identificação. Nas duas formas de coleta, foram seguidos os procedimentos de pesquisa com seres humanos, coerente com o que determina a legislação vigente (Resolução CNS 466/12).

Análise de Dados

Os dados foram tabulados no Microsoft Excel®, sendo exportados para o *software* R versão 2.15.1 (R Core Team, 2012). Foram utilizados os seguintes pacotes nas análises dos dados: *Mokken* (Van der Ark, 2007) para verificar a unidimensionalidade da escala; *ltm* (Rizopoulos, 2006) para estimar os parâmetros da TRI via modelo de Resposta Gradual de Samejima (1969); *lordif* (Choi, Gibbons, & Crane, 2011) juntamente com a interface gráfica em TCL/TK de Ladwig (2012) para detectar DIF e, por fim, o *Rcommander* (Fox, 2005) para as estatísticas descritivas.

A análise Mokken (*Mokken Scale Analisys* – MSA) busca verificar os pressupostos de homogeneidade monotônica e monotonicidade dupla. Para isso, os critérios adotados como indicativos de unidimensionalidade foram os índices de escalonabilidade *H* de Loevinger (*H* para a escala total e *Hs* para cada item), demandando-se valores acima de 0,30, o e *Rho* de Mokken, cujos valores maiores que 0,80 são recomendáveis (Van der Ark, 2007).

Para verificar o ajuste dos dados ao modelo de Samejima, compararam-se por meio da razão de verossimilhança dois modelos, sendo um alternativo com discriminação constante (*Rating Scale Model* de Andrich, 1978) e outro com discriminação variável (Modelo de Samejima). Também foram analisados os valores χ^2 das frequências dos padrões de resposta observados em relação às frequências preditas pelo modelo; Pasquali (2007) sugerem que valores abaixo de 3,0 indicam ajuste adequado.

Com o fim da checagem, observando se a escala apresenta itens com DIF, utilizou-se o método híbrido de Regressão Logística Ordinal/TRI, que utiliza os escores da TRI como critério de correspondência para a regressão. Este oferece uma escolha mais adequada para a detecção de DIF, já que não utiliza a simples soma das pontuações como na Teoria Clássica dos Testes (Crane, Gibbons, Jolley, & van Belle, 2006; Pasquali, 2007; Somerville, 2012). Para realizar a equiparação dos *thetas* dos grupos, foi utilizado o método de purificação. Este consiste em, inicialmente, estimar os itens que provavelmente apresentam DIF e, em seguida, retirá-los da pontuação da variável latente (*theta*), depois, com a nova pontuação, estimam-se novamente os itens com DIF; as purificações são realizadas até que nenhum item com DIF contribua para a pontuação.

Procurando detectar a existência e o tipo de DIF, utilizou-se uma Regressão Logística Ordinal hierárquica de três passos, em que o primeiro verificou se o critério de correspondência (traço latente) predizia a resposta ao item; posteriormente, inseriu-se a variável Grupo juntamente com Critério de correspondência; e, finalmente, no terceiro passo, foram consideradas as duas variáveis anteriores e a interação entre elas (grupo e a variável correspondente). Quando a inserção de uma nova variável explicava significativamente a resposta ao item, sugere-se DIF; porém, como esse critério é vulnerável a grandes amostras, optou-se por comparar as diferenças entre os três passos por meio do teste de razão de verossimilhança (Swaminathan & Rogers, 1990).

Se a diferença entre os modelos 1 e 2 da regressão for significativa, o item possui DIF uniforme; se o modelo 3 se diferenciar significativamente do 2, fala-se em DIF não uniforme. Nos casos em que apenas os modelos 1 e 3 se diferenciam estatisticamente, não é possível definir se o item possui DIF uniforme ou não uniforme. Também foi utilizado o Pseudo-R^2 como estimativa de magnitude das diferenças entre os modelos. Para fins de interpretação, adotou-se o critério de Zumbo (1999), que sugeriu um Pseudo-R^2 menor que 0,13 como irrelevante; entre 0,13 e 0,26 moderado; e maior do que 0,26 como grande (Zumbo, 1999). O último parâmetro utilizado para analisar o DIF foi o $\Delta\beta_1$ $\{\Delta\beta_1 = [(\beta_{1modelo2} - \beta_{2modelo1})] / \beta_{1modelo1}\}$ Crane, van Belle, e Larson (2004), relataram que o $\Delta\beta_1$ menores que 0,1 (10%) indicaria ausência de DIF, mas também sugeriram 0,05 (5%) e até mesmo 0,01 (1%) como critérios mais intensos e menos permissivos.

