Subtestes Semelhanças, Vocabulário e Compreensão do WISC-III: Pontuação Objetiva ou Subjetiva?

WISC-III Subtests of Similarities, Vocabulary and Comprehension: Objective or Subjective Scoring?

Vera Lucia Marques de Figueiredo*, Jaciana Marlova Gonçalves Araújo, Taise Costa Dias, & Marcela Vargas Busetti
Universidade Católica de Pelotas, Pelotas, Brasil

Resumo
Este trabalho visou analisar a pontuação dos subtestes Semelhanças, Vocabulário e Compreensão do WISC-III, tendo em vista que podem envolver maior subjetividade do avaliador. Participaram do estudo 42 psicólogos de diferentes Estados do Brasil, os quais corrigiram as respostas de 6 protocolos do teste selecionados aleatoriamente da amostra de padronização ao contexto brasileiro. Tomando-se como referência os escores totais, o subteste Vocabulário apresentou maior variabilidade nas pontuações, seguido por Compreensão. Considerando-se o total de itens analisados em cada subteste, Semelhanças apresentou a maior concordância entre os avaliadores. Entretanto, os resultados evidenciaram que os três subtestes envolvem a subjetividade do avaliador na pontuação das respostas.
Palavras-chave: WISC-III; Psicometria; Subtestes Verbais.

Abstract
In all psychological tests, scoring should be of concern for examiners because the accuracy of results depends, at some extent, on the quality of the correction. This work aims to examine the correction, by different psychologists, of the scores for the Wechsler Intelligence Scale for Children (WISC-III) subtests of Similarities, Vocabulary and Comprehension since these are the subtests where examiner's subjectivity seemingly most influences scoring. Forty two psychologists from different states in Brazil participated in this study. They corrected the answers of six test protocols randomly selected from a standardization sample for the Brazilian context. Taking as reference the total scores, the Vocabulary subtest showed greater variability in score, followed by the Comprehension one. Considering the total number of items tested in each subtest, Similarities had the highest agreement among raters. The results showed that all the three subtests involve subjectivity on behalf of the examiner to score the answers. Continuing in this study, we also aim to determine test reliability based on interrater agreement.
Keywords: WISC-III; Psychometric; Verbal Subtests.

O teste psicológico é uma ferramenta capaz de auxiliar no diagnóstico e na tomada de decisão do psicólogo. Entretanto, para garantir sua eficácia, é condição essencial que este instrumento de aferição esteja calibrado e que os usuários tenham domínio dos procedimentos de uso e interpretação. Identificar as limitações da técnica e as dificuldades provenientes de seu emprego é uma tarefa importante que vai auxiliar seus usuários a minimizar os erros de estimação. A proposta deste trabalho é apontar os problemas envolvidos na etapa de correção (pontuação) de uma técnica amplamente empregada para avaliar a inteligência de crianças e adolescentes.

* Endereço para correspondência: Universidade Católica de Pelotas, Escola de Psicologia, Rua Almirante Barroso, 1202, Centro, Pelotas, RS, Brasil, CEP 96010-280. E-mail: verafig@terra.com.br, jacianamga@hotmail.com, taise.cdias@hotmail.com e marcelabusetti@hotmail.com

Segundo Pasquali (1999), a melhor maneira de observar um fenômeno psicológico é através da medida. Esse pressuposto é reforçado por Alchieri e Cruz (2003), para os quais os testes são operações empíricas utilizadas pela Psicologia a fim de estudar as diferentes manifestações do seu objeto, os processos psíquicos e os comportamentos deles decorrentes.

A medida é um procedimento empírico e como tal não está isento de erro. O erro está sempre presente em qualquer medida, podendo ser creditado à própria observação, à amostragem ou ao próprio evento que foi avaliado (Pasquali, 2003). Uma das formas de minimizar os erros no processo de correção de um teste é fundamentar as decisões nas regras do manual. Para Cunha (2000), em relação aos escores para pontuar um teste, geralmente, as instruções que constam nos manuais são suficientes. A familiarização com o teste e o uso correto do manual auxiliarão na atribuição dos diferentes escores e serão

úteis para que o examinador saiba quando os dados de que dispõe são suficientes ou quando deve estimular o examinando a falar mais. Essas recomendações, no sentido de melhor entender o raciocínio subjacente à resposta, não implicam em quebra de instruções na administração. Por outro lado, mesmo que o examinando seja o foco principal, a testagem e a atribuição de escores devem também exigir atenção por parte do examinador.

