Boletim de Psicologia, 2013, Vol. LXIII, Nº 138: 001-010

FIDEDIGNIDADE DO WISC-III COM BASE NA CONCORDÂNCIA ENTRE AVALIADORES[1]

JACIANA MARLOVA GONÇALVES ARAUJO
VERA LÚCIA MARQUES DE FIGUEIREDO
Universidade Católica de Pelotas - RS - Brasil

RESUMO

Este trabalho pretendeu estabelecer os índices de fidedignidade do WISC-III, utilizando a técnica de concordância entre avaliadores, tendo em vista, que a subjetividade dos profissionais que utilizam o teste pode interferir diretamente em seus resultados. Foram selecionados, aleatoriamente, da amostra de padronização ao contexto brasileiro, protocolos dos testes de três crianças com a mesma idade, os quais foram pontuados por 42 psicólogos de diferentes estados do Brasil. Baseando-se nos escores totais, a precisão foi calculada através do índice de correlação intra-classe e os coeficientes, em geral, foram considerados fortes. Os subtestes não-verbais, com exceção de Completar Figuras, apresentaram os índices mais altos, sendo que as correlações dos subtestes verbais foram menores. Os resultados evidenciaram que a correção objetiva dos itens propicia maior consenso entre os avaliadores e quando as respostas a serem pontuadas envolvem verbalizações, existem maiores dificuldades.

Palavras-chave: *Concordância entre avaliadores; WISC-III; fidedignidade; avaliação da inteligência.*

ABSTRACT

WISC-III RELIABILITY BASED IN INTER-RATER AGREEMENT
This work aimed to establish Wechsler Intelligence Scale for Children – Third Edition (WISC-III) reliability indexes through the inter-rater agreement technique, keeping in mind that the scores could be influenced by subjective factors of the raters. Three WISC-III protocols of children of the same age were randomly selected from the Brazilian standardization sample and were scored by 42 psychologists from different States of Brazil. Reliability was assessed through the intraclass correlation index, on the base of total scores. Coefficients were considered strong in general. All nonverbal subtests, except Picture Completion, presented higher indexes than the verbal subtests. Results suggest that objective scoring propitiates greater consensus between the raters, and verbal answers produce more scoring difficulties.

Key words: *Inter-rater agreement; WISC-III; reliability; intelligence assessment.*

Endereço para correspondência: Universidade Católica de Pelotas, Curso de Psicologia. Rua Almirante Barroso, número 1202, Centro. Pelotas - RS, Brasil, CEP 96010-280. E-mail: jacianamga@hotmail.com; verafig@terra.com.br

[1] Este trabalho contou com o financiamento da Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul - FAPERGS.

INTRODUÇÃO

A utilização de medidas psicológicas padronizadas visa que a investigação tenha o máximo de proximidade com a realidade e é a maneira mais precisa e objetiva que se tem para estudar as diferenças individuais. Poucas décadas após o surgimento dos primeiros testes psicométricos que se propunham a medir a inteligência, já havia outros, cuja finalidade era medir aptidões específicas, interesses e personalidade. A larga utilização e o avanço rápido no desenvolvimento dos testes trouxeram, também, as críticas de que estes seriam passíveis de erro, já que uma indisposição do examinando seria capaz de alterar sensivelmente os resultados, assim como a influência do examinador e das próprias características do instrumento (Medeiros, 1999; Pasquali, 2010). Segundo Shrout e Fleiss (1979), a maior parte das medidas em ciências comportamentais envolve um erro de medição, e os julgamentos e avaliações, feitos por humanos sofrem de forma especial desse tipo de problema.

Identificada a possibilidade de erro, o primeiro cuidado é reconhecer suas fontes e buscar controlá-las (Vianna, 1976). Segundo Medeiros (1999) e McIntire e Miller (2000), entre as fontes de erro na psicometria identificam-se: o examinando, que pode estar fora de seu estado habitual na ocasião da prova; o instrumento de medida, por vezes inadequado para o caso ou impreciso, que oferecerá dados não confiáveis e, ainda, as condições de aplicação que podem influenciar o examinando. Quanto às condições para aplicação: cabe ao aplicador zelar para que elas sejam ótimas e certificar-se de que o examinando se encontre motivado pela execução da tarefa. Também é função do aplicador garantir a adequação do teste e ter conhecimento de sua precisão, para assegurar a confiança nos resultados.