Resultados

A análise Mokken foi empregada para checar a pertinência da unidimensionalidade da medida. A escala apresentou índices de escalabilidade acima dos indicados pela literatura como o mínimo satisfatório (H=0,54 e *Rho* de Mokken=0,91) (Van der Ark, 2007). Os itens 9 e 2 foram os que apresentaram os maiores Hs (ambos com 0,61), enquanto o item 10 foi o com menor índice (H=0,40). A média correspondente dos itens foi de 0,54 (DP=0,06). A escala apresentou coeficiente de consistência interna (α=0,91) acima do comumente aceito. Esses resultados corroboram a pertinência de tratar a Escala de Avaliação da Fadiga como unidimensional.

Quanto à comparação entre os ajustes dos modelos o de Samejima, no qual são estimados livremente os parâmetros de discriminação dos itens e limiares de resposta, apresentou melhor ajuste se comparado ao modelo alternativo de discriminação constante (*Rating Scale Model* de Andrich, 1978) [log.Lik (9)=-75160,61; p<0,001]. As frequências dos padrões de resposta observados se situaram próximos de zero, com exceção para 0,0006% dos casos, com todos os valores χ^2 abaixo de 3,00, sugerindo bom ajuste (Pasquali, 2007).

Considerando a unidimensionalidade e o ajuste dos dados ao modelo de Resposta Gradual, a Tabela 1 descreve os parâmetros dos itens gerados pela TRI. Para os itens da escala, a média de discriminação foi de 2,16 com baixa variância (DP=0,53). Os itens 1, 2, 5 e 9 apresentaram discriminação acima de 2 e, observando os limiares de resposta (j_i), verificou-se que estes mesmos itens apresentaram boa amplitude do *theta*, sugerindo que a boa discriminação se dá em um amplo contínuo. Além disso, esses itens foram os que proporcionaram maior quantidade de informação, especificamente os itens 2, 9 e 5, pois apresentaram informação no intervalo de -4

a +4 acima dos 9 pontos. Porém, esse comportamento não foi verificado em todos os itens; especificamente, os itens 6, 7 e 8 apresentaram amplitudes dos j_i mais baixas, ficando entre -0,65 / -0,24 no j_1 e 2,28 / 2,36 no j_4. De modo geral, as médias no j_1 foram de -0,85 (DP=0,40) e 2,79 (DP=0,51) no j_4.

Tabela 1
Descrição dos Parâmetros e da Quantidade de Informação dos Itens

	Limiares de resposta				A	Informação		
	j_1	j_2	j_3	j_4		Total	[-4; +4]	
Item1	-1,54	-0,31	0,91	2,16	2,62	8,96	[8,94]	99,72%
Item2	-1,13	0,11	1,13	2,28	2,85	9,77	[9,75]	99,78%
Item3	-1,01	0,56	1,69	3,09	1,75	5,29	[4,98]	94,17%
Item4	-0,63	0,99	2,03	2,83	1,98	5,89	[5,71]	96,95%
Item5	-1,28	-0,02	1,04	2,16	2,73	9,18	[9,16]	99,78%
Item6	-0,65	0,73	1,90	3,19	1,67	4,77	[4,42]	92,58%
Item7	-0,35	1,09	2,33	3,41	1,85	5,47	[5,00]	91,36%
Item8	-0,24	0,90	1,96	2,94	1,98	5,59	[5,37]	96,09%
Item9	-0,10	0,24	1,30	2,36	2,76	9,20	[9,17]	99,66%
Item10	-0,72	1,46	2,62	3,51	1,37	3,63	[3,15]	86,61%
M	-0,85	0,57	1,69	2,79	2,16	6,77	[6,56]	95,67
DP	0,40	0,55	0,57	0,51	0,53	2,24	[2,41]	4,47
Mín	-1,54	-0,31	0,91	2,16	1,37	3,63	[3,15]	86,61
Máx	-0,24	1,46	2,62	3,51	2,85	9,77	[9,75]	99,78

