Psicologia: Teoria e Pesquisa
Jan-Mar 2012, Vol. 28 n. 1, pp. 15-25

Funcionamento Diferencial dos Itens do Teste Não-Verbal de Inteligência SON-R 2½-7[a]

Camila Akemi Karino[1]
Jacob Arie Laros
Girlene Ribeiro de Jesus
Universidade de Brasília

RESUMO - A presença de uma quantidade considerável de itens com funcionamento diferencial (DIF) pode tornar um teste menos válido. Assim, este estudo investigou a existência de DIF no teste de inteligência SON-R 2½-7[a]. O teste é a versão abreviada do SON-R 2½-7, normatizado e validado em vários países da Europa. Os dados de 1.200 crianças da normatização brasileira foram utilizados para identificar a presença de DIF em relação à gênero e região, usando o método da TRI. Os resultados indicaram que, de um total de 60 itens, 5 itens apresentaram DIF entre os sexos e 13 itens apresentaram DIF entre as regiões. Conclui-se que há adequabilidade da maioria dos itens, o que viabiliza o uso do SON-R 2½-7[a] em contexto nacional.

Palavras-chave: DIF, SON-R 2½-7[a], teste de inteligência.

Differential Item Functioning of the SON-R 2½-7[a] Non-Verbal Intelligence Test

ABSTRACT - The presence of a considerable amount of items with differential functioning can make a test less valid. Therefore, the existence of Differential Item Functioning (DIF) in the intelligence test SON-R 2½-7[a] was investigated. This is the abridged edition of the SON-R 2½-7 with normatization and validation studies realized in various European countries. The data of the Brazilian normatization sample of 1,200 children were used to verify the presence of DIF in relation to gender and region using the IRT method. Of a total of 60 items, 5 items were indicated as having DIF for gender and 13 items as having DIF between regions. It was concluded that the majority of the items were adequate, which makes the use of the test feasible in a national context.

Keywords: DIF, SON-R 2½-7[a], intelligence test.

Em certa avaliação de inteligência de uma criança de 7 anos, residente na periferia de Brasília, a psicóloga pergunta: "O que você faria se encontrasse uma bolsa ou carteira de alguém em uma loja?". A criança responde: "procurava o telefone para ..., (pausa pensando), quer dizer eu pegava o dinheiro e as coisas para mim".

O exemplo apresentado refere-se a um item do subteste Compreensão do teste WISC-III (Wechsler, 2002). A criança recebeu uma pontuação zero, pois se esperava uma resposta que mostrasse o intuito dela devolver o pertence ao dono (2 pontos) ou de procurar identificar o dono (1 ponto). Contudo, pode-se questionar o quanto o item realmente mensurou inteligência. A resposta pode não estar de acordo com os padrões desejáveis da sociedade, mas está de acordo com a realidade e educação recebida pela criança. Sua resposta é coerente, apresenta adequado raciocínio lógico e verbal. Por que então atribuir zero à resposta? Essa é uma das fortes críticas que têm sido feitas aos testes tradicionais de inteligência.

De acordo com alguns autores (Weiss, 1982; Thorndike, Hagen & Sattler, 1986; Tellegen & Laros, 2004), os testes de inteligência geral, como o Stanford-Binet e os testes de inteligência Wechsler, focam mais na inteligência adquirida ao longo de um processo formal de aprendizagem, denominada inteligência cristalizada. Com isso, grupos sociais que não tiveram acesso às mesmas condições de aprendizagem podem ter seu desempenho no teste prejudicado.

Itens como identificar a parte faltante de um piano ou perguntas como "Qual a distância entre São Paulo e Lisboa?" do teste WISC-III são úteis para avaliar a inteligência de crianças de um determinado contexto social. Todavia, a resposta errada de uma criança que não teve acesso às mesmas condições de educação formal, não significa que ela é incapaz de adquirir certos conhecimentos ou habilidades. Assim, testes de inteligência que aferem, sobretudo, o resultado final de um processo formal de aprendizagem, podem acabar subestimando a habilidade de crianças que tiveram menos oportunidades - em geral, grupos de minorias étnicas ou de baixo nível socioeconômico (Tellegen & Laros, 2004; Tellegen & Laros, 2005). Pode-se ressaltar que uma avaliação ou um diagnóstico mal feito pode trazer consequências sérias para a criança, como a discriminação e o preconceito.

De acordo com Jesus (2009), o SON-R 2½-7[a] diferencia-se dos testes tradicionais de inteligência por buscar mensurar o potencial para aprendizagem da criança, ou seja, a inteligência fluida. A cada item é apresentada uma situação-problema para que a criança, a partir dos estímulos apresentados, busque a melhor solução. Esse teste foi normatizado para o Brasil em 2008 e consiste de quatro subtestes: Categorias, Situações, Mosaicos e Padrões (Laros, Tellegen,

1 Endereço para Correspondência: Departamento de Psicologia Social e do Trabalho, Instituto de Psicologia, Campus Darcy Ribeiro, Universidade de Brasília, Brasília, DF. *E-mail*: camilaakarino@gmail.com

Jesus & Karino, no prelo). Exemplos dos itens dos quatro subtestes são oferecidos no *Manual and Research Report* do SON-R 2½-7 que está disponível no web site dos testes SON (www.testresearch.nl).

Uma das grandes vantagens do teste SON-R 2½-7[a] diz respeito à maior facilidade para adequação a diferentes culturas, uma vez que o processo de adaptação de testes não-verbais é menos complicado do que o exigido para testes que utilizam linguagem escrita ou falada como parte do seu conteúdo. Não obstante, o fato de os testes não-verbais não exigirem tradução não significa que esses instrumentos possam ser utilizados sem estudos que verifiquem sua adequação à cultura na qual serão utilizados.

A esse respeito, Van de Vijver e Poortinga (1997) assinalam que as propriedades psicométricas como validade e fidedignidade de instrumentos psicológicos desenvolvidos em uma determinada cultura, não podem ser assumidas em outra sem ser empiricamente demonstradas. O viés, por exemplo, pode mudar as propriedades psicométricas de um instrumento quando ele é usado em uma cultura diferente.

Dessa forma, o presente estudo tem como foco a investigação de DIF no SON-R 2½-7[a], pois apesar de ser um teste não verbal, muitas vezes o DIF pode ser indício de viés cultural. Tal investigação também é importante, uma vez que se trata de um novo teste que está sendo utilizado no cenário nacional, com validação e normatização recentes.

