

Confiabilidade do Teste WPPSI-R na avaliação do desenvolvimento cognitivo de pré-escolares

Reliability of the WPPSI-R test in the evaluation of cognitive development in preschoolers

Maria Dalva Barbosa Baker Méio ¹
 Claudia S. Lopes ²
 Rosely Sichieri ²
 Denise Streit Morsch ¹

¹ Departamento de Neonatologia, Instituto Fernandes Figueira, Fundação Oswaldo Cruz, Av. Rui Barbosa 716, Rio de Janeiro, RJ 22250-020, Brasil. meio@openlink.com.br

² Instituto de Medicina Social, Departamento de Epidemiologia, Universidade Estadual do Rio de Janeiro, Rua São Francisco Xavier 524, 7º andar, Rio de Janeiro, RJ 20559-900, Brasil. lopes@uerj.br dsmorsch@gbf.com.br

Abstract *The WPPSI-R scale (Wechsler Preschool and Primary Scale of Intelligence – Revised) is a psychometric test chosen as the evaluation tool in a study on preschool-age cognitive development in a cohort of very low birth weight (VLBW) premature children from the Fernandes Figueira Institute (IFF), applied by four previously trained psychologists. The objective of this study was to verify inter-observer reliability in the test application. Two types of reliability study design were used: balanced incomplete blocks, to verify agreement in the application of the scale, and crossed design, to verify agreement in scoring of items. We studied 12 preschool children born at IFF (birthweight < 1,500g). The intraclass correlation coefficients (ICC) were: 0.82 (full-scale IQ), 0.89 (verbal IQ), and 0.91 (performance IQ), in the incomplete block design study, and 0.99, 0.98, and 0.99, respectively, in the crossed design study, indicating good reliability. Application of the WPPSI-R scale in the study of cognitive development of VLBW premature children at IFF proved adequate, as shown by these results.*

Key words *Reproducibility of Results; WPPSI-R Test; Cognition; Child Development*

Resumo *O Teste psicométrico WPPSI-R (Wechsler Preschool and Primary Scale of Intelligence – Revised) foi o instrumento de avaliação em uma pesquisa para estudar o desenvolvimento cognitivo, na idade pré-escolar, de uma população de crianças do Instituto Fernandes Figueira (IFF) nascidas prematuras e de muito baixo peso (MBP), aplicado por quatro psicólogas previamente treinadas. O objetivo deste estudo foi avaliar a confiabilidade inter-observador na aplicação do teste. Utilizaram-se dois desenhos: de blocos incompletos equilibrados, para avaliar a concordância na aplicação do teste, e cruzado, para avaliar a concordância na pontuação dos itens. Foram estudadas 12 crianças pré-escolares nascidas no IFF (Peso ao nascer < 1.500g). Os coeficientes de correlação intraclasses (CCI) encontrados foram: 0,82 (Escore Total), 0,89 (Escore Verbal) e 0,91 (Escore Executivo), no desenho de blocos incompletos equilibrados, e 0,99, 0,98 e 0,99, respectivamente, no desenho cruzado, indicando uma boa confiabilidade. Estes achados mostram que a aplicação do Teste WPPSI-R no estudo do desenvolvimento cognitivo de prematuros de MBP do IFF foi adequada.*

Palavras-chave *Reprodutibilidade de Resultados; Teste WPPSI-R; Cognição; Desenvolvimento Infantil*

Introdução

As mudanças nos cuidados perinatais – com a formação de ambulatórios para gestantes de alto risco e unidades de cuidados intensivos neonatais – possibilitaram o aumento da sobrevivência de recém-nascidos prematuros em todas as faixas de peso de nascimento, como relatado por Hack & Fanaroff (1988), Allen et al. (1993), Fanaroff et al. (1995) e Hack et al. (1996). A evolução futura destas crianças, egressas das unidades de terapia intensiva neonatal, tornou-se um problema de saúde pública e motivo de diversas pesquisas publicadas nos periódicos. Todavia, não apenas as anormalidades maiores – como surdez, cegueira e paralisia cerebral – são objetivos de pesquisas. Estudos recentes mostram a preocupação com o desenvolvimento intelectual deste grupo de crianças, já que a deficiência no desenvolvimento cognitivo pode interferir em seu rendimento escolar, de acordo com Resnick et al. (1998), McCarton et al. (1997), Kok et al. (1998) e Hutton et al. (1997).