Nota. [-4; +4] Informação no intervalo de -4 a +4 desvios padrões. % - Percentual de informação oferecida pelo item no intervalo de -4 a +4

Função diferencial do item (DIF)

Inicialmente foi feito a equiparação entre os *thetas* de cada subgrupo a saber: 1 – Sudeste; 2 – Nordeste; 3 – Sul; 4 – Centro-Oeste e 5 – Norte. Posteriormente, foram necessárias três interações para a purificação do critério de correspondência, de modo que os itens que apresentassem algum indício de DIF não participassem da pontuação da variável latente, já que ela seria testada quanto a sua capacidade preditiva de resposta aos itens.

A Tabela 2, apresenta os critérios utilizados para detectar o Funcionamento Diferencial dos Itens, o primeiro com a Razão de Verossimilhança, o segundo com as diferenças entre os Pseudo-R^2 e o terceiro com os $\Delta\beta$s resultantes dos modelos das regressões. Pelo teste da Razão das Verossimilhanças, foram encontrados quatro itens com provável DIF (1 (Sinto-me incomodado devido à fadiga), 3 (Não faço muitas coisas durante o dia), 5 (Sinto-me exausto fisicamente) e 8 (Não sinto vontade de fazer nada)).

Os itens 1, 3 e 5 apresentaram significância estatística na comparação entre os modelos 1 e 2 e na comparação entre os modelos 1 e 3, denotando que, caso haja DIF, ele seria uniforme. O item 8 apresentou significância estatística nas três comparações, sugerindo funcionamento diferencial não uniforme dos itens. Porém, quando se observa a diferença entre os valores do teste do Pseudo-R2 para os modelos das regressões, verificou-se que esses valores foram bastante baixos para os quatro itens. O mesmo aconteceu para a diferença entre os parâmetros β_1 das regressões.

Tabela 2
Critérios Utilizados para Detectar o Funcionamento Diferencial dos Itens

	Testes de Razão de Verossimilhança (p)			Diferença entre os Pseudo-R^2			$\Delta(\beta_1)$
	$Pr(\chi^2_{(1;2)})$	$Pr(\chi^2_{(2;3)})$	$Pr(\chi^2_{(1;3)})$	$R^2_{(1;2)}$	$R^2_{(1;3)}$	$R^2_{(2;3)}$	
Item 1	0,0005	0,40	0,0004	0,0017	0,0019	0,0002	0,0040
Item 3	0,0001	0,34	0,0003	0,0013	0,0015	0,0002	0,0017
Item 5	0,0002	0,31	0,0004	0,0018	0,002	0,0002	0,0046
Item 8	0,0012	0,0018	0,0001	0,0010	0,0020	0,0010	0,0005

Nota. $Pr(\chi^2_{(1;2)})$ – Comparação da razão de verossimilhança entre os passos um (01) e dois (02) da regressão. $Pr(\chi^2_{(2;3)})$ - Comparação da razão de verossimilhança entre os passos dois (02) e três (02) da regressão. $Pr(\chi^2_{(1;3)})$ - Comparação da razão de verossimilhança entre os passos um (01) e três (03) da regressão. $R2_{(1;2)}$ – Diferença entre os Pseudo-R2 dos passos um (01) e dois (02) da regressão. $R2_{(1;3)}$ - Diferença entre os Pseudo-R2 dos passos um (01) e três (03) da regressão. $R2_{(2;3)}$ – Diferença entre os Pseudo-R2 dos passos dois (02) e três (03) da regressão.

A Figura 1 sobrepõem as CCIs das cinco regiões para cada um dos quatro itens com suposição de DIF. Os valores presentes nos gráficos se referem respectivamente à discriminação do item e aos quatro limiares de resposta. A descrição dos parâmetros dos itens obedece a seguinte ordem: 1=Sudeste, 2=Nordeste, 3=Sul, 4=Centro-Oeste e 5=Norte.

