Revista Interamericana de Psicología/Interamerican Journal of Psychology - 2011, Vol. 45, Num. 3, pp. 419-428

Reliability of the Brazilian WAIS-III in Depression

Andre Goettems Bastos[1]
Bruna Monego Gomes
Denise Ruschel Bandeira
Clarissa Marceli Trentini
Federal University of Rio Grande do Sul, Brazil

Abstract

The third edition of the Wechsler Adult Intelligence Scale (WAIS-III) has been extensively studied since it was first developed. The WAIS-III is broadly used in psychological assessments in Brazil and other Latin American countries. WAIS-III has shown to have good psychometric properties, but its reliability has not been evaluated in clinical samples in Brazil. This study evaluated the reliability of WAIS-III for a sample of patients with depression using test-retest. The coefficients obtained were compared with those of standardization samples for the Brazilian and American versions of the scale. This study enrolled 83 adults aged 26 to 34 years. All participants were evaluated at two time points at an interval of 12 months. The Pearson correlation coefficient was used to analyze data and evaluate WAIS-III reliability, and the Fisher z test was used for the comparisons between samples. Results showed that WAIS-III has stability across time in depressed individuals.
Keywords: Wechsler Adult Intelligence Scale, reliability, depression

Estabilidad Temporal de la WAIS-III Brasileña en sujetos deprimidos

Resúmen

La tercera edición de la Escala de Inteligencia de Wechsler para Adultos (WAIS-III) ha sido ampliamente estudiada desde que fue desarrollado por primera vez. El WAIS-III es ampliamente uilizada en la evaluación psicológica en Brasil y en otros países de Latinoamérica. La WAIS-III ha demostrado tener buenas propiedades psicométricas, pero su confiabilidad no ha sido evaluada en muestras clinicas en Brasil. Este estudio evaluó la fiabilidad de la WAIS-III en una muestra de pacientes con depresión mediante el método de test y retest. Los coeficientes obtenidos fueron comparados con los de las muestras de estandarización para las versiones de Brasil y de Estados Unidos de la escala. Todos los participantes fueron evaluados en dos momentos en un intervalo de 12 meses. El coeficiente de correlación de Pearson fue usado para analizar los datos y evaluar la confiabilidad del WAIS-III, y la prueba exacta de Fischer (z test) fue utilizada para las comparaciones entre las muestras. Los resultados mostraron que la WAIS-III tiene una buena estabilidad a través del tiempo en las personas deprimidas.
Palabras clave: escala de inteligencia de Wechsler para Adultos, confiabilidad, depresion.

The study of the reliability of a psychological test is part of the fundamental psychometric aspects of psychological testing. Test reliability ensures the trustworthiness of its use and is defined as the quality of testing scores that confirm that they are sufficiently consistent, free of measurement error and, therefore, useful (Urbina, 2007).

Test reliability may be measured using different methods. Reliability, generally, has two aspects: the problem of stability across time points and the problem of internal consistency. Within the study of reliability, stability across time, or test-retest reliability, provides information about how much passing time affects the results of a test when it is reapplied later on in a different context. Internal consistency, in turn, is generally assessed using Cronbach alpha to detect it between test items (Hair, Black, Babin, Anderson, & Tatham, 2009).

This study evaluated the reliability of the third edition of the Wechsler Adult Intelligence Scale (WAIS-III) in a test-retest model with adult patients with depression. It is part of a larger project to compare cognitive changes of patients with depression who receive different types of treatment. All participants were evaluated cognitively using the WAIS-III before the beginning of the treatment and at subsequent time points at 6-month intervals. The results of the first two evaluations are presented here. The sample was a clinical group, and not healthy individuals. Moreover,

[1] Professor of Contemporary Institute of Psychoanalysis and Transdisciplinarity in Porto Alegre.
Email: andregbastos@terra.com.br

all were receiving some type of treatment. The effect of these interventions on test reliability is discussed.

Below we review studies about the psychometric stability of WAIS-III along time, the definitions of major depressive disorders, and the cognitive changes expected in patients with depression. After that, the method used in this study is described, followed by our results, discussion and final considerations.