Quando ocorrem respostas inusitadas, bem diversas das constantes entre os exemplos do manual, a adequação ou a correção de tais respostas deve ser julgada pelo examinador. Ele deve, também, consultar profissionais especializados ou obras específicas para se certificar de que a atribuição de tal escore pode ser razoável (Cunha, 2000).

A aplicação e pontuação corretas do teste irão determinar uma interpretação precisa da habilidade do examinando. O conceito de fidedignidade refere-se a quanto o escore obtido no teste se aproxima do escore verdadeiro do sujeito num traço qualquer. A fidedignidade de um teste está intimamente ligada ao conceito de variância-erro, sendo este definido como a variabilidade nos escores produzida por fatores estranhos ao construto (Pasquali, 2001).

Há vários métodos para estimar a precisão de um teste e, entre eles, está a fidedignidade dos avaliadores que, segundo Anastasi e Urbina (2000), pode ser determinada com base em uma amostra de protocolos dos testes pontuados independentemente por dois ou mais examinadores. Segundo McIntire (2000) e (Pasquali (2001), entre os fatores que podem influenciar os resultados dos testes não-objetivos, encontra-se a opinião do apurador, como fator de decisão. Nesse caso, é preciso mais de um apurador para garantir a precisão do resultado no teste.

O *Wechsler Intelligence Scale for Children-Third Edition* (WISC-III) é um teste de inteligência utilizado em crianças entre 6 e 16 anos. Sendo um dos mais utilizados para avaliar a inteligência, ele estima a capacidade intelectual, estabelecendo um perfil das habilidades cognitivas, e é administrado por psicólogos, tanto da área educacional quanto clínica. O desenvolvimento cognitivo do examinando é avaliado por meio dos escores obtidos nas diferentes tarefas do teste, os quais são comparados com os esperados para sua idade. A exatidão na pontuação dos resultados é essencial para garantir um diagnóstico preciso. Contudo, nenhum estudo sobre a precisão dos avaliadores foi feito na adaptação do instrumento ao contexto brasileiro.

Segundo Wechsler (1991), a correção da maioria dos itens do WISC-III é objetiva e exige pequena ou nenhuma interpretação de critérios por parte do examinador. Já em alguns subtestes, como Compreensão, Semelhanças, Vocabulário e Informação, os itens exigem julgamento do examinador. Por essa razão, especialistas neste instrumento sugerem que, para evitar a subjetividade do avaliador na correção e assegurar o menor número possível de erros na apuração dos resultados, os protocolos devem ser corrigidos por mais de um profissional, além do próprio aplicador (Sattler, 1992).

No WISC-III (Wechsler, 1991), a pontuação de itens dos subtestes verbais deve ser baseada nas respostas da amostra de padronização. Essas respostas estão incluídas no manual do teste abaixo de cada item dos respectivos subtestes, ilustrando os vários tipos e níveis de respostas. Também há critérios gerais para orientar o examinador, principalmente no subteste Vocabulário. Em algumas situações, entretanto, tais informações parecem insuficientes para que o profissional possa pontuar com certeza uma determinada resposta, posto que, muitas vezes, o examinando emite respostas bastante diferentes das registradas no manual.

Segundo Sattler (1992), a pontuação dos subtestes Semelhanças, Vocabulário e Compreensão exige considerável habilidade. Além disso, para o autor, um estudo cuidadoso deve ser feito nos critérios de pontuação para ajudar a reduzir possíveis erros. Na pesquisa desenvolvida por Sattler, Andres, Squire, Wisely e Maloy (1978) com o WISC-R, 110 psicólogos e estudantes de graduação pontuaram 726 respostas ambíguas nos subtestes Semelhanças, Vocabulário e Compreensão. Um nível de 80% de concordância na pontuação foi alcançado em somente 51% das respostas de Semelhanças, 49% de Compreensão e 38% de Vocabulário, evidenciando dificuldade na pontuação das respostas.

Na versão americana do teste WISC-III, foi estabelecida a fidedignidade entre avaliadores para os subtestes Semelhanças, Vocabulário e Compreensão. Sessenta protocolos foram selecionados aleatoriamente da amostra de padronização e quatro avaliadores pontuaram, independentemente, os três subtestes. No estudo, foi usado um tipo de correlação Intra-Classe para avaliar a concordância entre juízes. 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. Os resultados mostram que tais subtestes, apesar de exigirem maior julgamento do examinador, podem ser pontuados muito confiavelmente (Wechsler, 1991).