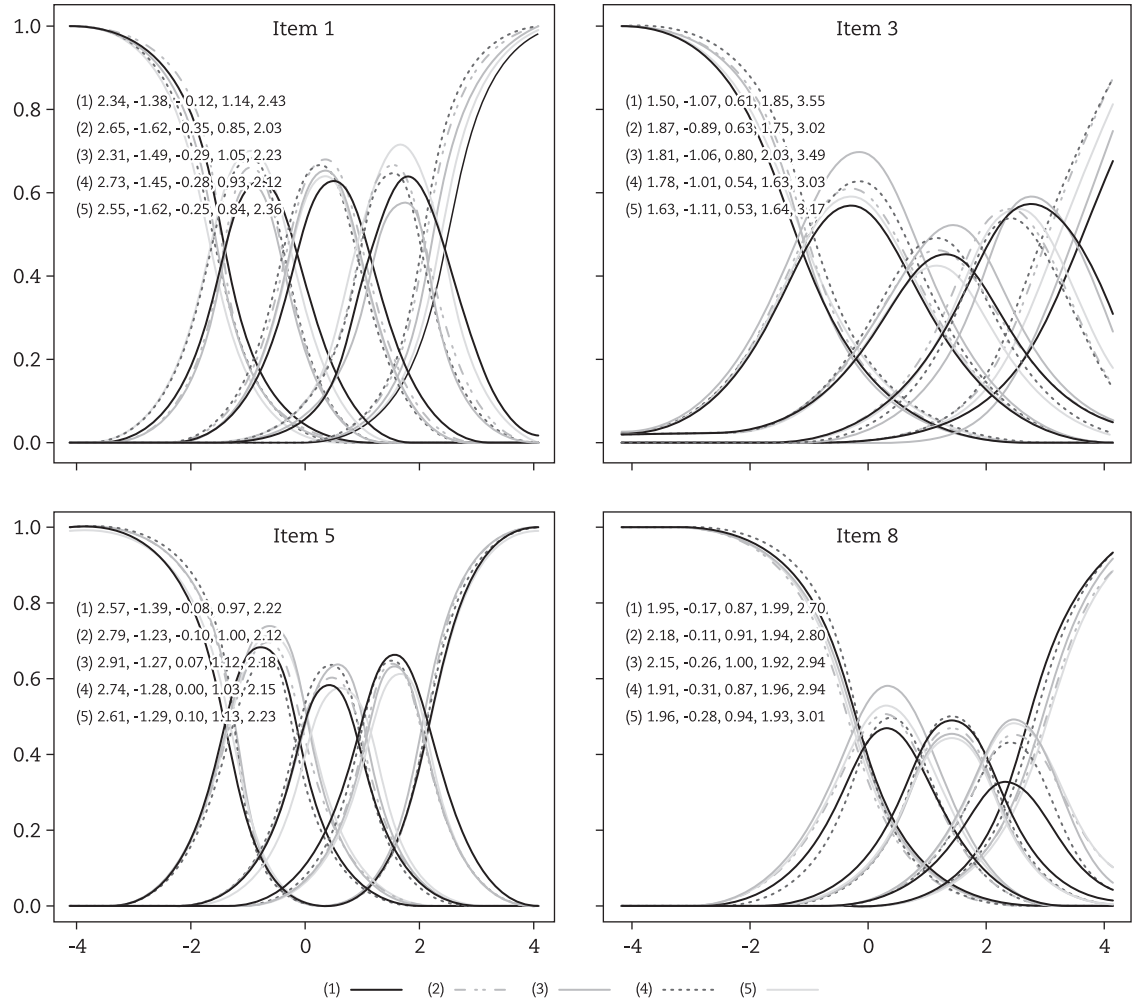

Figura 1. Curvas característica dos itens para as cinco regiões

Por fim, apresenta-se a Figura 2 com o impacto dos itens com presumível DIF nas curvas características do teste. A primeira curva é baseada nos parâmetros de todos os dez itens para cada grupo. A curva da direita é baseada apenas nos parâmetros dos quatro itens que apresentaram indícios de DIF. Percebe-se uma ligeira diferença na curva dos itens com DIF apenas para os sujeitos com *thetas* mais elevados.