No Brasil, vários autores verificaram a necessidade de testes psicológicos válidos e com normas nacionais, especificamente na área das habilidades cognitivas de crianças pré-escolares e escolares (Muñiz, Prieto, Almeida & Bartram, 1999; Oakland, Wechsler, Bensuan & Stafford, 1994; Hu & Oakland, 1991). Uma grande parte dos testes brasileiros são traduções de versões internacionais, não passaram por um adequado processo de validação e normatização e, mesmo assim, acabam sendo utilizados no país inteiro (Alves, 2002; Noronha & Alchieri, 2002; Noronha et al., 2003; Noronha, Primi & Alchieri, 2004). O teste SON-R 2½-7[a], por ser um instrumento desenvolvido para crianças entre 2 anos e meio a 7 anos e por ser o primeiro teste de inteligência no Brasil a possuir normas nacionais, vem de certo modo suprir parte dessa deficiência.

Contudo, mesmo com o adequado processo de normatização e validação de um teste, ainda não é possível garantir uma avaliação totalmente justa. Os itens do instrumento podem privilegiar um grupo em detrimento de outro. Por exemplo, um item que aborda conteúdos relacionados a futebol pode privilegiar meninos em relação às meninas. Deste modo, os itens podem apresentar vieses e/ou funcionamento diferencial (DIF). O DIF é um termo estatístico utilizado para descrever a situação na qual pessoas de um grupo respondem de forma correta um item mais frequentemente do que pessoas de outro grupo com mesma habilidade.

Nesse contexto, a introdução do termo 'differential item functioning' (DIF) possibilita a distinção entre impacto do item e viés do item. O impacto do item descreve a situação na qual o DIF existe porque existem diferenças reais entre os grupos no que tange ao construto que está sendo mensurado por meio do item. Por sua vez, o viés do item descreve a situação na qual existe DIF devido a algumas características dos itens do teste ou da situação de testagem, as quais não são relevantes para o construto de interesse (portanto, para o propósito do teste) (Zumbo, 2007). Dessa forma, DIF quer dizer que existem diferenças e só se pode falar de viés quando existem argumentos válidos que explicam o desempenho diferencial (Camilli & Shepard, 1994).

Os diferentes métodos utilizados para verificar a existência de DIF permitem analisar se os itens de um teste estão funcionando da mesma forma em vários grupos de respondentes. Dessa forma, configura-se como uma boa forma de verificar a invariância da mensuração, ou seja, o teste funciona da mesma forma para diferentes grupos de respondentes?

O termo DIF tem sido preferido pelo fato de fornecer uma informação sem preconceito, pois apenas indica propriedades estatísticas diferentes de acordo com o grupo ao qual foi aplicado (Alves, 2004). É o possível funcionamento diferencial dos itens de instrumentos de avaliação que guia este estudo, não necessariamente identificar os vieses. No entanto, é preciso admitir que esses dois construtos caminham, na maioria das vezes, juntos e o cuidado com os possíveis vieses pode auxiliar na construção de um teste com menos DIF.

Uma análise de vieses culturais já foi realizada no Brasil com a versão do SON-R para a faixa etária entre 5 e 17 anos (Tellegen & Laros, 2004). No referido estudo, alguns subtestes do teste SON-R 5½-17 foram aplicados em crianças brasileiras. Dois procedimentos foram seguidos para avaliar o viés do item: dificuldade do item e reconhecimento dos desenhos que compunham os itens. Os autores encontraram 14 itens com viés cultural, dos quais quatro favoreciam as crianças brasileiras.

Antes do estudo de normatização do teste SON-R 2½-7[a], foi realizado também um estudo piloto em Brasília com 130 crianças (Jesus, 2009). Este estudo teve como objetivo adequar o conteúdo do teste para a realidade brasileira. Desse estudo piloto, que teve como objetivo fazer a validação de conteúdo do SON-R 2½-7, participaram cerca de dez especialistas da área de psicologia e neuropsicologia infantil e estudantes de graduação e pós-graduação em Psicologia. Como o subteste Categorias foi apontado como problemático este foi aplicado em cinco crianças com idade na faixa etária do teste. No total, sete itens do subteste Categorias sofreram modificações da versão do SON-R 2½-7, holandesa, para a versão do SON-R 2½-7[a], versão reduzida brasileira.

Foi utilizado como parâmetro para a modificação dos itens, além da análise dos especialistas e estudantes de psicologia citados, a avaliação feita pelas crianças acerca das figuras. Por exemplo: (a) um chapéu muito utilizado na Europa foi confundido com um sofá e com ferro de passar roupas pelas crianças brasileiras; (b) a figura de uma xícara infantil holandesa foi confundida com um pinico pelas crianças brasileiras. Essas e outras figuras com problemas semelhantes foram modificadas.

Todos esses cuidados já apresentados contribuem para minimizar os possíveis vieses culturais. Contudo, ainda não imunizam o teste de possuir itens com funcionamento diferencial. Como já discutido, nem todo item com DIF possui claramente um viés (Camilli & Shepard, 1994).

A importância de uma avaliação justa e fidedigna, juntamente com o papel relevante do SON-R 2½-7[a] para o contexto brasileiro, justifica o presente estudo. Dessa forma,

pretende-se aqui avaliar o funcionamento diferencial dos itens que compõem as escalas de raciocínio e de execução do teste de inteligência SON-R 2½-7[a], considerando dois grupos de análise: gênero e região.

Funcionamento Diferencial do Item (DIF)

O fato de os meninos obterem um percentual de acerto maior em testes de raciocínio espacial do que meninas pode indicar DIF no teste ou uma maior habilidade dos meninos nesse tipo de tarefa (Hogan, 2006; Jardine & Martin, 1983).

É com essa preocupação que surge o paradoxo de Simpson (Andriola, 2001). De acordo com este paradoxo, temos que comparar o comparável. Assim, as diferenças na probabilidade de acerto do item somente podem ser comparadas se os sujeitos possuem a mesma magnitude do traço latente que está sendo medido.

Há vários métodos utilizados ultimamente para a análise do comportamento diferencial do item, que consideram o paradoxo de Simpson: Delta-plot, Mantel-Haenszel, regressão logística e uma técnica baseada na Teoria de Resposta ao Item (TRI) (Alves, 2004; Andriola, 2001). Neste estudo, optamos por utilizar o método da TRI.

Na TRI, um item possui DIF quando a Curva Característica do Item (CCI) difere para dois ou mais grupos, considerando o nível da variável latente (Lord, 1980). A CCI, para dados dicotômicos, expressa a partir de uma curva monotônica crescente a probabilidade de acerto ao item de acordo com o aumento da magnitude do construto avaliado (Hambleton, Swaminatham & Rogers, 1991).

Além da comparação das CCIs, na TRI, a análise de DIF pode ser feita a partir da comparação dos parâmetros dos itens (Cohen, Kim & Baker, 1993). Essa técnica foi proposta por Lord (1980) e define que um item apresentará DIF se o valor do parâmetro *b* estimado para uma população possuir diferenças significativas com o valor estimado em outra população. Para tanto, é preciso definir um grupo de referência, que servirá como base de comparação. Neste estudo, essas duas técnicas foram utilizadas para analisar DIF, considerando as variáveis sexo e região.