WAIS-III reliability

The Wechsler Adult Intelligence Scale is a general intelligence test first published in 1955 as a revised version of the Wechsler-Bellevue test (1939). Wechsler (1975, p.136) defines intelligence as "the global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment".

WAIS is one of the most important tests for the clinical evaluation of the intellectual capacity of adults aged 16 to 89 years. It is broadly used in Brazil and in other countries of Latin America, since not every country in Latin America has already validated for its population the WAIS-IV which is being used now in USA. Although it has essential characteristics common to its predecessors, WAIS-III, in its third edition, provides current normative data, both in the original American sample and in the Brazilian population, as well as standardized material and application procedures (Nascimento, 2004; Wechsler, 2004). It is a tool to measure an individual's intellectual and cognitive functioning broadly and globally (Nascimento, 2004). This instrument provides information to detect and evaluate the status and functioning of intellectual and cognitive conditions as a whole. WAIS-III was adapted for Brazilian populations (Nascimento, 2004) and is widely used because it is a closed battery of tests with subtests that evaluate different cognitive functions and compare individuals with their peers and with their own cognitive performance.

The test is made up of 14 subtests that provide measures for the following scales and factors: verbal IQ, performance IQ, full scale IQ, verbal comprehension index, perceptual organization index, working memory index, processing speed index. All WAIS-III subtests and indices appear to have good reliability.

The term "reliability" is defined as "temporal stability" in English. Urbina (2007) points out that the correlation between the scores in test-retest situations is a test-retest stability coefficient, or reliability. Therefore, the term to be used here is reliability.

Reliability and validity studies were carried out to evaluate the WAIS-III psychometric properties in Brazil (Nascimento, 2000). Results in general indicated that the WAIS-III is appropriate for use with adolescents older than 16 years and adults.

The reliability of the Brazilian version of WAIS-III was studied in the form of a subsample, according to Nascimento (2004). Results showed that the adapted WAIS-III had good reliability. The coefficients for the vocabulary, information and comprehension subtests were very high (above 0.90) and for similarities, arithmetic, picture completion, digit-symbol coding, block design, matrix reasoning and symbol search were also very good (above 0.80). In the other subtests, reliability coefficients were above 0.70, which is reasonable. The object assembly subtest had the lowest coefficient (0.65). IQ scores and factor indices were above 0.90, which is considered excellent (Nascimento, 2004).

Mean results of the test-retest conducted by Nascimento (2004) revealed that the results of the second application were higher than those in the first. Mean IQ differences were 1.03 for verbal IQ, 5.33 for performance IQ, and 3.03 for full scale IQ. The analysis of factor scales revealed that the perceptual organization index had the greatest increase at a mean 4.98 weighted points. The index with the lowest difference was the one for working memory, with a mean increase of 0.69. According to Nascimento (2004), the differences in results between the two applications may result from practice effects. Moreover, the author calls attention to the fact that differences are usually found in studies about reliability.

The reliability of the American version of WAIS-III was evaluated using test-retest at an interval of 2 to 12 weeks. Mean interval was 34.6 days. The WAIS-III sample comprised 349 participants aged 16 to 89 years. Of all the participants, 50.3% were women. Schooling ranged from 8 to over 16 years of formal education (Wechsler, 2004).

In the analysis of the American WAIS-III subtests, Wechsler (2004) reports that mean temporal reliability coefficients for most subtests of the full sample, except picture arrangement, search symbols and object assembly, ranged from 0.82 to 0.93. For the vocabulary, digit span, information and matrix reasoning, coefficients were very high (0.90). The arithmetic, comprehension, letter-number sequencing, picture completion, digit-symbol coding and block design coefficients ranged from 0.82 to 0.88. The symbol search subtest had a test-retest coefficient of 0.77. The picture arrangement and object assembly coefficients were lower (0.74 and 0.70), but equal to or higher than the reliability coefficients found for the American WAIS-R (Wechsler, 2004). In the analysis of reliability coefficients for the WAIS-III IQ scores and factor scales, Wechsler (2004) found that they ranged from 0.88 to 0.97. In general, WAIS-III coefficients are higher than the WAIS-R results.