Discussão

Este estudo teve como objetivo conhecer os parâmetros psicométricos e o Funcionamento Diferencial dos Itens da Escala de Avaliação da Fadiga por meio da Teoria de Resposta ao Item entre as cinco regiões geopolíticas do Brasil. Nesse sentido, inicialmente foi confirmada a unidimensionalidade dos itens e a plausibilidade do ajuste dos dados ao modelo de Resposta Gradual (GRM) de Samejima, o que possibilitou a estimação dos parâmetros de discriminação (*a*) e dificuldade que, pelo GRM, é dado por meio dos limiares de resposta (j_s).

O parâmetro de discriminação (*a*) apresentou média de 2,16 com baixa variância (*DP*=0,53), o que para Baker (2001), é uma discriminação muito alta, já que esse autor oferece o seguinte critério de classificação: de 0,65 a 1,34 corresponde à discriminação moderada; de 1,35 e 1,69 é discriminação alta; e, por fim, acima de 1,70 indica

discriminação alta. Segundo esse autor, a maioria das medidas politômicas tem apresentado discriminação alta ou muito alta. Nesse ponto, indica que. na delimitação desse critério. é importante considerar o tamanho da amostra, pois para amostras muito grandes (e.g., N>1.000), o poder de discriminação do item precisaria ser maior para que os participantes pudessem assumir valores distintos na mesma métrica. Sendo assim, entende-se que uma discriminação boa para uma amostra pequena talvez não seja tão útil para um processo seletivo com amostra maior.

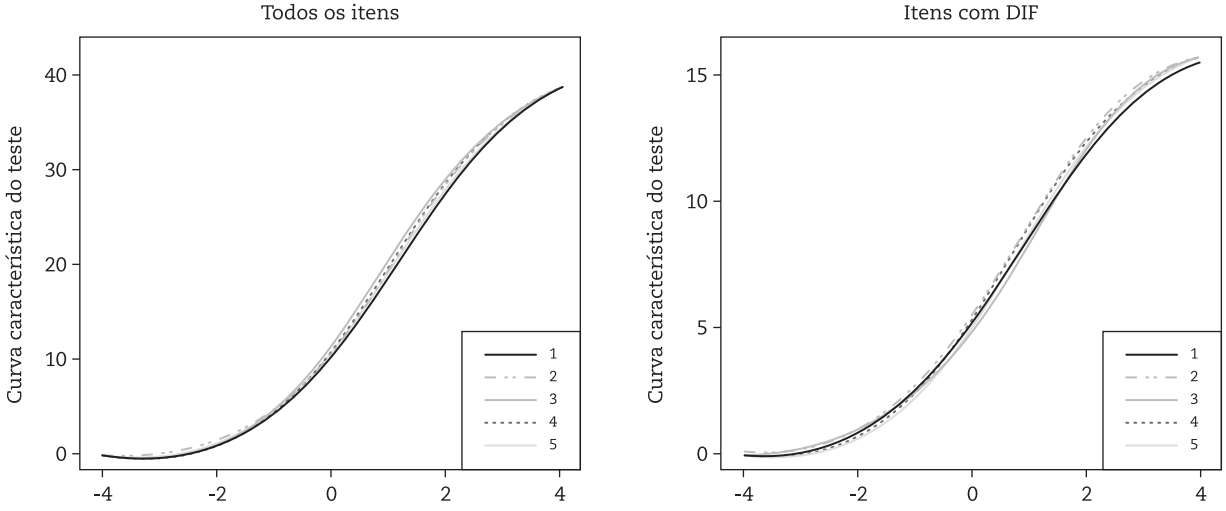

Figura 2. Impacto dos itens com presumível DIF nas curvas características do teste

Os itens 1 (Sinto-me incomodado devido à fadiga), 2 (Fico cansado muito rapidamente), 5 (Sinto-me exausto fisicamente) e 9 (Sinto-me exausto mentalmente) apresentaram discriminações superiores a 2, além de grande amplitude nos limiares de resposta, sugerindo que a boa discriminação se dá em um contínuo amplo. Isso atesta a possibilidade de utilizar esse instrumento ou pelo menos a maioria desses itens em pesquisas com amostras que apresentem *thetas* bastante diferentes, preferencialmente entre pessoas que estejam entre -1 e +3 de desvios padrões na curva normal da distribuição do *theta*.