A avaliação do desenvolvimento cognitivo é realizada através da aplicação de testes psicométricos, sendo que os testes mais comumente utilizados são as Escalas de Wechsler, o Teste de Stanford-Binet, o Teste de Griffiths e as Escalas de McCarthy. O Teste WPPSI-R (*Wechsler Preschool and Primary Scale of Intelligence – Revised*) faz parte das Escalas de Inteligência de Wechsler para Crianças (WISC) – Wechsler (1989). As Escalas de Wechsler compreendem, na realidade, duas escalas: o WISC, que abarca a faixa etária de 6 a 17 anos, e o WPPSI (Escalas Pré-escolares de Inteligência), que abrange a faixa etária de 4 a 7 anos. Esse teste foi revisado em 1989, permitindo a avaliação de crianças a partir da idade de 3 anos e 6 meses até o limite de sete anos de idade – WPPSI-R de acordo com o Manual WPPSI-R (Wechsler, 1989).

O Teste WPPSI-R, além de apresentar boa correlação com outros testes de avaliação do desenvolvimento cognitivo, é instrumento muito empregado em pesquisas para a avaliação do desenvolvimento cognitivo de prematuros, como nas pesquisas realizadas por Largo et al. (1989), Fedrizzi et al. (1993), Sommerfelt et al. (1996) e Bennet & Scott (1997). Foi, portanto, o instrumento escolhido para a avaliação do desenvolvimento cognitivo na idade pré-escolar de um grupo de crianças nascidas prematuras e de muito baixo peso no Instituto Fernandes Figueira (IFF).

O Teste WPPSI-R é composto por uma bateria de tarefas a serem executadas pela criança, agrupadas em dois grandes blocos – executivo

e verbal –, que possibilitam a avaliação de diferentes áreas do desenvolvimento cognitivo, resumidas, no final, por um escore global. O teste se baseia na apresentação da mesma tarefa com dificuldades crescentes, até alcançar o limite da criança. Nesta pesquisa foi utilizado um teste psicométrico, aplicado por um grupo de psicólogas, existindo grande preocupação com a precisão dos escores. Havia várias fontes potenciais de erros e discordância entre os examinadores, que abrangiam desde a própria complexidade do teste – a abordagem da criança, a oferta das tarefas, o registro do modo de execução das mesmas e a pontuação feita a partir desses registros – até o treinamento das psicólogas. Portanto, era necessário a realização de um estudo de confiabilidade inter-observador.

O objetivo deste estudo foi o de avaliar o grau de concordância das psicólogas previamente treinadas em relação à interpretação das respostas das crianças e à pontuação dos escores, fundamental para o estabelecimento do diagnóstico do desenvolvimento cognitivo nas diversas áreas.

Material e métodos

Em função do objetivo da pesquisa – ou seja, avaliar o desenvolvimento cognitivo de uma população de recém-nascidos prematuros de muito baixo peso ao nascer –, a população escolhida para a realização do estudo de confiabilidade foi de crianças pré-escolares nascidas prematuras, com peso de nascimento inferior a 1.500g, participantes da pesquisa realizada por Méio (1999). Os prematuros de muito baixo peso selecionados para a pesquisa original foram admitidos na Unidade de Terapia Intensiva Neonatal do Departamento de Neonatologia do IFF entre janeiro de 1991 e setembro de 1993, nascidos na maternidade do hospital ou transferidos até sete dias de vida para a Unidade e que sobreviveram à internação. Foram excluídas do estudo todas as crianças com malformações congênitas, síndromes genéticas e infecção congênita detectadas no período neonatal, transferidas de outras instituições com mais de sete dias de vida ou que tivessem nascido de parto domiciliar, assim como todas as crianças sem condições de serem testadas na idade pré-escolar em função de distúrbio emocional importante, autismo, surdez, retardo mental grave, tetraplegia espástica e cegueira.