Outra preocupação dos examinadores refere-se à influência de sua própria subjetividade como fonte de erro, pois ela não pode interferir na mensuração sob pena de invalidá-la. Para Medeiros (1999), entre as fontes de erro a subjetividade é a que pede maior atenção. Se ela for mantida sob controle, examinadores diferentes poderão aplicar o mesmo teste aos mesmos examinandos e julgar em separado as suas respostas, que chegarão a resultados iguais. Assim, também se eleva a fidedignidade de um teste restringindo a influência da subjetividade na sua construção, no seu emprego e especialmente na interpretação dos resultados.

Glasser e Zimmerman (1977) afirmam que as Escalas Wechsler de Inteligência para Crianças (WISC) não envolvem pontuações objetivas nas provas verbais, uma vez que exigem certo juízo do avaliador. O subteste Vocabulário, por exemplo, é difícil de corrigir e pontuar, devido à interferência da subjetividade (Glasser & Zimmerman, 1977; Cayssials, 2000). Segundo Wechsler (1991), no WISC-III a correção de determinados itens é objetiva e exige pequena ou nenhuma interpretação de critérios por parte do examinador. Entretanto, em subtestes como Compreensão, Semelhanças, Vocabulário e Informação, os itens exigem certo julgamento. Por essa razão, Cunha (2000) sugere que, para evitar a subjetividade do avaliador na correção e assegurar o menor número de erros possíveis na apuração dos resultados, os protocolos devem ser corrigidos por mais de um profissional, além do próprio aplicador. Segundo Sattler (1992), um estudo cuidadoso deve ser feito nos critérios de pontuação para ajudar a reduzir erros.

O estudo feito com relação à confiabilidade dos escores do WISC-III, para as crianças brasileiras (Wechsler, 2002) consistiu em uma única aplicação do teste, com exceção dos subtestes de

velocidade (Código e Procurar Símbolos), para os quais foram realizadas duas aplicações da mesma prova, empregando os índices de consistência interna, correlação teste-reteste, estimativa do erro padrão de medida (EPM) e determinação dos intervalos de confiança (IC). O método das duas metades utilizado na amostra de padronização americana (Wechsler, 1991), assim como a fidedignidade entre avaliadores obtida para os subtestes Semelhanças, Vocabulário e Compreensão não foram empregados para o estudo brasileiro.

As técnicas empregadas relacionam-se ao teste propriamente dito, sem considerar a influência do julgamento do avaliador, que como apontado anteriormente, é um dos fatores que afetam a fidedignidade do instrumento. Em testes não-objetivos, nos quais a opinião dos avaliadores é um fator relevante na determinação dos escores, é necessário o estabelecimento do nível de concordância entre os mesmos. O método consiste em um tipo de correlação entre os escores obtidos na avaliação do mesmo conjunto de dados, por dois ou mais examinadores (Guilford, 1954; Pasquali, 1999; McIntire & Miller, 2000; Anastasi & Urbina, 2000).