Estudos mostram que os examinadores que utilizam as escalas Wechsler de inteligência cometem erros de administração e de pontuação com grande freqüência. As pesquisas apontam que grande parte do percentual de erros encontrados no WISC-III e WISC-IV são cometidos na fase de pontuação, com tendência a superestimar os escores (Belk, LoBello, Ray, & Zachar, 2002; Loe, Kadlubek, & Marks, 2007). Por outro lado, Sherman e Taylor (2001) identificaram, no subteste de Compreensão, o efeito da experiência dos avaliadores, uma vez que

psicólogos mais experientes foram mais rigorosos na pontuação do que alunos do início da graduação. Simões et al. (2000) analisaram a concordância entre avaliadores nos subtestes Semelhanças, Vocabulário, Compreensão e Labirintos em um estudo no qual quatro profissionais pontuaram independentemente 40 protocolos do teste. O subteste identificado como mais difícil de pontuar objetivamente foi Compreensão.

No Brasil, não há trabalhos divulgados relacionados à dificuldade de pontuação das respostas do WISC-III. No manual do teste (Wechsler, 2002) a autora da adaptação aponta a necessidade de desenvolver estudos relativos à precisão na pontuação. Por essa razão, o presente estudo teve como objetivo identificar o nível de concordância entre os avaliadores na tarefa de pontuar itens dos três subtestes que exigem maior julgamento dos avaliadores.

Método

Participantes

Em contato com especialistas reconhecidos na comunidade científica por sua prática na utilização do WISC-III, solicitou-se a indicação de outros profissionais que também utilizassem o teste. Dessa forma, formou-se uma rede com 60 psicólogos de diferentes Estados do Brasil que foram convidados a participar do estudo. Entretanto, apenas 42 demonstraram interesse, devolvendo o material enviado.

Material

Foram selecionados aleatoriamente do banco de dados da pesquisa de padronização do WISC-III ao contexto brasileiro (Wechsler, 2002) seis protocolos do teste com o registro das respostas colhidas em tal ocasião e o material foi encaminhado via correio, com envelope selado e subscrito para facilitar a devolução. Os protocolos examinados pelos psicólogos eram de crianças e adolescentes de escolas públicas com idades entre 9 e 15 anos. Juntamente com os protocolos, foram enviadas aos profissionais: (a) uma ficha de dados de identificação; (b) uma carta de livre consentimento informado; e (c) uma carta contendo orientações sobre o procedimento para execução da tarefa. Os psicólogos deveriam pontuar, seguindo os critérios do manual do teste, as respostas registradas nos protocolos. Em caso de dúvidas na pontuação de algum item, poderia colocar entre parênteses a pontuação alternativa.

Procedimentos para Análise dos Dados

A digitação e análise dos dados foram realizadas nos programas SPSS e Excel, utilizando-se medidas de tendência central, análises de freqüências e teste *Qui-quadrado*. Foram calculados, com base no total de respostas analisadas, os percentuais de avaliadores (de 40 a 100%) que deram a mesma pontuação em cada um dos itens.

Para este estudo, foram considerados o gabarito formulado pelas pesquisadoras (escores esperados), as pontuações em cada item e os escores totais atribuídos pelos avaliadores aos subtestes (escores observados). Para a análise das pontuações, as somas dos escores de cada subteste foram comparadas ao gabarito. Calcularam-se as diferenças entre o gabarito e os escores observados e, pos-teriormente, as freqüências dessas diferenças, o que possibilitou identificar o percentual de profissionais que encontraram os mesmos escores previstos (diferença igual a zero foi interpretada como concordância); para tais diferenças consideraram-se valores positivos, como tendência a superestimação e escores negativos, como subestimação.

Resultados

Os psicólogos que participaram como avaliadores eram 95,3% (*N*=40) do sexo feminino, com idade média de 39 anos (*DP*=10,84), sendo que 42,8% (*N*=18) residiam na região sul, 47,6% (*N*=20) na região sudeste, 4,7% (*N*=2) na região centro-oeste e 4,7% (*N*=2) na região nordeste. Entre eles, 47,3% (*N*=20) tinham mestrado, 26,2% (*N*=11), alguma especialização e 21,4% (*N*=9), doutorado, sendo que 4,8% da amostra (*N*=2) não especificaram sua formação. Os contextos mais freqüentes de utilização do teste pelos avaliadores eram clínica e pesquisa, em atividades como docência e 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).