O Teste WPPSI-R foi aplicado por uma equipe constituída de quatro psicólogas previamente treinadas, que não faziam parte da equipe

do ambulatório de acompanhamento e não tiveram acesso ao prontuário das crianças, de forma a não serem influenciadas pelo conhecimento das intercorrências neonatais ou evolução durante o acompanhamento ambulatorial.

Havia duas perguntas a serem respondidas. A primeira, se as psicólogas, ao aplicarem o teste, estariam concordando na interpretação das respostas dadas pelas crianças. A segunda, se estariam fazendo a pontuação adequada, a partir das respostas registradas que seguem um padrão estabelecido pelo teste. Em função de tais perguntas, este estudo de confiabilidade foi dividido em dois, de modo a permitir a resposta a estas questões.

Para responder a primeira pergunta – se as psicólogas estariam concordando na interpretação das respostas das crianças – foi escolhido o desenho de blocos incompletos equilibrados, proposto por Yates (1936) e descrito por Fleiss (1981) e Dunn (1989). Neste desenho, um dos examinadores aplica a entrevista ou teste, enquanto o outro assiste ao exame como espectador neutro. Ambos fazem uma avaliação independente da entrevista ou teste. Não seria possível expor as crianças à presença de mais de dois examinadores por sessão, e a equipe era composta por quatro psicólogas; nesta situação, este tipo de desenho permite fazer a avaliação da confiabilidade entre observadores mediante a utilização de uma estratégia de duplas de examinadores alternados. Como havia interesse em detectar a existência de vieses entre as psicólogas no modo de abordagem das crianças e no registro das respostas, este seria o desenho mais adequado.

Foram formadas duplas com as quatro psicólogas, nas quais uma aplicava o teste e registrava as respostas e a outra funcionava como espectadora neutra, apenas registrando as respostas; em outro momento, com outra criança, os papéis da dupla eram invertidos. Deste modo, ao fim do estudo, todas teriam aplicado e registrado ou apenas registrado o mesmo número de crianças. As psicólogas foram rotuladas como A, B, C e D, formando as duplas AB, CD, AC, AD, BC, BD. As crianças participavam do estudo na medida em que eram chamadas para a avaliação e em que havia a possibilidade da presença da dupla de psicólogas. Foi determinado um total de 12 crianças a serem testadas em duplas alternadas. Segundo Dunn (1989), algumas condições devem ser respeitadas para que este modelo de desenho de confiabilidade seja válido:

$$mr = nk \quad m \leq n \quad \lambda(m - 1) = r(k - 1)$$

Considerando:

m = número de examinadores envolvidos

n = número de pacientes

k = número de examinadores examinando cada indivíduo

r = número de indivíduos examinados por cada examinador

λ = número de indivíduos examinados por um determinado par de examinadores

Neste estudo tivemos 4 examinadores (m), 12 pacientes (n), 2 examinadores examinando cada indivíduo (k), 6 indivíduos examinados por cada examinador (r) e 2 indivíduos examinados por um determinado par de examinador. Por exemplo, os pacientes 1 e 7 foram examinados pela dupla A e B, os pacientes 3 e 9 pela dupla A e C, e assim sucessivamente. Desta forma, as condições necessárias foram preenchidas:

$$mr = nr \quad 4 \times 6 = 12 \times 2 \quad m \leq n \quad 4 < 12$$

$$\lambda(m - 1) = r(\lambda - 1) \quad 2(4 - 1) = 6(2 - 1)$$

A distribuição das duplas está exposta na Tabela 1.