Por se tratar de uma técnica estatística bastante útil na verificação da qualidade psicométrica dos instrumentos, a fidedignidade entre avaliadores tem sido difundida e utilizada em diversas áreas (Hedricks, Robie & Oswald, 2013; Smith, et al., 2013; Steenson, Vivanti & Isenring, 2013; Storch, et al., 2012). Segundo Bartko e Carpenter (1976) há várias formas de estabelecer os índices de fidedignidade nesses casos. Os autores sugerem que para este fim algumas técnicas são mais eficazes que outras. Assim, são consideradas "pobres" na determinação do coeficiente de fidedignidade técnicas como: análise do percentual de concordância; chi-quadrado; correlação produto-momento e correlação de ordem, bem como a simples apresentação dos dados de determinado fenômeno e sua comparação com a avaliação obtida. Como técnicas mais "sofisticadas" ou mais precisas, os autores apresentam diferentes métodos, dependendo do número de avaliadores e do tipo de dados. Para os casos que envolvem dois avaliadores e dados qualitativos dicotômicos referem-se: ao Kappa; ao Kappa Ponderado; e ao Coeficiente de Correlação Intraclasse (internacionalmente identificado pela sigla ICC, abreviação de *Intraclass Correlation Coefficient*). No caso de dois avaliadores, o ICC é apropriado, apenas quando os totais marginais (somatório de linha e de coluna) são iguais. Quando os dados são dicotômicos e julgados, por mais de dois avaliadores, é apropriado o uso do ICC e do Kappa Generalizado. O Kappa Generalizado é útil também na análise, por mais de dois avaliadores, de dados politômicos. Quando se dispõe de dados quantitativos, o ICC é o cálculo mais indicado.

Segundo Shrout e Fleiss (1979), o ICC é obtido através de uma razão de variâncias. Há diferentes modos de calcular o coeficiente de correlação intraclasse, que produzem resultados bastante diversos, quando aplicados sobre os mesmos dados; sendo que cada uma destas formas é apropriada para uma situação específica, que depende da maneira como o experimento foi realizado e da intenção do estudo. Para o caso onde vários avaliadores identificam a presença ou ausência de uma categoria ou outra, arbitrariamente designadas (0 ou 1) o ICC é uma medida adequada da fidedignidade ou, caso não se possua o mesmo número de avaliadores para cada caso em avaliação, pode-se utilizar o método ANOVA ICC (ICC via análise de variância). Entre os diferentes tipos de ANOVA ICC encontra-se a ANOVA dupla do tipo Sujeito X Avaliador, nela cada sujeito recebe notas dadas pelo mesmo conjunto fixo de k avaliadores. Neste caso, é considerado o número de avaliadores e o tipo de dados, não

havendo exigência do mesmo número de avaliadores para cada caso, admitindo a existência de itens sem resposta (Bartko & Carpenter, 1976).

Para o estudo americano, sobre a concordância entre avaliadores do WISC-III com os subtestes Semelhanças, Vocabulário e Compreensão, foram selecionados 60 protocolos da amostra de padronização, os quais foram pontuados por quatro avaliadores independentemente. Na análise, foi usado um tipo de correlação intra-classe, que avaliou a concordância entre juízes, levando em conta a leniência do examinador (Shrout & Fleiss, 1979, citado por Wechsler, 1991). Os coeficientes foram de 0,94 para o subteste Semelhanças, 0,92 para Vocabulário e 0,90 para Compreensão. Quando os escores totais dos subtestes foram utilizados nas análises, as concordâncias obtidas entre avaliadores foram de 0,98 para Semelhanças, 0,98 para Vocabulário e 0,97 para Compreensão (Wechsler, 1991). Segundo o autor, os subtestes, apesar de exigirem julgamento do examinador, podem ser pontuados de forma confiável. No estudo de Van Noord e Prevatt (2002) coeficientes entre 0,97 e 1 foram encontrados na análise de 110 protocolos do WISC-III, por dois avaliadores. Os subtestes, em que ocorreram mais erros de pontuação foram: Compreensão, Vocabulário, Semelhanças e Informação. Resultados semelhantes foram encontrados na pesquisa de Loe, Kadlubek e Marks (2007), em que dois avaliadores pontuaram 51 protocolos do WISC-IV e obtiveram concordância entre as pontuações em 92% dos casos.

O objetivo do presente estudo foi determinar os coeficientes de concordância entre avaliadores, por meio do índice de correlação intraclasse, para todos os subtestes do WISC-III BR como forma de dar continuidade aos estudos sobre o instrumento. Este tipo de evidência é recomendado, quando o teste envolve subjetividade na avaliação, devendo os manuais sempre relatar essa informação, quando apropriado.