Esses itens ainda proporcionaram a maior quantidade de informação no intervalo de -4 a +4, indicando sua precisão nos limiares de resposta, demonstrando sua aplicabilidade em pesquisas com populações diversas, uma vez que é na amplitude de -4 a +4 DPs que estão 99,99% das pessoas. Isso parece convalidar as diversas pesquisas em que esse instrumento tem sido utilizado (de Vries & Drent, 2004; Hinz, et al., 2011)

Quanto ao parâmetro de dificuldade, a média para o j_1 foi relativamente pequena, não contemplando a parte da distribuição mais baixa da fadiga, apesar de ter apresentado pouca variabilidade, o que é recomendável segundo Pasquali (2003). Albuquerque e Tróccoli (2004) chamam a atenção para a necessidade de se desenvolver critérios de avaliação do parâmetro b para diferentes faixas de itens politômicos.

Os itens 6 (Tenho problemas para começar coisas), 7 (Tenho problemas em pensar claramente) e 8 (Não sinto vontade de fazer nada) apresentaram as amplitudes nos limiares de respostas mais baixas; os j1 desses itens foram os mais baixos, ficando entre -0,65 e -0,24, estando dentro do que Baker (2001) considera como alta ou muito alta, porém a informação proporcionada pelos itens ficou entre as mais baixas. Os resultados indicam que esses três itens apresentam pouca precisão para participantes com baixa fadiga, especificamente para a porção de 15,9% da população que tem menos fadiga, ou seja, quem está -0,10 DP abaixo da curva normal.

Quando se comparou a semântica dos quatro itens citados com melhor desempenho e os três com menor precisão psicométrica, verificou-se que os melhores itens apresentaram conteúdos mais diretamente relacionados com a exaustão e a fadiga, porém os três itens com desempenho inferior mostraram conteúdos relacionados com a compreensão da fadiga, como falta de vontade de fazer coisas (itens 6 e 8) e dificuldade de pensar (item 7). Isso poderia explicar a pouca precisão desses itens para aqueles 15,9% de pessoas que apresentaram menos fadiga, pois as consequências desta se dariam em níveis mais elevados.

Para verificar a presença de DIF, foi utilizada a razão de verossimilhança, que destacou quatro itens com provável DIF, sendo que três com DIF uniforme e um item sem tipo detectável. Porém, esse resultado pode advir do tamanho elevado da amostra. Sendo assim, outros critérios foram utilizados considerando a magnitude dos efeitos das regressões, como a estatística Pseudo-R^2, e as

diferenças entre os βs resultantes dos modelos 1 e 2 da regressão ($\Delta\beta_1$).

Em relação ao Pseudo-R^2, os efeitos de magnitude encontrados neste estudo estiveram abaixo dos preconizados por Zumbo (1999), variando de $R^2_{(1;\ 3)}$=0,0019 a $R^2_{(2;\ 3)}$=0,0002. Outro critério adotado para dirimir qualquer dúvida a respeito da presença de DIF nos itens da medida de fadiga foi o $\Delta\beta_1$. Mesmo considerando os critérios menos permissivos, adotados por Crane et al. (2004), os $\Delta\beta_1$ encontrados para os itens da citada escala foram bem abaixo.

Os itens também foram examinados graficamente, sendo sobrepostas as CCIs de cada região. De modo geral, percebeu-se que, além de um comportamento análogo dos parâmetros dos itens entre os grupos, houve uma relativa sobreposição das curvas para todos os itens, conforme pode ser verificado em razão da pouca variação dos parâmetros *a* e *b*. Por fim, a curva característica do teste evidenciou apenas uma pequena ausência de sobreposição entre os sujeitos que apresentam maiores *thetas*. Considerando esses resultados, os itens da medida de fadiga provavelmente não apresentam DIF, corroborando sua utilização no contexto brasileiro com independência de região geopolítica. Não obstante, este estudo apresenta algumas limitações que precisam ser consideradas, a exemplo da amostra, que não pode ser representativa do Brasil, uma vez que os participantes foram médicos, pessoas com nível alto de escolaridade e renda, além de constituírem um grupo homogêneo, características que podem interferir na ausência de DIF. Mesmo com tais limitações potenciais, confia-se que este estudo tenha oferecido suporte psicométrico que justifique a utilização da Escala de Avaliação da Fadiga no Brasil, ao menos no âmbito da pesquisa.