Data about reliability for the American WAIS-III suggest that it is appropriate for all the age groups

already evaluated. Mean reliability coefficients for vocabulary and information are excellent (about 0.90); reliability of similarities, arithmetic, digit span, comprehension, digit-symbol coding and block design is very good (around 0.80); reliability for the other subtests is good (close to 0.70). As the data reported by Wechsler (2004) confirmed, mean retest scores are higher than initial test scores. According to Wechsler, these differences, which result primarily from practice effects (Wechsler, 2004), are found in the final scores of verbal IQ, performance IQ and full scale IQ. Differences ranged from 2.0 to 9.2 weighted points. All the correlations were corrected for the variability of the standardization sample, as recommended by Allen and Yen (1979) and Magnusson (1967).

According to Wechsler (2004), WAIS-III has a greater psychometric strength than WAIS-R. Several studies using WAIS-R showed that the test has good reliability both for normal individuals and clinical populations (Atkinson et al., 1990; Parker, Hanson, & Hunsley, 1988). In contrast, a recent study conducted by Matarazzo and Herman (2008) found significant changes in the verbal IQ and performance IQ results using WAIS-R in the context of retesting the same individuals. The authors recommend that this should be taken into consideration when interpreting results in a clinical context.

Lemay, Bedard, Rouleau and Tremblay (2004) conducted a test-retest study using WAIS-III in Canada and found good reliability, particularly for the subtests whose scores were associated with processing speed. Iverson (2001) confirmed it and pointed out that WAIS-III has excellent reliability. According to the author, this may be useful in clinical contexts to find out exactly which cognitive functions have been affected.

One exclusion criterion used by Wechsler (2004) for the participants in the standardization sample was psychopharmacological treatment with antidepressants, anti-anxiety medications, or antipsychotic drugs. According to Wechsler (2004) such exclusion was justified by the fact that these medications may impair an individual's cognition, which would affect reliability in a standardization study. Several other clinical conditions were adopted as exclusion criteria; for example: exaggerated consumption of alcoholic beverages, epilepsy, and bipolar mood disorder. Although Wechsler (2004) did not mention major depressive disorder, studies using WAIS-III in specific clinical groups with different disease found that WAIS-III can be used satisfactorily in clinical populations (Curtis, Greve, & Bianchini, 2009; Ryan, Tree, Morris, & Gontkovsky, 2006).

Lezak (2004) called attention to the fact that reliability might only be ensured when dealing with normal healthy individuals. When individuals in specific clinical groups are examined, attention to the test reliability is even more important. For healthy individuals, retest results are expected to be similar, which is not necessarily the rule when testing clinical groups. Several "good" tests that meet statistical criteria for reliability may lack clinical measurement validity for a certain disease.

Reliability of test-retest results for individuals with neuropsychiatric diseases may be practically absent (Lezak, 2004) because of the particular progression of a certain disease or even the effects of treatment. Lezak (2004) also points out that the best test for the clinical evaluation of neuropsychiatric patients is the one that is more sensitive to fluctuations in patient performance.

In this study, the pathology under study is depression. This neuropsychiatric disease and its expected effects on cognition should be better known to determine the level of reliability that may be expected.

Major Depressive Disorder: Concept and Associated Cognitive Aspects

Depression is the psychopathology with the greatest incidence in the statistics of mental diseases according to the World Health Organization (WHO) (http://www.who.int, 2009). Moreover, according to the WHO, in 2020 it will be one of the diseases with the greatest global impact on quality of life, second only to cardiovascular diseases. The Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR, American Psychiatry Association, 2002) uses the 296.2x codes for depressive disorders.

According to the WHO (2009), depression is a frequent mental disease that is usually characterized by sadness, loss of interest and motivation for activities and lack of energy for daily tasks. It is estimated that 5-10% of the population in all age groups has detectable depression that requires treatment. The risk of having depression along life is 10-20% among women and a little less among men (http://www.who.int, 2009).