MÉTODO

Participantes

O estudo foi realizado com uma amostra de conveniência, obtida a partir de contatos com psicólogos da área de avaliação, que também indicaram outros profissionais. No total, foram contatados 60 psicólogos de diferentes estados do Brasil, com prática na utilização do teste WISC-III. Considerando as perdas, apenas 70% deles (N= 42) efetivamente participaram do estudo. Os colaboradores foram de ambos os sexos e residentes em quatro diferentes regiões do país. Decidiu-se comparar as pontuações de vários avaliadores com a intenção de que os resultados desse estudo se aproximassem ao máximo de um panorama real de como os profissionais brasileiros pontuam os protocolos do WISC-III em sua prática cotidiana. Diferentemente, nos estudos feitos usualmente nas pesquisas de padronização, muitos protocolos são pontuados por um pequeno grupo de especialistas alcançando, em geral altos índices de fidedignidade, por se tratar de uma situação ideal.

Material

Foram selecionados aleatoriamente, do banco de dados da pesquisa de padronização do WISC-III, ao contexto brasileiro, seis protocolos do teste com o registro das respostas colhidas em

tal ocasião. Optou-se por um reduzido número de questionários para que a tarefa de pontuar não foi demasiadamente exaustiva para os avaliadores, posto que a intenção era de que o maior número possível de avaliadores participasse da pesquisa. Não foram considerados os subtestes de Velocidade, Códigos e Procurar Símbolos, que exigiriam maior tempo dos avaliadores no processo de correção e o uso de crivos. Juntamente com os protocolos, foram enviadas aos profissionais: a) uma ficha para dados de identificação pessoal e informações profissionais; b) termo de consentimento livre e esclarecido; c) uma carta contendo orientações sobre o procedimento para execução da tarefa.

Procedimento

O estudo foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Católica de Pelotas, protocolo nº. 0601-2/05, em 30/12/2006. Todos os participantes assinaram um Termo de Consentimento Livre e Esclarecido antes de sua inclusão na amostra.

O material foi encaminhado via correio e a tarefa dos psicólogos consistiu em pontuar as respostas registradas nos seis protocolos. No presente estudo foram utilizados somente três protocolos cujos respondentes tinham a mesma idade (nove anos). Os dados considerados para o cálculo do coeficiente de fidedignidade foram os escores totais de cada subteste, ou seja, a soma dos pontos atribuídos pelos avaliadores a cada item do subteste. Quando as somas não foram realizadas pelos juízes, elas foram efetuadas pelas pesquisadoras.

Foram usadas medidas de tendência central e análise de frequência para descrever a amostra. Como forma de análise exploratória, calculou-se o percentual de concordância, entre avaliadores, considerando o número de juízes que encontraram o mesmo escore total, em cada subteste. Para estimar os coeficientes de fidedignidade foi utilizado o método de Correlação Intraclasse e ANOVA dupla do tipo Sujeito x Avaliador, calculados nos programas *Microsoft Office Excel* 11.0 e *Statistical Package for the Social Sciences* 13.0. A interpretação dos coeficientes seguiu os seguintes critérios: r= 0,10 a 0,30 fraca; r= 0,40 a 0,60 moderada; r= 0,70 a 0,99 forte e r=1 perfeito (Dancey & Reidy, 2006).

RESULTADOS E DISCUSSÃO

Os psicólogos, que participaram como avaliadores, eram 95% (N=40) do sexo feminino, com idade média de 39 anos (DP=10,84), sendo que 47% (N=20) residiam na região sudeste, 43% (N=18) no sul, 5% (N=2) no centro-oeste e 5% (N=2) no nordeste. Entre eles, 47% (N=20) possuíam mestrado, 26% (N=11) alguma especialização e 22% (N=9) doutorado, sendo que 5% da amostra (N=2) não responderam a este item do questionário. Os contextos mais frequentes de utilização do teste, pelos avaliadores, foram clínica, pesquisa em atividades como docência e na prática de avaliação psicológica. Um índice de 85% (N= 35) dos avaliadores relatou experiência tanto na aplicação, quanto na correção do WISC-III, sendo que 60% (N=24) deles utilizam somente o manual como recurso para a correção. Os demais recorrem a material complementar, como: artigos e apostilas de cursos. Os psicólogos tinham em média 15,5 anos de formados (DP=10,5).