Por fim, como estudos futuros, será importante checar a invariância dos parâmetros dos itens em amostras distintas e representativas, assim como explorar o DIF da EAF entre gênero de diferentes contextos interculturais. Considerando que a TRI oferece flexibilidade na escolha dos itens (Irwin et al., 2012), outra demanda será investigar as propriedades psicométricas dessa escala, estratificando por itens que avaliem com menor taxa de erro pessoas com alta e baixa fadiga. Mais especificamente, retirando os itens que apresentaram pouca confiabilidade para as pessoas com baixa fadiga, a exemplo dos itens 6, 7 e 8, requerer-se-á testar suas propriedades.

Referências

Albuquerque, A. S., & Tróccoli (2004). Desenvolvimento de uma escala de bem-estar subjetivo. *Psicologia: Teoria e Pesquisa, 20*(2), 153-164.

Alvarenga Filho, H., Carvalho, S. R. D. S., Dias, R. M., & Alvarenga, R. M. P. (2010). Principais testes utilizados na avaliação de fadiga na esclerose múltipla: revisão sistemática. *Revista Brasileira de Neurologia, 46*(2), 37-43.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Baker, F. B. (2001). *The basics of item response theory* (2ª ed.). Washington, DC: Eric Clearinghouse on Assessment and Evaluation.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R package for detecting differential item Functioning using iterative hybrid ordinal logistic regression/Item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(8), 1- 30.

Crane, P. K. van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*(2), 241-256.

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIF detect and dif with par. *Medical Care, 44*(3), 115-123.

de Vries, J., & Drent, M. (2004). Relationship between perceived stress and sarcoidosis in a Dutch patient population. Sarcoidosis, vasculitis, and diffuse lung diseases: *Official Journal of WASOG/World Association of Sarcoidosis and Other Granulomatous Disorders, 21*(1), 57-63.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence.

Fini, A., & Cruz, D. D. A. L. M. (2010). Psychometric properties of the Dutch Fatigue Scale and the Dutch Exertion Fatigue Scale: brazilian version. *Revista Brasileira de Enfermagem, 63*(2), 216-221.

Fox, J. (2005). The R Commander: A basic statistics graphical user interface to R. *Journal of Statistical Software, 14*(9), 1-42.

Gouveia, V. V., Oliveira, G. F., Mendes, L. A., Souza, L. E. C., Cavalcanti, T. M., & Melo, R. L. P. (2015). Escala de Avaliação da Fadiga: adaptação para profissionais da saúde. *Revista Psicologia: Organizações e Trabalho, 15*(3), 246-256.

Hauck-Filho, N., & Teixeira, M. A. P. (2013). Funcionamento diferencial do item no Alcohol use Disorders Identification Test. A*valiação Psicológica, 12*(1), 19-25.

Havlikova, E., Rosenberger, J., Nagyova, I., Middel, B., Dubayova, T., Gdovinova, Z., ... Groothoff, J. W. (2008). Impact of fatigue on quality of life in patients with Parkinson's disease. *European Journal of Neurology, 15*(5), 475-480.

Hinz, A., Fleischer, M., Brähler, E., Wirtz, H., & Bosse-Henck, A. (2011). Fatigue in patients with sarcoidosis, compared with the general population. *General Hospital Psychiatry, 33*(5), 462-468.

Irwin, D. E., Stucky, B. D., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J., … DeWalt, D. A. (2012). PROMIS pediatric anger scale: An item response theory analysis. *Quality of Life Research, 21*(4), 697-706.

Ladwig, R. (2012). *Detecção de funcionamento diferencial do item através da regressão logística e da teoria de resposta ao item – uma interface gráfica* (Monografia não publicada). Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.