Método

Participantes

Para esse estudo utilizamos a base de dados composta por 1.200 crianças participantes da pesquisa de normatização do teste SON-R 2½-7[a] no Brasil (Jesus, 2009). As crianças tinham entre 3 e 7 anos, com cerca de 240 crianças em cada uma das cinco idades contempladas. Também foi mantida uma divisão igualitária entre os sexos.

A amostra foi desenhada considerando: o IDHM (Índice de Desenvolvimento Humana composto de três fatores: renda, longevidade e educação) das cidades brasileiras, a densidade de crianças na faixa etária do teste em cada região, a localização das cidades (capital e interior), o tipo de escola (pública ou particular) e a escolaridade. No total, foram 36 municípios participantes de 13 estados brasileiros: Amazonas, Bahia, Ceará, Distrito Federal, Goiás, Maranhão, Minas Gerais, Pará, Paraná, Rio de Janeiro, Rio Grande do sul, São Paulo e Tocantins. Um detalhamento maior sobre a amostra da pesquisa pode ser encontrado em Jesus (2009) e Laros, Tellegen, Jesus e Karino (no prelo).

Instrumento

O instrumento utilizado foi o teste não-verbal de inteligência SON-R 2½-7[a], que é a versão reduzida e normatizada para a população brasileira do teste não-verbal de inteligência SON-R 2½-7, versão européia. O teste SON-R 2½-7 foi normatizado e validado na Holanda em 1998 com base em uma amostra de 1.124 crianças (Tellegen, Winkel, Wijnberg-Williams & Laros, 1998). Em 2008, foi realizada a normatização brasileira da versão reduzida do teste em uma amostra de 1.200 crianças de todas as regiões do país (Jesus, 2009). O teste é de aplicação individual e sua versão reduzida é composta por quatro subtestes: Mosaicos, Categorias, Situações e Padrões.

No subteste Mosaicos, a criança precisa copiar padrões de mosaicos em uma moldura utilizando quadrados vermelhos, amarelos e vermelhos/amarelos. No subteste padrões, a criança precisa copiar com um lápis uma forma geométrica. Esses dois subtestes possuem, respectivamente, 15 e 16 itens, e compõem a escala de execução (SON-EE).

O subteste Categorias é constituído por 15 itens. Nesse subteste a criança precisa escolher as figuras que possuem algo em comum com a categoria apresentada. Por exemplo, são apresentadas três figuras de boneca que formam uma categoria e a criança precisa escolher entre 5 figuras, as duas que complementam esta categoria. Por fim, o subteste Situações é composto por 14 itens e exige que a criança escolha a opção que melhor complete uma figura ou situação, de modo a deixá-la consistente e coerente. Esses dois últimos subtestes compõem a escala de raciocínio (SON-ER), raciocínio concreto e abstrato.

As análises de qualidade psicométrica, fidedignidade e validade do instrumento foram realizadas por Jesus (2009). Nessas análises, verificou-se boa consistência interna do instrumento (fidedignidade média de 0,92), adequada validade de construto (confirmação da estrutura e cargas fatoriais em média superiores a 0,40) e validade convergente (correlação de 0,75 e 0,67 com WIPPSI-III e WISC-III, respectivamente) (Laros, Tellegen, Jesus e Karino, no prelo).

Procedimento

A coleta de dados foi realizada na casa das crianças ou em instituições como: igrejas, escolas e creches. As instituições foram selecionadas por conveniência e as crianças foram selecionadas de acordo com o critério de idade. Os pais recebiam uma carta de explicação da pesquisa e, caso concordassem com a participação do seu filho, assinavam o Termo de Consentimento Livre e Esclarecido.

No dia marcado, um aplicador treinado comparecia à instituição para a realização da avaliação. Todas as aplicações

Tabela 1. Quantidade de Participantes, Média e Desvio-Padrão nas Escalas e no QI total do SON-R 2½-7[a] por Grupo de Análise.

Grupo	N	ER		EE		QI Total	
		Média	DP	Média	DP	Média	DP
Masculino	600	99	15	100	16	99	15
Feminino	600	101	15	100	14	101	15
T		-1,98		-0,96		-1,61	
Sig.		0,05		0,34		0,11	
Sudeste	459	104	14	103	14	103	14
Norte	120	95	13	98	13	96	13
Nordeste	381	97	15	97	16	97	16
Sul	160	98	14	99	14	98	14
Centro-Oeste	80	106	16	104	17	106	17
F		19,93		9,42		17,17	
Sig.		<0,01		<0,01		<0,01	

ER = Escala de Raciocínio, EE=Escala de Execução, N=número de participantes no grupo, DP=Desvio-padrão

Tabela 2. Valores dos Parâmetros b dos Itens da Escala de Raciocínio por Sexo.

Item	Dificuldade dos itens Parâmetro *b*		Diferença entre os grupos	Erro Padrão da Diferença
	Masculino (GR)	Feminino (GF)	(GF-GR)	
CAT1	-1,60	-1,73	-0,13	0,10
CAT2	-1,49	-1,62	-0,14	0,09
CAT3	-1,53	-1,55	-0,02	0,11
CAT4	-1,10	-1,15	-0,05	0,09
CAT5	-0,58	-0,74	-0,16	0,09
CAT6	-0,50	-0,74	-0,24	0,07
CAT7	-0,22	-0,27	-0,06	0,08
CAT8	0,07	0,02	-0,05	0,05
CAT9	0,18	0,14	-0,03	0,05
CAT10	0,58	0,62	0,04	0,06
CAT11	0,78	0,75	-0,03	0,04
CAT12	0,99	0,98	-0,01	0,05
CAT13	1,06	1,05	-0,02	0,05
CAT14	1,30	1,21	-0,09	0,07
CAT15	1,48	1,51	0,03	0,10
SIT1	-2,43	-2,07	0,36	0,13
SIT2	-2,03	-1,89	0,15	0,12
SIT3	-1,77	-1,54	0,24	0,13
SIT4	-1,50	-1,38	0,12	0,11
SIT5	-0,56	-0,41	0,15	0,08
SIT6	-0,58	-0,52	0,06	0,09
SIT7	0,13	0,28	0,15	0,07
SIT8	-0,25	-0,11	0,13	0,06
SIT9	0,40	0,18	-0,22	0,07
SIT10	0,90	0,89	-0,02	0,07
SIT11	1,37	1,37	0,00	0,11
SIT12	1,31	1,29	-0,02	0,08
SIT13	1,50	1,43	-0,07	0,09
SIT14	1,57	1,51	-0,07	0,09