Com o objetivo de evidenciar a dispersão dos escores, emitidos pelos avaliadores, as Tabelas 1 e 2 demonstram a amplitude dos totais mínimos e máximos de pontos, em cada subteste, para cada

um dos três protocolos considerados na análise. Também apresenta a Moda (Mo) indicando o escore de maior frequência, com o respectivo percentual de avaliadores que obteve tal pontuação. Apesar das frequências serem consideradas medidas pobres de fidedignidade por Bartko e Carpenter (1976), elas fornecem algum subsídio para a compreensão da variabilidade dos escores produzida por fatores estranhos ao construto, que, neste caso, se refere ao julgamento do avaliador. Os dados demonstram que os subtestes de Execução apresentam maior consenso nas pontuações, do que os do conjunto Verbal, e consequentemente menor influência de subjetividade.

Tabela 1. Estimativa da dispersão e concordância nas pontuações dos avaliadores, quanto ao escore total para os subtestes verbais

Caso	Informação			Semelhanças			Aritmética			Vocabulário			Compreensão			Dígitos		
	a	Mo	%	a	Mo	%	a	Mo	%	a	Mo	%	a	Mo	%	a	Mo	%
1	2	10	54	5	13	46	0	14	100	14	20	20	12	16 e 18	20	0	11	100
2	1	9	79	5	11	38	3	14	98	9	22	33	4	10	41	0	9	100
3	2	12	81	5	9	41	1	15	98	14	24	29	7	14	49	2	9	95

a= amplitude; Mo= moda; %= percentual de avaliadores

Tabela 2. Estimativa da dispersão e concordância nas pontuações dos avaliadores, quanto ao escore total para os subtestes de execução

Casos	Completar Figuras			Arranjo de Figuras			Cubos			Armar Objetos		
	a	Mo	%	a	Mo	%	a	Mo	%	a	Mo	%
1	7	16	85	4	25	95	0	23	100	14	24	81
2	2	15	93	0	15	100	0	31	100	10	26	88
3	2	19	76	3	23	93	0	26	100	2	35	95

a= amplitude; Mo= moda; %= percentual de avaliadores

A amplitude (a) média para cada subteste foi calculada como indicador da variabilidade das pontuações. Entre os subtestes verbais, em Dígitos (a= 0,66), Aritmética (a=1,33) e Informação (a=1,66) observou-se maior consenso na pontuação dos escores. Maior variabilidade ocorreu em Vocabulário (a = 12,33), Compreensão (a = 7,66) e Semelhanças (a = 5) demonstrando serem os que sofrem maior influência da subjetividade do avaliador. Segundo Wechsler (1991), a pontuação desses três subtestes e de alguns itens do subteste Informação exigem julgamento. Nos subtestes Vocabulário, Semelhanças e Compreensão as dúvidas geralmente ocorrem devido à pontuação variar de 0 a 2, dependendo da qualidade da resposta. Também, muitas respostas dadas pelas pessoas avaliadas não constam no manual, fazendo com que a pontuação destas envolva a subjetividade do avaliador. As respostas-modelo (disponíveis no manual do teste após cada item) consistem nas respostas reais dadas pela amostra de padronização americana e que foram traduzidas para o manual brasileiro.

Entretanto, na aplicação do teste, surgem muitas respostas diferentes, difíceis de avaliar sua equivalência com as respostas-modelo.

Nos subtestes de execução que, segundo Wechsler (1991), são objetivos, também houve variabilidade nos escores. Cubos apresentou a menor variabilidade (a = 0), seguido por Arranjo de Figuras (a = 2,33) e Completar Figuras (a = 3,66), enquanto em Armar Objetos observou-se a maior dispersão (a = 8,66). Os erros mais comuns dos avaliadores, na correção dos protocolos, em Arranjo de Figuras e Armar Objetos foram somas erradas, assinalamentos de pontuações inadequadas em relação ao tempo e pontuações apesar do fracasso no desempenho do item, problemas que são decorrentes de falta de atenção.