Na revisão dos protocolos recebidos, identificaram-se alguns problemas como, por exemplo, itens e subtestes não pontuados e escores totais não somados ou somados incorretamente. Entre as razões que podem ter contribuído para estas falhas, pode-se pensar na falta da instrução, por parte das pesquisadoras, de que os escores individuais de cada subteste deveriam ser somados por suporem que essa tarefa seria automaticamente realizada pelos avaliadores. A falta de conclusão da tarefa, ou seja, deixar itens sem pontuação por parte de alguns avaliadores, pode estar associada à pouca disponibilidade de tempo, uma vez que o trabalho de pontuação consistiu em uma atividade exaustiva. Os avaliadores não registraram algumas somas em função de dúvidas na pontuação de determinados itens. Por esse motivo, os cálculos que utilizaram as somas dos escores nem sempre representaram a totalidade dos avaliadores (*N*=42).

Na pontuação dos itens, os psicólogos foram solicitados a indicar quando tivessem dúvidas no escore a ser atribuído a cada resposta. Ao analisar a freqüência de

tais dúvidas, identificaram-se no subteste Vocabulário os itens 7, 11, 15 e 21; em Semelhanças, os itens 2, 7, 8 e 11; e em Compreensão, os itens 2, 6, 7 e 12 como os que causaram mais dúvidas na pontuação, considerando os seis protocolos. Exemplificando as respostas que foram mais dúbias em cada subteste, temos em Vocabulário (item 21 – aflição) respostas como "inquieta e nervosa", suscitando dúvidas para a pontuação em 33% dos profissionais; em Semelhanças (item 11 – família-tribo) foram "um monte de pessoas, pai, irmãos", "conjunto de pessoas", e "união", enquanto que, em Compreensão (item 7 – briga), "chamava ou falava com a mãe" e "conversava com ela". Tais respostas foram pontuadas com escores entre 0 e 1 ponto. Entretanto, os profissionais registraram que tinham dúvidas de que a pontuação poderia variar de 0 a 2 pontos. Nos registros dos avaliadores, as dificuldades ocorreram tanto pelo fato de as respostas não constarem nos exemplos oferecidos pelo manual, como pela falta de inquéritos, que resultou em respostas incompletas e/ou duvidosas. Segundo Sattler (1992), o manual do WISC-III apresenta um limitado número de exemplos, causando dificuldade no estabelecimento de critérios precisos que se apliquem às diferentes respostas dadas pelos examinandos, o que possibilita que cada profissional utilize critérios próprios para decidir sobre a pontuação.

Ao pontuarem, os avaliadores não mostraram tendência na estimação, exceto em Vocabulário, onde, dos seis protocolos analisados, quatro apresentaram um percentual maior de avaliadores que superestimaram o escore total em relação ao gabarito. No item 4, (guarda-chuva) por exemplo, a resposta "Não deixa a gente se molhar" foi avaliada com dois pontos por 88% dos juízes, mostrando tendência à supervalorização, uma vez que, segundo o manual, a pontuação deveria ser um. O manual contém a orientação de que dois pontos devem ser atribuídos a respostas que dão idéia de proteção e de situação em que se usa o objeto, tal como "Mantém a pessoa seca quando chove". Os dados confirmam, em parte, os estudos de Belk et al. (2002) e Loe et al. (2007), que encontraram uma tendência dos avaliadores em superestimar os escores dos subtestes verbais.

Com base nos escores totais de cada subteste estimados pelos avaliadores, as medidas de tendência central encontradas para cada um foram calculadas e os dados aparecem nas Tabelas 1, 2 e 3.

Tabela 1
Escores Esperados e Observados no Subteste Semelhanças

Protocolos	*N*	Escores esperados	Concordância (%)	Escores Observados				
				Moda	Média	(*DP*)	Mín-Máx	(a)
1	38	13	46,3	13	12,93	1,21	10-15	5
2	39	11	40,4	11	10,38	1,13	8-13	5
3	39	23	26,3	23	22,52	2,07	17-26	9
4	39	8	40,5	8	7,86	1,41	4-11	7
5	39	9	35,6	9	9,71	1,17	8-13	5
6	38	22	31,7	23	22,46	1,27	19-26	7

Nota. *N* = número de respostas válidas constantes em cada protocolo; a = amplitude.