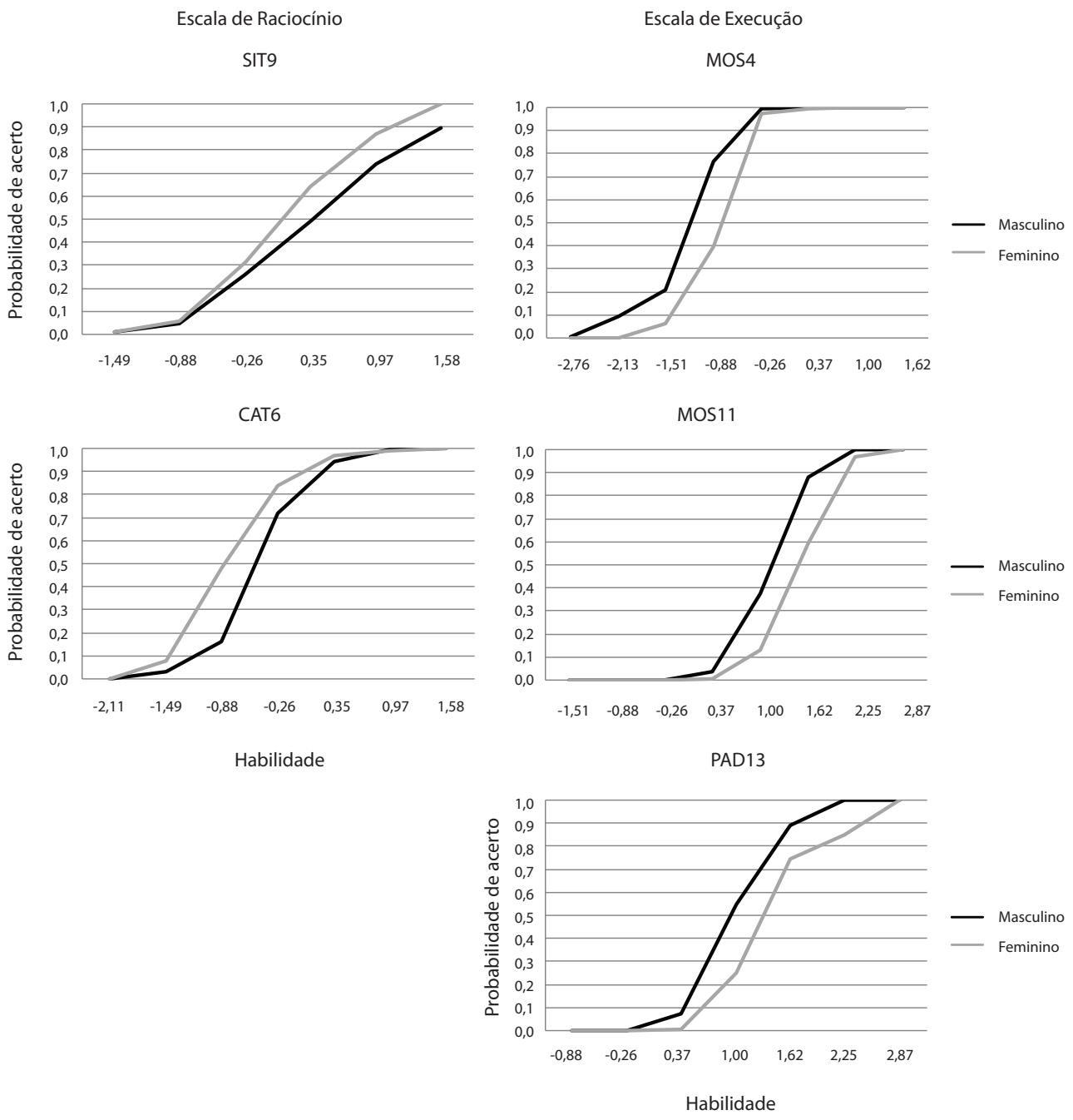

Figura 1. CCI dos Itens da Escala de Raciocínio e Execução que apresentaram DIF por Sexo.

foram individuais e realizadas em uma única seção. Ao final, foi encaminhado aos pais ou responsáveis um relatório com o desempenho da criança.

Análises dos Dados

As análises foram realizadas considerando o sexo e a região onde as crianças residiam. Quando considerada a variável sexo, o masculino foi o grupo de referência e o feminino o grupo focal. Na análise por região, o Sudeste foi o grupo de referência e as demais regiões, os grupos focais.

As análises foram feitas com base no modelo logístico de dois parâmetros da Teoria de Resposta ao Item (TRI) e para tanto foi utilizado o programa Bilog-MG 3.0. Para verificação de DIF foram comparados os parâmetros dos itens, e para aqueles itens que apresentaram DIF foi analisada também a Curva Característica do Item (CCI).

Ao definir o grupo de referência no Bilog-MG 3.0 e realizando a calibração conjunta, isto é, a equalização com grupos múltiplos, a escala de proficiência do teste, bem como os resultados dos diferentes grupos, se tornam comparáveis em uma métrica comum.

Tabela 3. Valores dos Parâmetros b dos Itens da Escala de Execução por Sexo.

Item	Dificuldade dos itens Parâmetro *b*		Diferença entre os grupos	Erro Padrão da Diferença
	Masculino (GR)	Feminino (GF)	(GF-GR)	
MOS1	-2,47	-2,55	-0,09	0,28
MOS2	-1,43	-1,51	-0,07	0,10
MOS3	-1,21	-1,09	0,12	0,10
MOS4	-1,19	-1,00	0,20	0,06
MOS5	-1,00	-1,00	0,00	0,07
MOS6	-0,63	-0,64	-0,01	0,06
MOS7	-0,32	-0,31	0,00	0,06
MOS8	-0,13	-0,14	-0,02	0,05
MOS9	0,57	0,57	0,00	0,05
MOS10	0,92	0,98	0,06	0,05
MOS11	1,12	1,32	0,20	0,06
MOS12	1,53	1,72	0,19	0,10
MOS13	1,98	2,10	0,13	0,17
MOS14	1,82	1,94	0,12	0,11
MOS15	2,10	2,15	0,05	0,17
PAD1	-1,93	-2,14	-0,21	0,14
PAD2	-1,89	-1,96	-0,07	0,12
PAD3	-2,06	-2,11	-0,05	0,14
PAD4	-1,87	-2,08	-0,22	0,12
PAD5	-1,81	-2,15	-0,33	0,15
PAD6	-1,25	-1,30	-0,06	0,07
PAD7	-1,07	-1,25	-0,17	0,08
PAD8	-0,59	-0,75	-0,16	0,06
PAD9	-0,38	-0,46	-0,08	0,05
PAD10	0,17	0,13	-0,03	0,05
PAD11	0,48	0,56	0,08	0,04
PAD12	0,57	0,67	0,10	0,05
PAD13	0,97	1,15	0,17	0,06
PAD14	1,03	1,11	0,07	0,06
PAD15	1,43	1,61	0,18	0,08
PAD16	1,96	1,84	-0,12	0,12

Resultados

Para apresentação dos resultados, primeiramente são mostradas as médias de cada grupo na escala de QI (média=100 e desvio-padrão=15), para que se tenha uma noção geral da amostra utilizada. Posteriormente, serão descritos os resultados da análise de DIF por sexo e, por fim, os resultados da análise de DIF por região. Nas análises de DIF, iremos apresentar as comparações dos parâmetros e a análise das diferenças. Somente para os itens que apresentaram DIF serão realizadas as análises da CCI.