Em Completar Figuras, os avaliadores mencionaram que tiveram dúvidas para pontuar as respostas, em função da falta de registro pelos examinadores, tanto da expressão "apontar correto" (AC) como, de inquéritos (Q) para esclarecimento das respostas. As queixas foram pertinentes mostrando a relevância da padronização na aplicação dos testes, procedimento que exige uniformidade ao aplicar e pontuar o instrumento (Anastasi & Urbina, 2000). Entretanto, no mesmo subteste, os avaliadores questionaram de forma não pertinente algumas respostas facilmente identificáveis como incorretas ou corretas, respectivamente: item 21 (telefone) – "a coisa de ligar na tomada" e item 22 (banheira) – "o furo de onde a água sai". Tais verbalizações não necessitariam de questionamento, pois expressam claramente a resposta. Segundo o manual do teste algumas crianças podem não saber ou não serem capazes de nomear corretamente a parte omitida e podem usar um sinônimo ou usar suas próprias palavras para descrevê-la (Wechsler, 2002). Outro problema foi a pontuação com zero ponto em itens onde aparecia apenas o registro "AC", apesar da regra de que, quando a criança indicar corretamente a parte omitida apontando, deve-se pontuar como resposta correta (Wechsler, 2002). Nesse sentido é importante lembrar que o domínio do teste pelo examinador é um dos fatores que afetam a fidedignidade e, por essa razão, o manual do WISC-III (Wechsler, 1991 e 2002) apresenta no terceiro capítulo, as considerações gerais para aplicação do teste que, constituem regras a serem automatizadas pelos profissionais.

A Tabela 3 apresenta os coeficientes de fidedignidade, resultantes da análise pelo índice de correlação intra-classe e suas respectivas interpretações, segundo Dancey e Reidy (2006). Calculou-se os índices para todos os subtestes, exceto os de velocidade de processamento (Códigos e Procurar Símbolos).

Tabela. 3 Coeficientes de correlação intraclasse na pontuação do teste WISC-III

Subtestes	ICC (r)	Interpretação*
Cubos	1,00	Perfeito
Arranjo de Figuras	0,99	Forte
Dígitos	0,97	
Armar Objetos	0,91	
Completar Figuras	0,88	
Informação	0,88	
Aritmética	0,79	
Compreensão	0,76	
Semelhanças	0,74	
Vocabulário	0,71	

*Interpretação dos coeficientes de correlação (Segundo Dancey e Reidy, 2006)

Na Tabela 3 os subtestes foram dispostos em ordem decrescente, em função dos coeficientes de fidedignidade obtidos. Os valores variaram, indicando maior ou menor subjetividade na correção, entretanto todos, com exceção do subteste Cubos que mostrou correlação perfeita, apresentaram índices considerados fortes. Os dados não podem ser comparados aos do manual americano do WISC-III, pois este não apresenta os coeficientes para todos os subtestes, a não ser para os que exigem maior julgamento do avaliador: Vocabulário; Semelhanças e Compreensão.

Os subtestes não-verbais apresentaram coeficientes mais altos, demonstrando que, a correção objetiva dos itens propicia maior consenso entre os avaliadores e consequentemente, menor variância erro. A expectativa inicial era de que nos subtestes de execução fossem encontradas correlações perfeitas, uma vez que a pontuação depende apenas da execução correta ou incorreta do item e do tempo de execução, entretanto, observou-se em Armar Objetos, erros devido à falta de atenção dos examinadores, na tarefa de pontuar os itens. Quanto a Completar Figuras, os erros ocorreram por dúvidas, em função das respostas envolverem verbalizações.