Para responder a segunda pergunta – se as psicólogas estariam concordando na pontuação a partir das respostas registradas – foi escolhido o modelo de desenho cruzado, de acordo com Dunn (1989). Neste modelo, todas as psicólogas pontuaram, de forma independente, a totalidade dos registros dos testes aplicados. Este desenho permite avaliar a concordância da pontuação diagnóstica tanto dos escores obtidos nas subáreas quanto do escore global. Uma quinta psicóloga – que não participou da aplicação do teste – transcreveu todas as respostas obtidas nas 12 crianças para folhas identificadas apenas por códigos; as identificações dos desenhos feitos pelas crianças rece-

Tabela 1

Distribuição das duplas de psicólogas entre os 12 pacientes do estudo de confiabilidade (desenho de blocos incompletos equilibrados).

Paciente	Aplica/Registra	Registra
1	A	B
2	C	D
3	A	C
4	A	D
5	B	C
6	B	D
7	B	A
8	D	C
9	C	A
10	D	A
11	C	B
12	D	B

beram tarjas. A partir dos mesmos registros conseguidos nos testes realizados com as 12 crianças, não identificados, as psicólogas pontuaram não apenas os testes realizados ou assistidos por elas, mas também os realizados pelas outras duplas. No final foram alcançados 96 escores: 12 crianças x 2 registros independentes x 4 pontuações independentes, para cada área – global, executiva e verbal –, sendo feita a análise do estudo de confiabilidade.

Estratégia de análise

Armitage & Berry (1994) definem o coeficiente de correlação intraclasse como a correlação entre duas medidas feitas no mesmo indivíduo, efetuadas por examinadores ou instrumentos escolhidos aleatoriamente. Utilizado para avaliar a concordância entre examinadores ou instrumentos quando os resultados são variáveis contínuas, o coeficiente de correlação intraclasse (CCI) é calculado através da análise de variância, considerando-se excelente a concordância entre valores superiores a 0,75. Este coeficiente permite também analisar situações em que diferentes examinadores avaliam diferentes indivíduos e nas quais há mais de dois examinadores, levando em consideração a variação entre os indivíduos, a variação entre os instrumentos ou observadores e a variação decorrente do acaso.

A concordância entre os observadores na aplicação do Teste WPPSI-R foi avaliada em duas etapas. Na primeira, em que foi empregado o desenho de blocos incompletos equilibrados, foram calculados os CCIs para os três escores, total, verbal e executivo, de acordo com a análise de variância para este desenho (Fleiss, 1981); os intervalos de confiança foram calculados de acordo com Bartko (1994).

Na análise de variância proposta por Fleiss, é utilizada a estatística a_i , que funciona como um fator de ponderação, calculada através da fórmula:

$$a_i = 1/E (X_i - M_i),$$

onde:

E = fator de eficiência do desenho, que reflete a precisão do estudo;

M_i = média das médias dos escores (X_{ij}) de todos os indivíduos (j) avaliados por um determinado examinador (i);

X_i = média dos r resultados do examinador i.

O fator de eficiência do estudo é calculado através da fórmula:

$$E = [r (k - 1) + \lambda] / rk,$$

onde:

k = número de examinadores examinando cada indivíduo;

r = número de indivíduos examinados por cada examinador;

λ = número de indivíduos examinados por um determinado par de examinadores.

Na segunda etapa, quando as psicólogas pontuaram, de forma independente, todos os registros não identificados dos testes aplicados nas 12 crianças, o coeficiente de correlação intraclasse para cada um dos escores foi calculado através da análise de variância (ANOVA), utilizando-se o programa *SPSS for Windows*, versão 7.0 (SPSS Incorporation, 1995).