Depression is a rapidly growing disease (Branco et al., 2009). In addition, it is cause of functional and behavioral disability because it affects life motivational aspects, shakes personality and impairs cognition (Zorzetto Filho, 2009).

Studies in the literature show that cognitive changes are usually found in patients with depression (Chepenik, Cornew, & Farah, 2007). They are primarily changes in processing and organization of perceptual contents, working memory, attention, executive functions, cognitive control and inhibitory processes, and cognitive processing speed (Harvey et al., 2004; Marvel & Paradiso, 2004; Rogers et al., 2004; Steele & Lawrie, 2004). The DSM-IV-RT (2002) includes, among other symptoms of depression, psychomotor retardation, attention deficits and impaired concentration. The

International Classification of Diseases version 10 (ICD-10, WHO, 1995) also lists reduced attention and concentration.

Stefanis and Stefanis (1999) found that the most frequent cognitive changes in depression are attention and concentration deficits and memory and learning impairments. Moreover, Del Porto (1999) points out that processing speed and reasoning processes are slower in these patients.

Studies have found associations between more severe depressive conditions and impairments in the performance of neuropsychological tests (Robbins, Joyce, & Sahakian, 1992; Zorzetto Filho, 2009). However, Miller (1975) expanded this concept and quoted studies that found an association between moderate depressions and significant impairment in several cognitive, motor, perceptual and communicative tasks. Deficits were also found in full scale intelligence scores, visuospatial abilities, memory, learning, abstract reasoning, language and processing speed. Their conclusion was that, in all types of depression, there is some cognitive impairment.

Several studies found that depression has a broad and consistent influence on cognitive aspects. This raises the possibility that cognition as a whole is affected due to the onset of depression. Findings that a cognitive improvement takes place before improvement of the depressive mood during treatment support this possibility (Dunkin et al., 2000). In clinical populations with moderate to severe depression, significant cognitive deficits are estimated to be high and to affect at least 21% of the patients (Gualtieri & Morgan, 2008).

Moreover, findings of neuropsychological changes in cognition among patients with depression are corroborated by neuroimaging studies (Davidson, Pizzagalli, Nitschke, & Putnam, 2002; Marvel & Paradiso, 2004). Apparently, there are variable changes in the functional neuroanatomy of patients with depression (Murphy, Nimmo-Smith, & Lawrence, 2003). Such findings have been confirmed in several neuroimaging studies that found changes in brain functioning among patients with depression (Rogers et al., 2004; Steele & Lawrie, 2004).

The search for the most indicated type of treatment for depression is still the focus of several studies (Greenberg & Goldman, 2009; Tomba & Fava, 2009). Among the options are psychotherapy, pharmacotherapy (Ingram, 2009; Kaplan & Sadock, 2008; Greenberg & Goldman, 2009; Tomba & Fava, 2009) and combinations of both (Greenberg & Goldman, 2009; Tomba & Fava, 2009).

Some authors still investigate what types of depression may be treated with drugs in combination with psychotherapy (Greenberg & Goldman, 2009, De Jonghe et al., 2004). Combined treatments using psychotherapy and drugs have been prescribed particularly in cases of severe depression with risk of death or refractory to monotherapy (Tomba & Fava, 2009).

The cognitive changes caused by depressive disorders demand the use of adequate instruments for evaluations. Therefore, this study evaluated the reliability of WAIS-III in a sample of patients with depression and compared results with those obtained in the standardization samples of the Brazilian and American versions.

Method

Participants

The group of participants comprised 83 adults with a diagnosis of moderate major depressive disorder. Participant age ranged from 26 to 34 years (M=29.86; SD=2.47), and 64.2% were women. Most had a college education (54.2%) and were single (55.4%).

Instruments

The instruments used in this study estimated the level of depression and evaluated the broad neuropsychological functioning of patients with depression.

Questionnaire about sociodemographic data. The authors developed a questionnaire to explore information that might provide a better sample description and that included age, educational level and marital status.