Ainda conforme a Tabela 3, os valores abaixo de 0,90 sugerem maior dificuldade dos avaliadores na pontuação dos itens dos subtestes verbais, evidenciando a necessidade de consultar os pares, para minimizar as dúvidas. Os coeficientes mais baixos, relacionados aos subtestes Vocabulário, Semelhanças e Compreensão diferem dos obtidos por Wechsler (1991), Van Noord e Prevatt (2002), e Loe, Kadlubek e Marks (2007) (utilizando o WISC-IV), os quais encontraram valores próximos à correlação perfeita. Torna-se importante ressaltar, o número pequeno de avaliadores envolvidos nos referidos estudos (seis, dois e dois respectivamente), resultando em baixa variabilidade e consequentemente maior precisão. Os dados observados no presente estudo vão ao encontro dos resultados de Glasser e Zimmerman (1977), Sattler (1992) e Cayssials (2000) que identificaram os mesmos subtestes, como sendo os mais difíceis de avaliar, recomendando um estudo cuidadoso nos critérios para a pontuação. Quanto ao subteste Aritmética, a correção é objetiva, uma vez que, depende de uma resposta numérica decorrente do cálculo executado pelo examinando. A própria folha de registro apresenta a resposta certa, possibilitando a correção imediata e minimizando a possibilidade de erros. Entretanto, observou-se um coeficiente abaixo do esperado, posto que, houve grande variabilidade nos escores. Em Aritmética os sete primeiros itens não são aplicados para os sujeitos de 9 anos ou mais, sem suspeita de atraso mental, porém, eles devem ser computados na soma do escore total, o que não foi considerado por alguns juízes do presente estudo.

CONCLUSÃO

Os dados da presente pesquisa proporcionaram evidências de fidedignidade do WISC-III ao contexto brasileiro, comprovadas pela técnica de concordância entre avaliadores. Quanto aos participantes, contou-se com uma maior representatividade de profissionais das regiões Sul e Sudeste, todos com experiência no referido instrumento.

Os resultados mostraram que existe certa variabilidade na pontuação dos escores do WISC-III e que ela é maior nos subtestes verbais, principalmente em: Vocabulário, Semelhanças e Compreensão. Os itens nesses subtestes suscitam um grande espectro de respostas difíceis de corrigir, tanto devido às pontuações politômicas (0, 1 ou 2, conforme a qualidade da resposta), quanto pela dificul-

dade de fazer equivalência entre as respostas dadas pelo examinando e as poucas respostas-modelo disponíveis no manual. Também nos subtestes não-verbais, considerados objetivos para a correção, observou-se falta de concordância, essencialmente por desatenção dos avaliadores. As falhas individuais, na pontuação e na subjetividade contribuem como fontes de erros nos escores totais do teste. Entretanto, os coeficientes de fidedignidade foram elevados, indicando índices significativos de precisão.

Um aspecto positivo do estudo foi o número maior de avaliadores envolvidos, em relação à pesquisa realizada por Wechsler (1991), oportunizando um dimensionamento da situação real de manejo do teste, pelos profissionais brasileiros. Pode-se considerar a inclusão dos subtestes não verbais no presente estudo uma contribuição significativa, pois eles não foram incluídos nas análises da versão americana e, conforme os dados, também são influenciados pela subjetividade dos avaliadores. Uma limitação do estudo refere-se a dados omissos: alguns protocolos não apresentavam o escore total dos itens, provavelmente pelo fato de as instruções aos avaliadores não alertarem sobre a necessidade dessa tarefa. A decisão das pesquisadoras foi de efetuar a somatória. Ainda nas situações em que os itens iniciais não haviam sido aplicados, em função da idade do sujeito, alguns avaliadores não os pontuaram. Nesses casos, optou-se por considerá-los como respostas corretas, com a soma realizada pelas pesquisadoras. Apesar de reconhecer que esse procedimento poderia diminuir a variabilidade na soma dos escores, optou-se por esse artifício, para minimizar *missings* na análise dos dados, uma vez que o cálculo do coeficiente intraclasse é baseado nos escores totais. Sugere-se que os profissionais busquem capacitação para utilizar de forma correta o WISC-III e consultem os pares, ou materiais suplementares ao manual, quando se depararem com dúvidas de pontuação.

O presente estudo complementou a pesquisa de padronização do WISC-III ao contexto brasileiro, que utilizou os métodos de consistência interna, estabilidade temporal e erro padrão de medida (para estimar a variância de erro), técnicas de análise relacionadas ao próprio teste. Todos os coeficientes, independentemente do método utilizado, mostraram-se satisfatórios, sugerindo que o WISC-III é um instrumento confiável para estimar a inteligência de crianças e de adolescentes.