Na Tabela 1 são apresentadas as estatísticas descritivas das escalas de execução (EE), de raciocínio (ER) e do QI total do SON-R 2½-7[a] de cada um dos grupos. Nota-se que o quantitativo de participantes em relação à variável sexo é igualitariamente dividido. Já o quantitativo de participantes por região seguiu proporcional à densidade de crianças nessa faixa etária em cada localidade.

Ainda em relação à Tabela 1, verificamos médias e desvios-padrão muito semelhantes entre os dois sexos, não apresentando diferenças significativas (Teste *t*, considerando

Tabela 4. Valores dos Parâmetros b dos Itens da Escala de Raciocínio por Região.

	Dificuldade dos itens Parâmetro *b*					Erro Padrão da Diferença			
Item	**Sudeste (GR)**	**Nordeste (GF1)**	**Norte (GF2)**	**Sul (GF3)**	**C.Oeste (GF4)**	**(GF1-GR)**	**(GF2-GR)**	**(GF3-GR)**	**(GF4-GR)**
CAT1	-2,12	-2,40	-1,90	-2,05	-2,08	0,14	0,19	0,18	0,24
CAT2	-1,96	-2,15	-1,94	-2,14	-2,05	0,12	0,17	0,18	0,22
CAT3	-1,95	-2,20	-1,91	-2,04	-2,02	0,14	0,20	0,19	0,26
CAT4	-1,52	-1,64	-1,59	-1,63	-1,18	0,12	0,17	0,15	0,20
CAT5	-0,92	-1,17	-0,89	-0,95	-1,15	0,12	0,17	0,15	0,23
CAT6	-0,98	-0,94	-0,88	-1,03	-1,26	0,10	0,13	0,12	0,19
CAT7	-0,42	-0,62	-0,41	-0,66	-0,42	0,11	0,15	0,14	0,19
CAT8	-0,25	-0,10	0,09	-0,33	-0,07	0,08	0,11	0,10	0,14
CAT9	-0,05	-0,02	0,02	-0,08	0,06	0,08	0,11	0,10	0,13
CAT10	0,45	0,54	0,57	0,58	0,49	0,09	0,13	0,12	0,15
CAT11	0,64	0,76	0,71	0,88	0,61	0,06	0,10	0,09	0,11
CAT12	0,96	0,97	0,99	1,11	0,91	0,06	0,11	0,09	0,11
CAT13	1,01	1,06	1,04	1,12	1,20	0,07	0,11	0,10	0,12
CAT14	1,14	1,28	1,22	1,51	1,35	0,08	0,11	0,12	0,12
CAT15	1,50	1,45	1,41	1,81	1,40	0,12	0,16	0,21	0,14
SIT1	-2,64	-3,08	-3,37	-3,17	-2,65	0,20	0,28	0,27	0,34
SIT2	-2,27	-2,80	-3,00	-2,46	-2,46	0,17	0,28	0,25	0,30
SIT3	-2,13	-2,41	-2,29	-2,03	-2,16	0,18	0,26	0,22	0,31
SIT4	-1,91	-1,95	-1,87	-2,33	-1,86	0,15	0,22	0,23	0,27
SIT5	-0,93	-0,63	-0,71	-0,95	-0,60	0,12	0,16	0,15	0,20
SIT6	-0,91	-0,84	-0,63	-0,98	-1,19	0,12	0,17	0,16	0,24
SIT7	-0,11	0,25	0,04	-0,05	-0,08	0,11	0,15	0,14	0,19
SIT8	-0,51	-0,34	-0,44	-0,45	-0,42	0,10	0,13	0,12	0,17
SIT9	0,09	0,35	0,02	0,07	-0,07	0,10	0,15	0,13	0,18
SIT10	0,82	0,98	0,84	0,85	0,75	0,10	0,15	0,14	0,16
SIT11	1,37	1,60	1,52	1,36	1,32	0,15	0,22	0,18	0,20
SIT12	1,30	1,44	1,19	1,47	1,32	0,11	0,14	0,14	0,15
SIT13	1,45	1,54	1,52	1,67	1,36	0,13	0,18	0,17	0,14
SIT14	1,49	1,72	1,33	1,67	1,60	0,15	0,14	0,17	0,13

p=0,005). Já a análise das médias por região mostra diferenças mais relevantes, sobretudo na escala de raciocínio. A análise de variância (ANOVA) indicou a existência de alguma diferença entre as médias das cinco regiões (p=0,005). Essa diferença era esperada, considerando as diferenças regionais existentes em nosso país e independentemente de DIF nos instrumentos de mensuração.

Na Tabela 2 apresentamos o parâmetro *b* dos itens da escala de raciocínio de acordo com o grupo de referência (GR= masculino) e focal (GF=feminino). Podemos observar que dos 29 itens que compõem a escala de raciocínio, 11 possuem diferenças positivas, o que indica favorecimento ao GR e 18 possuem diferenças negativas, o que beneficia o GF. Entretanto, apenas 2 itens (CAT6 e SIT9) possuem diferença significativa entre os parâmetros (p=0,001).

A análise da CCI dos itens CAT6 e SIT9 mostra que a probabilidade de acerto foi maior para o grupo feminino do que para o grupo masculino (Figura 1). Nota-se também que o funcionamento diferencial é uniforme, ou seja, em nenhum ponto do eixo da habilidade o grupo masculino tem maior probabilidade de acerto do que o grupo feminino.

Já na Tabela 3, mostramos o parâmetro *b* dos itens da escala de execução, ainda de acordo com a variável sexo. Nesta escala, temos uma divisão mais igualitária entre os grupos, sendo que 16 itens parecem favorecer o GR (diferença positiva) e 15 o GF (diferença negativa). Dos 31 itens que compõem a escala de execução, apenas três apresentaram diferenças de parâmetro significativas (p=0,001): MOS4, MOS11 e PAD13. Ao contrário do ocorrido na escala de raciocínio, os três itens favorecem o GR, como pode ser visto na Figura 1.

Os DIFs encontrados na escala de execução também são uniformes, favorecendo ao longo de toda a escala de habilidade o grupo masculino (Figura 1).

Na Tabela 4, as análises de DIF dos itens da escala de raciocínio por região são apresentadas. Utilizamos como grupo de referência a região Sudeste. A partir do cálculo de significância (p=0,001), três itens apresentaram DIF: CAT8, CAT14 e SIT7. Em todos os casos, o GR (sudeste) foi beneficiado em relação às regiões norte, sul e nordeste, respectivamente. Vale destacar que nenhum item apresentou DIF para todas as regiões, simultaneamente nessa escala.