Tabela 2
Escores Esperados e Observados no Subteste Vocabulário

Protocolos	*N*	Escores esperados	Concordância (%)	Escores Observados				
				Moda	Média	(*DP*)	Mín-Máx	(a)
1	37	18	12,3	20	19,37	2,84	10-24	14
2	40	22	33,3	22	22,0	1,79	15-24	9
3	39	34	11,9	32	32,45	1,97	27-36	9
4	38	26	22,0	26	25,0	2,82	19-31	11
5	39	24	28,5	24	24,21	2,31	16-30	14
6	38	34	11,8	35	34,79	2,89	29-43	14

Nota. *N* = número de respostas válidas constantes em cada protocolo; a = amplitude.

Tabela 3
Escores Esperados e Observados no Subteste Compreensão

Protocolos	*N*	Escores esperados	Concordância (%)	Escores Observados				
				Moda	Média	(*DP*)	Mín-Máx	(a)
1	37	16	20,0	16	15,85	2,58	9-21	12
2	37	10	39,0	10	10,0	1,10	8-12	4
3	39	20	31,0	20	21,21	2,49	18-28	10
4	37	15	17,5	14	15,10	2,74	10-23	13
5	37	14	50,0	14	13,3	1,45	10-16	6
6	37	25	25,0	25	26,68	2,77	21-32	11

Nota. N = número de respostas válidas constantes em cada protocolo; a = amplitude.

Nas Tabelas 1, 2 e 3 observa-se que, quanto à concordância dos avaliadores, em relação aos escores totais, Vocabulário evidenciou o menor e Semelhanças o maior percentual de concordância em relação ao gabarito. Estes resultados não coincidem com os de Wechsler (1991), que apontou Compreensão como o de menor índice de concordância entre juízes, ao estabelecer o coeficiente de fidedignidade (0,90). Para Sattler et al. (1978), Vocabulário foi identificado como o subteste que envolve maior subjetividade na correção. Deve-se levar em conta que a pontuação dada pelas pesquisadoras não teve a pretensão de ser considerada como a correta, mas como uma referência que, ao mostrar-se discrepante dos escores dos avaliadores, permitiu identificar a influência do julgamento na correção.

As modas, valores mais freqüentes de uma distribuição, foram próximas aos valores esperados, enquanto as médias mostraram maior variabilidade devido à influência dos valores extremos, observados na amplitude das pontuações. Entretanto, a proximidade das médias e das modas sugere uma distribuição normal. Pelo cálculo do *Qui-quadrado*, compararam-se os escores esperados com os da média observada; os valores foram menores que o tabelado χ^2 (5, N = 42) = 16,75, p< 0,05, demonstrando que as diferenças não foram estatisticamente significativas, ou seja, que os escores observados foram bastante próximos aos esperados.

Apesar de estatisticamente não se observarem diferenças significativas, sob o ponto de vista qualitativo, o desejado seria que pesquisadoras e avaliadores encontrassem a mesma pontuação, uma vez que esta determinará os coeficientes de inteligência (QIs) dos examinandos, pois escores diferentes resultariam em QIs distintos para um mesmo sujeito. Entretanto, os escores mínimos e máximos e, conseqüentemente, a amplitude, demonstram variabilidade no julgamento, resultando em valores considerados *outliers*. Ao calcular os valores extremos, tomando-se dois desvios-padrão em relação à média, identificaram-se em Vocabulário, Semelhanças e Compreensão, respectivamente 10, 9 e 6 pontuações fora da faixa média.

Ainda nas Tabelas 1, 2 e 3, em relação à variabilidade dos escores, observa-se que Vocabulário (média=11,8; *DP*=2,48) e Compreensão (média=9,3; *DP* =3,55) foram os subtestes que apresentaram maior variabilidade e Semelhanças (média=6,3; *DP* =1,63), apesar da menor amplitude, também evidenciou dispersão. Talvez, por Vocabulário e Compreensão propiciarem maior ocorrência de respostas diferentes das apresentadas no manual, aumente a possibilidade de a pontuação ser influenciada pela subjetividade do avaliador.