REFERÊNCIAS

Anastasi, A. & Urbina, S. (2000). Fidedignidade. In: A. Anastasi & S. Urbina, *Testagem psicológica*. (pp. 84-105). Porto Alegre: Artes Médicas.

Bartko, J. & Carpenter, W. (1976). On the methods and theory of reliability. *Journal of Nervous and Mental Disease, 163* (5), 307-317.

Cayssials, A. N. (2000). *La escala de Inteligencia WISC-III en la evaluación psicológica infanto-juvenil*. Buenos Aires: Paidós.

Cunha, J. A. (2000). Escalas Wechsler. In: J. A. Cunha et al., *Psicodiagnóstico V.* (pp. 529-602). Porto Alegre: Artes Médicas.

Dancey, C.P. & Reidy, J. (2006). *Estatística sem Matemática para Psicologia usando SPSS para Windows*. Porto Alegre: Artes Médicas.

Glasser, A.J. & Zimmerman, I.L. (1977). *Interpretación clínica de la Escala de Inteligencia de Wechsler para Niños (WISC).* Madrid: Tea.

Guilford, J. (1954). *Psychometric methods*. New York: McGraw-Hill.

Hedricks, C.A.; Robie, C. & Oswald, F.L. (2013). Web-based multisource reference checking: An investigation of psychometric integrity and applied benefits. *International Journal of Selection and Assessment, 21 (1), 99-110*.

Loe, S.A.; Kadlubek, R.M. & Marks, W. J. (2007). Administration and scoring errors on the WISC-IV among graduate student examiners. *Journal of Psychoeducational Assessment, 25* (3), 237-247.

McIntire, S. & Miller, L. (2000). *Foundations of psychological testing*. Boston: McGraw-Hill.

Medeiros, E.B. (1999). *Medidas psico e lógicas: Introdução à psicometria.* Rio de Janeiro: Ediouro.

Pasquali, L. (1999). Testes referentes a construto. In: L. Pasquali (Org.), *Instrumentos psicológicos: Manual prático de elaboração*. (pp. 37-71). Brasília: LABPAM;IBAPP.

Pasquali, L. (2010). *Instrumentação psicológica: Fundamentos e práticas.* Porto Alegre: Artmed.

Sattler, J.M. (1992). *Assessment of children: WISC-III and WPPSI-R supplement*. San Diego: Jerome M. Sattler.

Smith, T.O.; Cogan, A.; Patel, S.; Shakokani, M.; Toms, A.P. & Donell, S.T. (2013). The intra- and inter-rater reliability of X-ray radiological measurements for patellar instability. *The Knee, 20*, 133-138.

Steenson, J.; Vivanti, A. & Isenring E. (2013). Inter-rater reliability of the subjective global assessment: A systematic literature review. *Nutrition, 29*, 350-352.

Storch, E.A.; Wood, J.J.; Ehrenreich-May, J.; Jones, A.M.; Park, J.M.; Lewin, A.B. & Murphy, T. K. (2012). Convergent and discriminant validity and reliability of the Pediatric Anxiety Rating Scale in youth with autism spectrum disorders. *Jounal Autism Devevelopment Disorder, 42*, 2374-2382.

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86* (2), 420-428.

Van Noord, R.G. & Prevatt, F.F. (2002). Rater agreement on IQ and achievement tests effect on evaluations of learning disabilities. *Journal of School Psychology, 40* (2), 167-176.

Vianna, H. (1976). *Testes em Educação*. São Paulo: Ibrasa / MEC

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children-Third Edition (WISC-III): Manual*. San Antonio: Psychological Corporation.

Wechsler, D. (2002). *WISC-III: Escala de Inteligência Wechsler para Crianças Terceira Edição: Manual.* (V. L. M. Figueiredo, adaptação e padronização brasileira). São Paulo: Casa do Psicólogo.

Recebido em 12/11/12
Revisto em 5/03/13
Aceito em 10/03/13