Wechsler Adult Intelligence Scale, 3rd edition (WAIS-III). WAIS-III is an instrument to measure intellectual and cognitive functioning broadly and globally (Nascimento, 2004). It is made up of 14 subtests, in which vocabulary, similarity, arithmetic, digit span, information, comprehension and letter-number sequencing belong to the verbal IQ scale, and picture completion, digit-symbol coding, block design, matrix reasoning, picture arrangement, symbol search and object assembly are part of the performance IQ scale. The full scale IQ score is obtained by adding the verbal and performance IQ scores. In addition to these measures, it is also possible to calculate indices for verbal comprehension, perceptual organization, working memory and processing speed. WAIS was first published in 1955 as a revision of the Wechsler-Bellevue test (1939) and the adaptation for Brazilian populations was conducted by Nascimento (2004). In her national study, subtest internal consistency was calculated according to age groups. Mean values ranged from 0.82 to 0.92, except for object assembly, which had a mean value of 0.66.

Beck Depression Inventory. Beck Depression Inventory (BDI) assesses the presence and intensity of depression symptoms. It is a self-report inventory with 21 items; scores from 0 to 11 indicate minimal symptoms of depression; from 12 to 19, mild depression;

from 20 to 35, moderate depression; and from 36 to 63, severe depression. This instrument was developed by Beck, Ward, Mendelson, Mock and Erbaugh (1961) and validated for use in Brazil by Cunha (2001).

Data collection and ethical concerns

This study started in a mental health outpatient service in Porto Alegre/BR. Screening for selection and participant identification were conducted together with the team of local specialized psychologists and took into consideration the list of patients that sought assistance, had clear signs of depression and moderate depression according to the BDI score.

After participants were identified, one of the authors contacted them for a personal introduction and to find out whether they would accept to participate in a study whose purpose was to evaluate and understand some cognitive aspects that might be implicated in the treatment that they had sought. Further information was provided about the objectives of the study, and an informed consent term was prepared to explain ethical concerns and the confidentiality of testing results. After the consent term was signed by the patient, data collection started. No material used in tests had any direct patient identification, and all forms received identification codes. All the ethical aspects have been followed, and the study was approved by the Ethics in Research Committee of the institution where it was conducted.

Data analysis procedures

Descriptive analyses used frequencies, means and standard deviations. The Pearson correlation coefficient was used to evaluate WAIS-III reliability. Data were analyzed using the Statistical Package for the Social Sciences 17.0 (SPSS). In addition, reliability coefficients for the clinical sample and the coefficients for the American and Brazilian versions were tested to evaluate significant differences between them. For that purpose, the Fisher z test was used (Bruning & Kintz, 1968). This procedure is valid and appropriate for comparisons between correlation coefficients between independent samples (Shavelson, 1996).

Results and Discussion

Table 1 shows descriptive statistical results, which consist of means, standard deviation, subtest score ranges, WAIS-III factor indices and IQ scores for the two evaluation time points for the clinical sample of patients with depression.

Table 1
WAIS-III descriptive data: mean, standard deviation and score range for clinical sample at two time points

	Test			Retest		
	M(SD).	Min	Max	M(SD).	Min.	Max
Subtest						
Vocabulary:	12.82(3.45)	4	19	13.05(3.24)	5	19
Similarity	13.34(3.23)	5	19	13.48(2.97)	6	19
Arithmetic	11.81(3.36)	5	19	12.43(2.75)	8	19
Digit span	9.80(2.39)	5	14	10.57(1.96)	6	15
Information	11.67(3.09)	6	19	11.92(2.85)	7	19
Comprehension	12.63(3.59)	5	19	13.02(3.16)	6	19
Letter-number sequencing	9.64(2.33)	5	15	10.61(1.87)	7	15
Picture completion	12.06(2.54)	5	19	12.53(2.06)	7	19
Digit-symbol coding	10.16(2.27)	3	15	10.88(1067)	5	14
Block design	12.59(2.83)	6	19	12.87(2.45)	8	18
Matrix reasoning	11.11(2.06)	6	15	11.55(1.55)	8	16
Picture arrangement	10.94(2.72)	6	18	11.25(2.30)	7	18
Symbol search	11.14(2.39)	5	17	11.37(1.97)	6	15
Object assembly	11.02(2.70)	5	17	11.16(2.50)	6	16
IQ						
Verbal IQ	111.77(14.93)	78	143	114.05(12.88)	87	143
Performance IQ	108.28(10.41)	84	128	110.94(8.11)	93	128
Full scale IQ	110.69(12.02)	84	136	113.19(9.89)	92	135
Factor scale						
Verbal comprehension	113.99(15.65)	78	145	115.17(14.16)	86	145
Perceptual organization	110.92(10.88)	83	132	113. 18(8.90)	92	134
Working memory	102.36(12.21)	78	128	106.78(8.82)	87	126
Processing speed	103.46(10.42)	71	127	105.88(7.64)	82	121