Para cada subteste foi realizada a análise da freqüência das pontuações de cada item, independentemente de o escore ser 2, 1 ou 0. Considerando o total de itens pontuados, calculou-se a percentagem dos que apresentaram concordância entre avaliadores igual ou superior a 40%. A percentagem de concordância foi calculada a partir das pontuações dadas pelos avaliadores a todas as respostas de um mesmo subteste. Assim, como havia diferentes protocolos do teste a serem pontuados, a resposta dada a cada item foi considerada independentemente; por exemplo, o item 1 de qualquer subteste, considerando-se os 6 protocolos, foi analisado como tendo 6 respostas distintas.

Conforme a Tabela 4, dentro de um total de 79 respostas analisadas no subteste Compreensão, apenas em 20%, todos os juízes (100%) deram a mesma pontuação. Tomando-se como referência o critério de 80%, valor que pode ser considerado como leniente, em Semelhanças, a concordância na pontuação ocorreu em 87% dos 84 itens analisados, enquanto, a menor concordância pôde ser observada no subteste Compreensão, em que os juízes concordaram em apenas 65% das pontuações. Os dados diferem dos de Sattler et al. (1978), que encontraram 80% de concordância em 51% dos itens de Semelhanças, 49% de Compreensão e 38% de Vocabulário no WISC-R. O percentual de itens em que houve concordância na pesquisa de Sattler et al. (1978) foi menor, provavelmente, pelo fato de terem sido analisadas somente respostas ambíguas, enquanto, no presente estudo, todas as respostas registradas nos protocolos foram consideradas.

Tabela 4
Percentagem Acumulativa de Itens Com a Mesma Pontuação e os Diferentes Critérios de Concordância

Critérios de concordância entre avaliadores	Percentagem acumulativa de itens nos Subtestes		
	Compreensão (*N*=79)	Semelhanças (*N*=84)	Vocabulário (*N*=110)
100%	20	30	20
90%	51	71	56
80%	65	87	74
70%	73	93	85
60%	87	95	88
50%	100	99	93
40%	100	100	100

Observa-se que o subteste Semelhanças foi o que apresentou maior concordância entre avaliadores, em praticamente todos os níveis de critérios considerados e os menores índices foram observados, de maneira geral, em Compreensão. Os resultados coincidiram com os de Simões et al. (2000), para os quais o subteste mais difícil de pontuar objetivamente foi Compreensão.

Os resultados demonstram a influência da subjetividade na pontuação dos subtestes Semelhanças, Vocabulário e Compreensão do WISC-III, corroborando a afirmação de Pasquali (2001) de que, em testes não-objetivos, quando a opinião do apurador entra como fator de decisão, esta pode influenciar os resultados. No caso do WISC-III, as conseqüências podem ser funestas, uma vez que o desenvolvimento cognitivo do examinando (expresso em QI) é determinado por meio dos escores obtidos no teste segundo a atribuição do avaliador.

Conclusão

Os subtestes do WISC-III são considerados por Wechsler (1991) como provas objetivas. Entretanto, os três subtestes analisados apresentaram divergências nas pontuações, mesmo com o estabelecimento de um critério de 80% de concordância entre avaliadores. Os subtestes Vocabulário e Compreensão apresentaram maior possibilidade de discordância na pontuação, enquanto Semelhanças, a menor, sendo este o que eliciou respostas menos complexas dos examinandos e, conseqüentemente, respostas mais fáceis de pontuar.

A relevância da etapa de correção dos dados de um teste psicológico está associada à interpretação dos resultados, uma vez que escores divergentes produzirão, no caso das escalas de inteligência, QIs falseados. Os resultados da pesquisa reforçam que a correção dos subtestes Vocabulário, Compreensão e Semelhanças do WISC-III exige familiaridade com os critérios de pontuação, uma vez que envolvem subjetividade. Portanto, é importante que os psicólogos consultem os pares sempre que ocorrerem dúvidas, para que a análise seja revisada na tentativa de minimizar possíveis erros.

Com base nos resultados, conclui-se ser necessária a aplicação rigorosa dos testes, com os inquéritos pertinentes, e, ainda, a elaboração de um material complementar que disponibilize critérios para analisar outras verbalizações freqüentes no contexto brasileiro, auxiliando os psicólogos na correção dos subtestes mais subjetivos. Sua finalidade seria a ampliação das respostas-modelo que consistem atualmente na tradução das respostas da amostra de padronização americana, as quais exemplificam as pontuações de cada item, apresentadas no manual brasileiro.