Resultados

A média dos 24 escores das 12 crianças foi de 72,0 (desvio padrão – DP = 13,01) para escore total, de 72,9 (DP = 13,06) para escore executivo e de 76,0 (DP = 12,67) para escore verbal. A amostra de crianças do estudo de confiabilidade foi, inclusive, um pouco mais homogênea do que a população geral da pesquisa (n = 79) cognitivo dos recém-nascidos prematuros de muito baixo peso ao nascer. Estes valores foram 75,6 (DP = 11,9) para o escore total, 77,0 (DP = 12,9) para o escore executivo e 78,6 (DP = 11,1) para o escore verbal. A Figura 4 mostra a distribuição dos escores das crianças que participaram do estudo de confiabilidade.

O valor do fator de eficiência (E) do estudo de confiabilidade realizado com o desenho de blocos incompletos equilibrados foi 0,67. Na análise de variância para cada um dos três escores (total, verbal e executivo), os CCIs encontrados foram 0,82 (IC 95% = 0,56 - 1,00) para o escore total, 0,89 (IC 95% = 0,71 - 1,00) para o escore verbal e 0,91 (IC 95% = 0,24 - 1,00) para o escore executivo. Os resultados dos coeficientes de correlação intraclasse obtidos no segundo estudo foram 0,99 (IC 95% = 0,94 -) para o escore total, 0,98 (IC 95% = 0,96 -) para o escore verbal e 0,99 (IC 95% = 0,87 -) para o escore executivo. Estes resultados estão na Tabela 2.

Discussão

Os estudos de confiabilidade são realizados com vários propósitos, desde a elaboração de instrumentos até a verificação da precisão de resultados obtidos com determinado instrumento. Entretanto, todos têm em comum a preocupação de verificar potenciais fontes de erros e de discordância, que resultariam em informações incorretas ou pouco precisas.

Dunn (1989) enfatiza que o estudo de confiabilidade de um instrumento, para ser válido,

Tabela 2

Coefficientes de correlação intraclasse obtidos nos dois estudos de confiabilidade realizados.

Escores	CCI	IC 95%*
1º estudo		
Escore total	0,82	0,56 -
Escore verbal	0,89	0,71 -
Escore executivo	0,91	0,24 -
2º estudo		
Escore total	0,99	0,94 -
Escore verbal	0,98	0,96 -
Escore executivo	0,99	0,87 -

* Cálculo do intervalo de confiança do CCI, segundo Bartko (1994).
 $F_0 = MSB / MSE$ com $(N-1)$, $(N-1)$ gl;
 $F_0 = 2 S_{xx} / 0,5 S_{yy}$; $F_L = F_0 / [F_{1-\alpha/2}(n-1), (n-1)]$;
 $F_U = F_0 [F_{1-\alpha/2}(n-1), (n-1)]$;
 $(F_L - 1) / [F_L + 1] < p (ICC) < (F_U - 1) / [F_U + 1]$