Lawrence, D. P., Kupelnick, B., Miller, K., Devine, D., & Lau, J. (2004). Evidence report on the occurrence, assessment, and treatment of fatigue in cancer patients. *Journal of the National Cancer Institute. Monographs, 32*, 40-50.

Makransky, G., & Glas, C. A. W. (2013). Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application. *Measurement, 46*(9), 3228-3237.
Michielsen, H. J., de Vries, J., van Heck, G. L., van de Vijver, F. J. R., & Sijtsma, K. (2004). Examination of the dimensionality of fatigue: The construction of the Fatigue Assessment Scale (FAS). *European Journal of Psychological Assessment, 20*(1), 39-48.
Mills, R. J., Young, C. A., Pallant, J. F., & Tennant, A. (2010). Development of a patient reported outcome scale for fatigue in multiple sclerosis: The Neurological Fatigue Index (NFI-MS). *Health Quality Life Outcomes, 8*(22), 1-10. Recuperado de http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2834659/pdf/1477-7525-8-22.pdf
Moore, R. P. G., & Gastão Neto, F. D. (2013). Occupational fatigue: Impact on anesthesiologist's health and the safety of surgical patients: As anesthesiologists we are frequently working in a stressful environment. Do you disagree with this? *Revista Brasileira de Anestesiologia, 63*(2), 167-169.
Mukherjee, S., Gibbons,L. E., Kristjansson, E., & Crane, P. K. (2013). Extension of an iterative hybrid ordinal logistic regression/item response theory approach to detect and account for differential item functioning in longitudinal data. *Psychological Test and Assessment Modeling, 55*(2), 127-147.
Oliveira, G. F., Gouveia, V. V., Peixoto, G. P., & Soares, M. A. L. (2010). Análise fatorial da Escala de Avaliação da Fadiga em uma amostra de universitários de instituição pública. *ID on-line Revista de Psicologia, 4*(11), 51-60.
Pasquali (2003). *Psicometria: teoria dos testes na psicologia e na educação.* Petrópolis, RJ: Editora Vozes.
Pasquali, L. (2007). *Teoria de resposta ao item: teoria, procedimentos e aplicações.* Brasília, DF: Laboratório de Pesquisa em Avaliação, LabPAM/UNB.
Pereira, M. G., & Duarte, S. (2010). Fadiga intensa em doentes com lúpus eritematoso sistémico: estudo das características psicométricas da escala da intensidade da fadiga. *Psicologia, Saúde & Doenças, 11*(1), 121-136.
R Development Core Team (2012). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Recuperado de http://www.R-project.org/.
Rizopoulos, D. (2006). LTM: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1-25.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4), 100-114.
Sanches, K. C., & Cardoso, K. G. (2012). Estudo da fadiga e qualidade de vida nos pacientes com doença de Parkinson. *Journal Health Science Institute, 30*(4), 391-394.
Sisto, F. F. (2006). O funcionamento diferencial dos itens. *Psico-USF, 11*(1), 35-43.
Somerville, J. T. (2012). *Detection of differential item functioning in the generalized full-information item bifactor analysis model* (Tese de doutorado não publicada). University of California, Los Angeles, CA.
Swaminathan, H., & Rogers H. J. (1990). Detecting Differential Item Functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.
Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1-19.
Wang, W., Tay, L., & Drasgow, F. (2013). Detecting differential item functioning of polytomous items for an ideal point response process. *Applied Psychological Measurement, 37*(4), 316-335.
Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF).* Ottawa: National Defense Headquarters.

recebido em junho de 2015
reformulado em novembro de 2015
aprovado em dezembro de 2015

Sobre os autores

Thiago Medeiros Cavalcanti é psicólogo, mestrando pela Universidade Federal da Paraíba.
Romulo Lustosa Pimenteira De Melo é psicólogo, doutorando pela Universidade Federal da Paraíba.
Emerson Diogenes De Medeiros é psicólogo, doutor em psicologia social e professor adjunto da Universidade Federal do Piauí.
Layrtthon Carlos De Oliveira Santos é psicólogo, doutorando pela Universidade Federal da Paraíba.
Valdiney Veloso Gouveia é psicólogo, doutor em psicologia social e professor titular na Universidade Federal do Paraíba.