O estudo teve algumas limitações: (a) Percentual de recusas, que impossibilitou uma maior representatividade dos profissionais que trabalham com o WISC-III; (b) Protocolos entregues incompletos ou não revisados pelos avaliadores, provavelmente pela tarefa ter exigido muito tempo para sua realização, o que dificultou algumas análises. Por outro lado, como ponto positivo, destaca-se a amplitude do estudo, ou seja, o aumento do número de avaliadores, da abrangência geográfica, e da quantidade de protocolos analisados, quando comparado aos procedimentos utilizados nos trabalhos consultados (Belk et al., 2002; Loe et al., 2007; Sherman & Taylor, 2001; Simões et al., 2000; Wechsler, 1991).

Priorizaram-se como foco do presente estudo, análises qualitativas concernentes aos subtestes que causam mais dúvidas na pontuação, tentando apresentar aos profissionais as dificuldades e os erros mais freqüentes. Dando continuidade a este trabalho, pretende-se, em um próximo estudo, apresentar as análises psicométricas relativas aos coeficientes de fidedignidade com base nos avaliadores (coeficiente de correlação Intra-Classe), para complementar os dados sobre a precisão do instrumento.

Referências

Alchieri, J. C., & Cruz, R. M. (2003). *Avaliação psicológica: Conceitos, métodos e instrumentos* (Vol. 1). São Paulo, SP: Casa do Psicólogo.

Anastasi, A., & Urbina, S. (2000). Fidedignidade. In A. Anastasi & S. Urbina, *Testagem psicológica* (7. ed., pp. 84-

105). Porto Alegre, RS: Artes Médicas.

Belk, M. S., LoBello, S. G., Ray, G. E., & Zachar, P. (2002). WISC-III administration, clerical, and scoring errors made by student examiners. *Journal of Psychoeducational Assesment, 20*, 290-300.

Cunha, J. A. (2000). Escalas Wechsler. In J. A. Cunha (Ed.), *Psicodiagnóstico V.* (5. ed., Vol 1, pp. 529-602). Porto Alegre, RS: Artes Médicas.

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 Assesment, 25*(3), 237-247.

McIntire, S., & Miller, L. (2000). Foundations of Psychological Testing. Boston: McGraw-Hill.

Pasquali, L. (1999). Testes referentes a construto: Teoria e modelo de construção. In L. Pasquali, *Instrumentos psicológicos: Manual prático de elaboração* (pp. 21-52). Brasília, DF: LabPAM.

Pasquali, L. (2001). *Técnicas de exame psicológico - TEP. Manual*. São Paulo, SP: Casa do Psicólogo.

Pasquali, L. (2003). Teoria da medida. In L. Pasquali, *Psicometria: Teoria dos testes na Psicologia e na Educação* (Vol. 1, pp. 23-51). Petrópolis, RJ: Vozes.

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

Sattler, J. M., Andres, J. R., Squire, L. S., Wisely, R., & Maloy, C. F. (1978). *Examiner scoring of ambiguous WISC-R responses. Psychology in the schools* (4nd ed., Vol. 15, pp. 486-489). San Diego, CA: Verlag Chemie.

Sherman, L., & Taylor, A. (2001). *Experimentally manipulated bias in school psychologists' scoring of WISC-III protocols.* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago, IL. Retrieved November 22, 2008, from http://www.users.muohio.edu/shermalw/mwera_version5_files/mwera_version5.htm

Simões, M. R., Santos, M. J. S., Pereira, M., Albuquerque, C. P., Vilar, M., Lança, C., et al. (2000, ago.). Estudos de precisão com a WISC-III: Acordo intercotadores e teste-reteste [Abstract]. In *Anais da VII Conferência Internacional "Avaliação Psicológica: Formas e Contextos"* (pp. 25). Belo Horizonte, MG.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children - WISC-III: Manual* (3th ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (2002). *Escalas de Inteligência Wechsler para Crianças: Manual* (3. ed., V. L. Marques de Figueiredo, Adaptação e padronização brasileira). São Paulo, SP: Casa do Psicólogo.

Recebido: 20/02/2009
1ª revisão: 03/08/2009
Aceite final: 02/09/2009