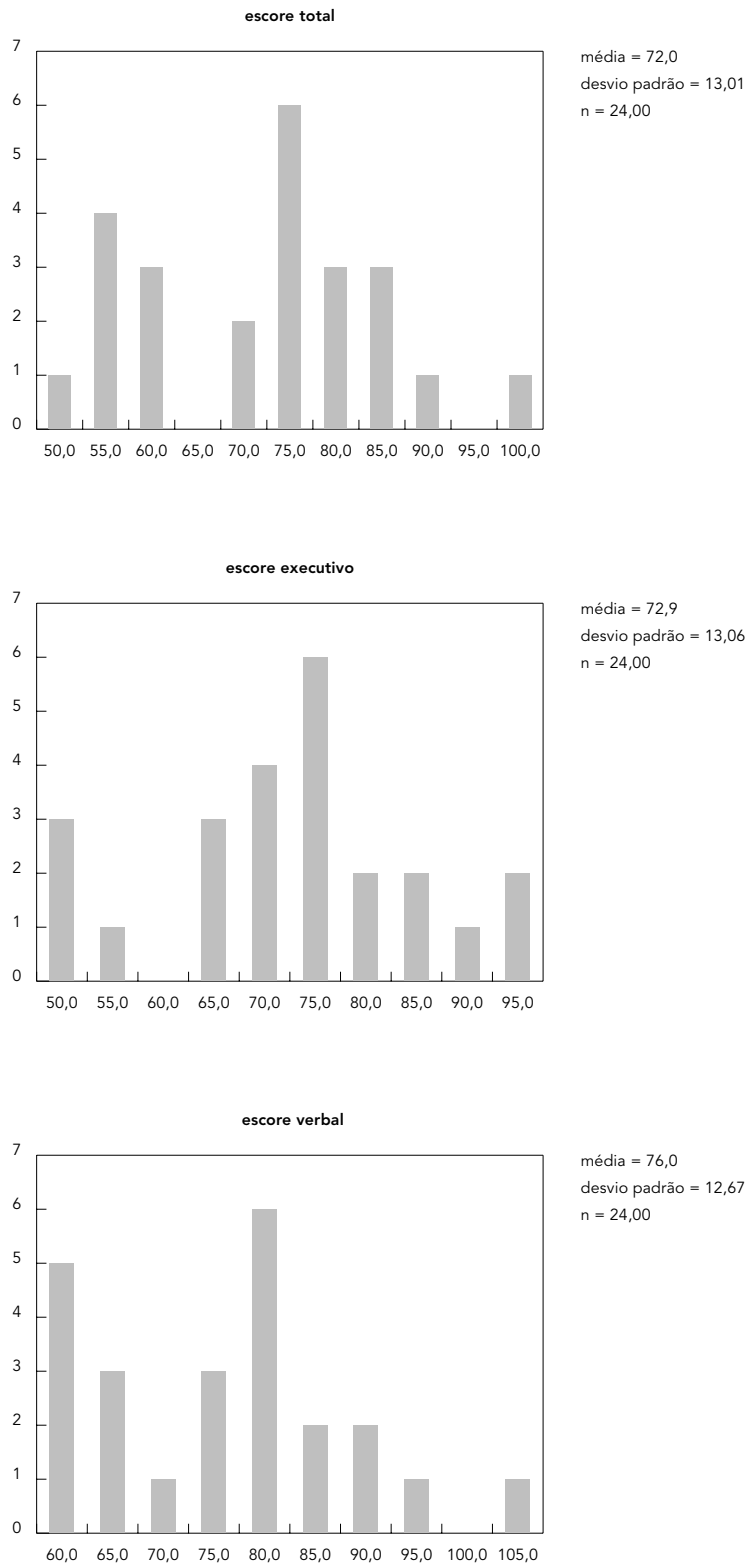
deve ser realizado na mesma população de indivíduos em que será empregado rotineiramente. O valor de uma técnica diagnóstica ou de um instrumento de medida, principalmente sua confiabilidade, não é apenas dependente de uma característica própria do instrumento ou técnica, mas é também resultante do meio em que é usado, do modo como é usado e da população que será avaliada por ele. Neste estudo de confiabilidade, as crianças faziam parte do grupo que completou a pesquisa original.

A utilização do desenho de blocos incompletos equilibrados permitiu contornar uma das grandes dificuldades para avaliar a confiabilidade na aplicação de um teste psicométrico por uma equipe de psicólogas. Por razões técnicas, não é possível colocar na mesma sala mais de duas psicólogas para aplicar um teste na mesma criança, já que isto poderia influenciar as respostas das crianças ou mesmo inibi-las. A vantagem deste tipo de desenho é permitir a aplicação dos teste por grupos de apenas duas psicólogas por vez, não influenciando os resultados dos escores e permitindo a avaliação da concordância entre observadores. Além disto, permite que o estudo seja realizado com pequeno número de pacientes, o que é importante quando se tem, como nas pesquisas com prematuros, a participação de número reduzido de pacientes.