Apesar da análise de comparação dos parâmetros ter apontado diferença significativa para esses três itens da escala de raciocínio, a análise da CCI indica que as diferenças são bastante sutis (Figura 2). Em especial no item CAT14 verificamos a ausência de crianças na faixa de habilidade acima de 0,98, o que poderia explicar a diferença de funcionamento encontrada.

Na Tabela 5 estão presentes os parâmetros *b* dos itens da escala de execução por região. Verificamos a existência de dez itens com DIF (p=0,001): MOS11, PAD1, PAD4, PAD8, PAD9, PAD10, PAD11, PAD12, PAD13 e PAD14, sendo que o PAD14 apresentou DIF nas quatro regiões em relação à região Sudeste.

A análise gráfica dos itens com DIF na escala de execução (Figura 3) nos permite observar que dos dez itens, apenas dois apresentam DIFs mais relevantes (PAD4 e PAD14). O PAD4 possui um DIF uniforme, pois a probabilidade de acertar o item da região Centro-Oeste é sempre maior do que das demais regiões ao longo do espectro de habilidade. Já o PAD14, parece ter probabilidades de acerto bastante diferente para as 5 regiões. Em ordem, as regiões mais favorecidas foram: Norte, Sul, Nordeste, Sudeste e por fim, Centro-Oeste.

Nos demais oito itens que apresentaram DIF, a análise gráfica nos indica curvas muito semelhantes, o que nos leva a questionar se há realmente um funcionamento diferencial. No Item MOS11, por exemplo, houve DIF da região Centro-Oeste porque ela foi comparada com a região Sudeste, caso mudasse o grupo de referência, a diferença poderia não ser significativa. Outro fator que poderia estar relacionado é o quantitativo de crianças que responderam aos itens. Nota-se que a maioria dos itens com DIF são os do final dos subtestes e, portanto, mais difíceis e menos respondidos. Inclusive os itens PAD15 e PAD 16 foram retirados automaticamente das análises pelo programa Bilog-MG, por possuírem poucas respostas.

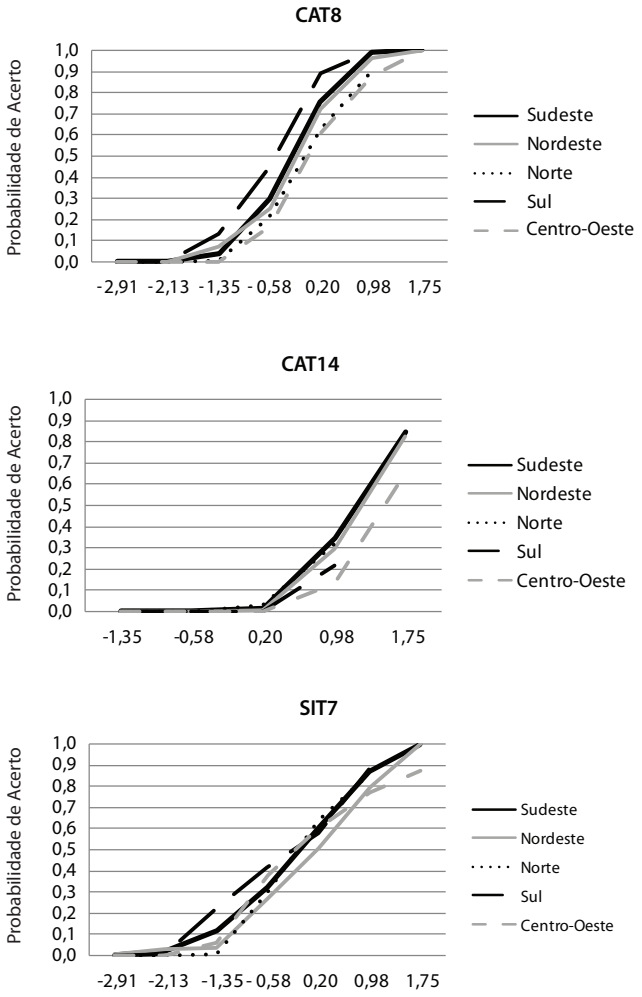

Figura 2. CCI dos Itens da Escala de Raciocínio que apresentaram DIF por Região.

Discussão

De modo geral, na análise por sexo, encontramos dois itens com DIF na escala de raciocínio e três itens na escala de execução. Considerando o fato que dificilmente uma prova ou teste serão totalmente isentos de DIF (Muñiz, 1997), esse quantitativo baixo de itens com funcionamento diferencial nos permite concluir que o teste SON-R 2½-7[a] é adequado para o uso tanto com meninas quanto com meninos.

Nota-se que nos itens com DIF, na escala de raciocínio, as meninas possuem uma probabilidade mais alta de acerto e, na escala de execução, os meninos passam a ter uma probabilidade mais alta. Isso nos remete a uma discussão teórica, uma vez que existe a concepção de que há uma superioridade masculina nos testes de habilidade espacial, enquanto que as mulheres possuem vantagem nas tarefas relacionadas à habilidade verbal (Hogan, 2006; Jardine & Martin, 1983). No SON-R 2½-7[a] não há subtestes verbais, mas a escala de raciocínio possui similaridades com a escala de processamento verbal (ambas avaliam raciocínio abstrato) e a escala de execução exige predominantemente o raciocínio espacial, o que condiz com a literatura e não necessariamente sugere impacto.

Com relação ao DIF por região, verificamos a existência de três itens com DIF na escala de raciocínio e 10 itens na escala de execução. Como se tratam de cinco regiões, a mudança do grupo de referência pode levar a resultados dife-

Tabela 5. Valores dos Parâmetros b dos Itens da Escala de Execução por Região.