Neste estudo, encontramos excelente concordância entre as psicólogas tanto na aplicação do teste quanto na pontuação dos registros obtidos, evidenciado pelos valores dos CCIs. O re-

Figura 1

Distribuição dos valores dos escores total, executivo e verbal.



sultado da análise do primeiro estudo demonstrou que, apesar das várias fontes potenciais de discordância na aplicação de um teste psicométrico, os escores obtidos, tanto para o escore total quanto para as áreas específicas, foram bastante concordantes, refletindo tanto uniformidade na abordagem das crianças como na interpretação das respostas dadas pelas crianças.

O cálculo do fator de eficiência para o desenho de blocos incompletos equilibrados leva em consideração o número de examinadores que examina cada indivíduo (k), o número de indivíduos examinados por cada examinador (r) e o número de indivíduos examinados por determinado par de examinadores (λ). A redução proporcional máxima na eficiência do estudo, ou seja, a precisão, é dada pelo valor $1 - E$. O fator de eficiência seria igual a um se houvesse um desenho de estudo completamente balanceado. Na estratégia utilizada, este será sempre inferior à unidade, e o limite inferior proposto por Fleiss (1981) como aceitável é de 0,67. Assim, o fator de eficiência encontrado neste estudo (0,67) significa que a perda de eficiência de um estudo com quatro examinadores e 12 pacientes é 33%, o que está no limite inferior de tolerância.

Os resultados do segundo estudo – realizado para verificar a concordância das psicólogas na pontuação dos itens do teste – mostram CCIs próximos a 1. Estes valores foram concordantes com o do estudo de confiabilidade do teste feito com crianças de um país desenvolvido.

O Teste WPPSI-R é instrumento bem elaborado e estruturado; a maioria dos seus subtestes envolve pontuação bem objetiva e direta. Entretanto, alguns subtestes implicam uma pontuação subjetiva, como: Compreensão, Vocabulário, Semelhanças, Quebra-cabeça e Desenho Geométrico. A confiabilidade interobservador para estes subtestes foi avaliada, pelos autores do Teste WPPSI-R, em duas amostras de crianças: uma de 151 crianças, sendo 83 meninos e 68 meninas, para os quatro primeiros subtestes, e outra de 188 crianças, sendo 105 meninos e 83 meninas; o coeficiente de correlação intraclasse foi calculado tomando em conta diferenças na tolerância dos examinadores, assim como os erros aleatórios, de acordo com Shrout & Fleiss (1979). Os coeficientes encontrados nestes estudos foram: 0,96 para Compreensão, 0,94 para Vocabulário, 0,96 para Semelhanças, 0,94 para Quebra-cabeça e 0,88 para Desenho Geométrico. Estes resultados indicam, portanto, que as regras de pontuação destes subtestes são suficientemente objetivas para que diferentes examinadores obtenham escores semelhantes.

Em conclusão, os resultados deste estudo indicaram que o Teste WPPSI-R apresentou boa confiabilidade quando aplicado por psicólogas treinadas, mostrando-se bastante adequado para ser utilizado no estudo sobre avaliação do desenvolvimento cognitivo de prematuros de muito baixo peso do IFF.

Referências

- ALLEN, M. C.; DONOHUE, P. K. & DUSMAN, A. E., 1993. The limit of viability – Neonatal outcome of infants born at 22 to 25 weeks' gestation. *New England Journal of Medicine*, 329:1597-1601.
- ARMITAGE, P. & BERRY, G., 1994. Further experimental designs. In: *Statistical Methods in Medical Research* (P. Armitage & G. Berry, eds.), pp. 237-282, Cambridge: Blackwell Science Ltd./University Press.
- BARTKO, J. J., 1994. General methodology II. Measures of agreement: A single procedure. *Statistics in Medicine*, 13:737-745.
- BENNET, F. C. & SCOTT, D. T., 1997. Long-term perspective on premature infant outcome and contemporary intervention issues. *Seminars in Perinatology*, 21:190-201.
- DUNN, G., 1989. *Design and Analysis of Reliability Studies – The Statistical Evaluation of Measurements Errors*. New York: Hodder & Stoughton/Oxford University Press.

- FANAROFF, A. A.; WRIGHT, L. L.; STEVENSON, D. K.; SHANKARAN, S.; DONOVAN, E. S.; EHRENKRANZ, R. A.; YOUNES, N. & KORONES, S. B., 1995. Very-low-birth-weight outcomes of The National Institute of Child Health and Human Development Neonatal Research Work, May 1991 through December 1992. *American Journal of Obstetrics and Gynecology*, 173:1423-1431.
- FEDRIZZI, E.; INVERNO, M.; BOTTEON, G.; ANDERLONI, A.; FILIPPINI, G. & FARINOTTI, M., 1993. The cognitive development of children born preterm and affected by spastic diplegia. *Brain Development*, 15:428-432.
- FLEISS, J. L., 1981. Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement*, 5:105-112.
- HACK, M. & FANAROFF, A. A., 1988. How small is too small? Considerations in evaluating the outcome of the tiny infant. *Clinics in Perinatology*, 15:773-788.
- HACK, M.; FRIEDMAN, H. & FANAROFF, A. A., 1996. Outcomes of extremely low birth weight infants. *Pediatrics*, 98:931-937.
- HUTTON, J. L.; PHAROAH, P. O. D.; COOKE, R. W. I. & STEVENSON, R. C., 1997. Differential effects of preterm birth and small gestational age on cognitive and motor development. *Archives Disease of Childhood*, 76:F75-F81.
- KOK, J. H.; DEN OUDEN, L.; VERLOOVE-VANHORICK, S. P. & BRAND, R., 1998. Outcome of very preterm small for gestational age infants: The first nine years of life. *British Journal of Obstetrics and Gynaecology*, 105:162-168.
- LARGO, R. H.; PFISTER, D.; MOLINARI, L.; KUNDU, S.; LIP, A. & DUC, G., 1989. Significance of prenatal, perinatal and postnatal factors in the development of AGA preterm infants at five to seven years. *Developmental Medicine and Child Neurology*, 31:440-456.
- McCARTON, C. M.; BROOKS-GUNN, J.; WALLACE, I. F.; BAUER, C. R.; BENNETT, F. C.; BERNBAUM, J. C.; BROYLES, S.; CASEY, P. H.; McCORMICK, M. C.; SCOTT, D. T.; TYSON, J.; TONASCIA, J. & MEINERT, C. L., 1997. Results at age 8 years of early intervention for low-birth-weight premature infants. *JAMA*, 277:126-132.
- MÉIO, M. D. B. B., 1999. *Como Estão Sobrevivendo os Pequenos Prematuros? Um Olhar Sobre a População do Instituto Fernandes Figueira*. Dissertação de Mestrado, Rio de Janeiro: Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro.
- RESNICK, M. B.; GOMATAN, S. V.; CARTER, R. L.; ARIET, M.; ROTH, J.; KILGORE, K. L.; BUCCIARELLI, R. L.; MAHAN, C. S.; CURRAN, J. S. & EITZMAN, D. V., 1998. Educational disabilities of neonatal intensive care graduates. *Pediatrics*, 102:308-314.
- SHROUT, P. & FLEISS, J. L., 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86:420-428.
- SOMMERFELT, K.; TROLAND, K.; ELLERTSEN, B. & MARKESTAD, T., 1996. Behavioral problems in low-birth-weight preschoolers. *Developmental Medicine and Child Neurology*, 38:927-940.
- SPSS INCORPORATION, 1995. *SPSS for Windows. Statistical Package for the Social Sciences. Release 7.0*. Chicago: SPSS Inc.
- STREINER, D. L., 1995. Learning how to differ: Agreement and reliability statistics in psychiatry. *Canadian Journal of Psychiatry*, 40:60-66.
- WECHSLER, D., 1989. *WPPSI-R – Wechsler Preschool and Primary Scale of Intelligence – Revised*. San Antonio: The Psychological Corporation.
- YATES, F., 1936. Incomplete randomized blocks. *Annals of Eugenics*, 7:121-140.