Item	Dificuldade dos itens Parâmetro *b*					Erro Padrão da Diferença			
	Sudeste (GR)	Nordeste (GF1)	Norte (GF2)	Sul (GF3)	C.-Oeste (GF4)	(GF1-GR)	(GF2-GR)	(GF3-GR)	(GF4-GR)
MOS1	-3,00	-2,83	-2,49	-2,45	-2,91	0,33	0,38	0,36	0,48
MOS2	-1,63	-1,89	-1,59	-1,64	-1,75	0,12	0,17	0,15	0,21
MOS3	-1,44	-1,35	-1,12	-1,49	-1,45	0,12	0,16	0,16	0,23
MOS4	-1,23	-1,46	-1,15	-1,40	-1,48	0,08	0,10	0,10	0,15
MOS5	-1,20	-1,31	-1,19	-1,17	-1,22	0,09	0,12	0,11	0,18
MOS6	-0,85	-0,78	-0,81	-1,02	-0,80	0,08	0,11	0,10	0,16
MOS7	-0,46	-0,51	-0,53	-0,59	-0,42	0,07	0,10	0,09	0,15
MOS8	-0,26	-0,38	-0,32	-0,34	-0,09	0,06	0,09	0,08	0,12
MOS9	0,49	0,43	0,42	0,26	0,54	0,06	0,09	0,08	0,12
MOS10	0,81	0,93	0,65	0,70	0,96	0,06	0,09	0,08	0,10
MOS11	1,02	1,13	0,95	0,91	1,48	0,07	0,12	0,09	0,11
MOS12	1,48	1,62	1,40	1,14	1,79	0,12	0,21	0,13	0,14
MOS13	1,79	2,39	2,19	2,08	2,06	0,22	0,49	0,30	0,21
MOS14	1,72	1,91	1,98	2,36	2,14	0,15	0,70	0,21	0,20
MOS15	2,09	2,15	2,01	2,34	-	0,24	0,61	0,29	-
PAD1	-2,37	-2,28	-1,81	-2,49	-2,23	0,15	0,16	0,23	0,33
PAD2	-2,23	-2,13	-1,95	-2,18	-	0,14	0,19	0,18	-
PAD3	-2,38	-2,43	-2,23	-2,25	-2,48	0,16	0,22	0,19	0,21
PAD4	-2,30	-2,24	-2,10	-2,01	-2,78	0,13	0,17	0,13	0,10
PAD5	-2,26	-2,24	-2,02	-1,99	-3,13	0,16	0,21	0,18	0,29
PAD6	-1,54	-1,47	-1,45	-1,50	-1,77	0,09	0,12	0,11	0,64
PAD7	-1,46	-1,34	-1,62	-1,11	-1,64	0,10	0,15	0,12	0,47
PAD8	-0,77	-0,89	-1,04	-0,87	-1,22	0,08	0,11	0,10	0,15
PAD9	-0,50	-0,71	-0,67	-0,64	-0,45	0,06	0,09	0,08	0,18
PAD10	0,15	-0,14	-0,14	0,04	-0,05	0,06	0,09	0,08	0,16
PAD11	0,46	0,40	0,13	0,29	0,58	0,06	0,08	0,07	0,12
PAD12	0,60	0,46	0,22	0,34	0,71	0,06	0,08	0,07	0,12
PAD13	1,01	0,91	0,56	0,82	1,00	0,07	0,08	0,09	0,11
PAD14	1,05	0,86	0,49	0,65	1,53	0,06	0,07	0,08	0,11

No Centro-Oeste, o item MOS15 foi errado por todos e o items PAD2 acertado por todos, por isso foram automaticamente retirados das análises. Os itens PAD15 e PAD16 foram retirados das análises, pois poucas crianças chegaram a responder esses itens.

rentes. Assim, apesar de um quantitativo maior de itens com DIF entre as regiões, esses resultados devem ser interpretados com cautela, uma vez que se trata de um teste validado para a realidade brasileira com parâmetros psicométricos muito robustos (ver Jesus, 2009).

Na escala de execução houve um número maior de itens com DIF do que na escala de raciocínio. Esse resultado parece indicar que os estudos anteriores com o SON-R 5½-17 (Tellegen & Laros, 2004) e SON-R 2½-7[a] (Jesus, 2009), que buscaram analisar vieses nos itens dos testes SON, tiveram função importante para o aprimoramento do instrumento. Nesses estudos, foram feitas diversas modificações justamente nos itens da escala de raciocínio a fim de diminuir os vieses culturais.

Por outro lado, esses resultados também confirmam a afirmação de que nem todo item com DIF possui claramente um viés (Camilli & Shepard, 1994), uma vez que os itens da escala de execução são constituídos basicamente por figuras abstratas e geométricas, que geralmente são tidas como imparciais. Nesse caso, as diferenças entre os grupos podem-se dever a fatores aleatórios, não necessariamente relacionados a viés cultural.

Neste estudo, tínhamos como objetivo identificar a presença de DIF, de acordo com o critério de comparação do

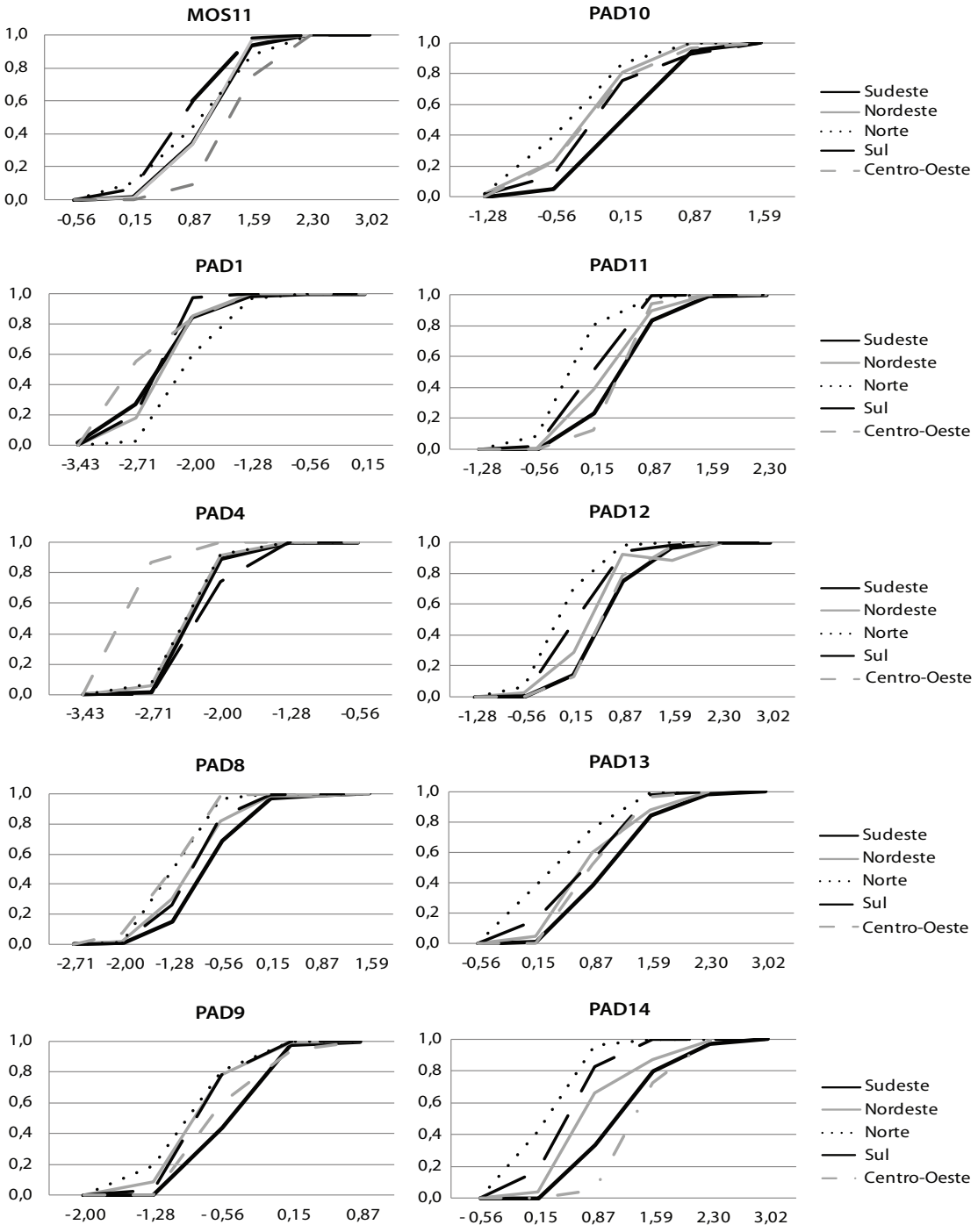

Figura 3. CCI dos Itens da Escala de Execução que apresentaram DIF por Região.

parâmetro *b* e com a análise da CCI (Cohen, Kim & Baker, 1993). Não obstante, seria interessante que novos estudos fossem realizados buscando aumentar o quantitativo de crianças em cada região ou utilizando outra técnica de detecção de DIF. É por meio da detecção de DIF que podemos aperfeiçoar os testes e evitar que grupos sejam desfavorecidos (Hambleton, 1989).

Em suma, esse estudo visa contribuir para o campo da avaliação psicológica de crianças, pois se verifica nessa área muitas lacunas que precisam ser atendidas (Noronha, Primi & Alchieri, 2004; Oakland, Wechsler, Bensuan & Stafford, 1994). Há lacunas metodológicas, que dependem de nós pesquisadores para serem sanadas. Dessa forma, é preciso reconhecer possíveis falhas existentes nos testes e buscar aprimorar. Existem também lacunas na prática profissional, relacionadas à escolha e ao modo de utilização dos testes. Com relação a essa última, é preciso chamar a atenção para o fato de que ela somente será suprida com o apoio de toda a classe de psicólogos usuários de testes.

A melhoria da qualidade dos testes psicológicos deve ser uma exigência de todos.

Referências

Alves, C. B. (2004). *Diferentes técnicas no estudo do Funcionamento Diferencial dos Itens: uma análise com os dados do Exame Nacional de Cursos.* Dissertação de mestrado, Universidade de Brasília, Brasília.

Alves, I. C. B. (2002). Instrumentos disponíveis no Brasil para avaliação da inteligência. In R. Primi (Ed.), *Temas em avaliação psicológica* (pp. 80-102). Campinas, SP: Impressão Digital do Brasil Gráfica e Editora Ltda.

Andriola, W. B. (2000). Funcionamento diferencial dos itens (DIF): estudo com analogias para medir raciocínio verbal. *Psicologia: Reflexão e Crítica, 13*, 475-483.

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

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publishers.

Cohen, A. S., Kim, S., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17,* 335-350.

Hambleton, H. K., Swaminatham, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (pp. 147-200). New York: Macmillan.

Hambleton, R. K. (1990). Item response theory: introduction and biography. *Psicothema, 11*, 97-107.

Hogan, T. P. (2006). *Introdução à prática de testes psicológicos*. Rio de Janeiro: LTC – Livros e Técnicos e Científicos Editora S.A.

Hu, S., & Oakland, T. (1991). Global and regional perspectives on testing children and youth: an empirical study. *International study of psychology*, *26*, 329-344.

Jardine, R., & Martin, N. G. (1983). Spatial ability and throwing accuracy. *Behavior Genetics, 13*, 331-340.

Jesus, G. R., de (2009). *Normatização e validação do teste não-verbal de inteligência SON-R 2½-7[a] para o Brasil.* Tese de Doutorado, Universidade de Brasília, Brasília.

Laros, J. A., Tellegen, P. J., Jesus, G. R., de & Karino, C. A. (no prelo). *SON-R 2½-7[a], Teste não-verbal de inteligência. Manual com normatização e validação brasileira.*

Lord, E. M. (1980). *Applications of item response theory to practical testing problems.* New Jersey: Lawrence Erlbaum.

Muñiz, J. (1997). *Introducción a la teoría de respuesta ao los items.* Madrid: Pirâmide.

Muñiz, J., Prieto, G., Almeida, L., & Bartram, D. (1999). Test use in Spain, Portugal and Latin American Countries. *European Journal of Psychological Assessment, 15*, 151-157.

Noronha, A. P. P., & Alchieri, J. C. (2002). Reflexões sobre os instrumentos de avaliação psicológica. In R. Primi (Ed.), *Temas em avaliação psicológica* (pp. 7-16). Campinas, SP: Impressão Digital do Brasil Gráfica e Editora Ltda.

Noronha, A. P. P., Primi, R., & Alchieri, J. C. (2004). Parâmetros psicométricos: uma análise dos testes psicológicos comercializados no Brasil. *Psicologia: Ciência e Profissão. 24*, 88-99.

Noronha, A. P. P., Vendramini, C. M. M., Canguçu, C., Souza, C. V. R. de, Cobêro, C., Paula, L. M. de, Franco M. O. de, Lima, O. M. P. de, Guerra, P. B. C. de, & Filizatti, R. (2003). Propriedades psicométricas em manuais de testes de inteligência. *Psicologia em Estudo, 8*, 93-99.

Oakland, T., Wechsler, S., Bensuan, E., & Stafford, M. (1994). The construct of intelligence among Brazilian children: An exploratory study. *School Psychology International, 15,* 361-370.

Tellegen, P. J., & Laros, J. A. (2004). Cultural bias in the SON-R test: Comparative study of Brazilian and Dutch children. *Psicologia: Teoria e Pesquisa, 20*, 103-111.

Tellegen, P., Winkel, M., Wijnberg-Williams, B. J., & Laros (1998). *Snijders-Oomen Nonverbal Intelligence Test SON-R 2½-7: Manual and research report*. The Netherlands: Swets Test Publishers.

Tellegen, P. J., & Laros, J. A. (2005). *Fair assessment of children from cultural minorities: a description of the SON-R non-verbal intelligence tests*. Slovakia: Paper presented at the UNESCO seminar.

Thorndike, R. L., Hagen, E. P., & Sattler J. M. (1986). *The Stanford-Binet intelligence scale: Fourth edition technical manual.* Chicago: The Riverside Publishing Company.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Testing in culturally heterogeneous populations: when are cultural loadings undesirable? *European Journal of Psychological Assessment, 8*, 17-24.

Wechsler, D. (2002). *WISC-III – Escala de Inteligência Wechsler para Crianças*: Manual/David Wechsler, 3ª. Ed.; Vera Lúcia Marques de Figueiredo. São Paulo: Casa do Psicólogo.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473-492.

Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

Recebido em 14.03.2010
Primeira decisão editorial em 24.08.2010
Versão final em 30.09.2010
Aceito em 26.10.2010 ■