Mean score of subtests in the first test was 11.4 weighted points, whereas for the second it was 11.9 points. The weighted mean factor scale index in the first test was 107.7 points, and in the second, 110.3 points. IQ results had a mean weighted score of 110.2 points in the first test. In the second test, mean IQ was 112.7 weighted points.

Table 2 describes the reliability coefficients (r) of the clinical sample and of the Brazilian and American standardizations. The reliability coefficients of the clinical sample ranged from 0.72 (digit span) to 0.99 (information, verbal IQ, full scale IQ, and verbal comprehension factor). Reliability coefficient median was 0.96. Fifteen coefficients were significantly different from the coefficients in at least on of the standardization samples. Vocabulary, similarity, arithmetic, information, comprehension, matrix reasoning, picture arrangement, object assembly, full scale IQ and verbal comprehension factor were different from both standardization samples. Digit span, symbol search, verbal IQ and processing speed factor were significantly different from the results of the American standardization sample, whereas block design differed from the Brazilian standardization sample.

Table 2
WAIS-III reliability coefficients in clinical sample, Brazilian standardization sample and American standardization sample.

	Reliability coefficients (*r*)		
	This study (n=83)	Brazilian standardization (n=43)	American standardization (n=80)
Subtest			
Vocabulary:	0.98*	0.93	0.92
Similarity	0.97*	0.89	0.86
Arithmetic	0.96*	0.85	0.89
Digit span	0.72*(USA)	0.66	0.92
Information	0.99*	0.95	0.91
Comprehension	0.97*	0.90	0.82
Letter-number sequencing	0.73	0.73	0.77
Picture completion	0.92	0.80	0.86
Digit-symbol coding	0.88	0.85	0.81
Block design	0.96*(BR)	0.87	0.90
Matrix reasoning	0.78*	0.81	0.91
Picture arrangement	0.97*	0.76	0.79
Symbol search	0.95*(USA)	0.89	0.74
Object assembly	0.98*	0.65	0.71
IQ			
Verbal IQ	0.99*(USA)	0.97	0.97
Performance IQ	0.95	0.94	0.95
Full scale IQ	0.99*	0.97	0.98
Factor scale			
Verbal comprehension	0.99*	0.96	0.96
Perceptual organization	0.93	0.93	0.95
Working memory	0.91	0.88	0.93
Processing speed	0.95*(USA)	0.93	0.87

Note: * significant difference between this study and the other two standardization samples ($p<0.001$, $z>3.09$) according to Fisher z test for comparisons of correlation coefficients for independent samples. Significant differences found for only one of the standardization samples are indicated by "BR" (when different from Brazilian sample) and "USA" (when different from the American sample).

In general, results showed that WAIS III has adequate reliability when administered to individuals with depression. Nine of the 14 subtests had a high reliability coefficient, considering that the minimum value acceptable is 0.80 (Sattler, 1992; Urbina, 2007). In addition, there was an excellent reliability between factor scales and IQ score in the clinical sample, as all coefficients were ≥0.91. Reliability in most of the subtests seems to be strong enough to justify clinical decisions about the individual examined, probably

suggesting evidence of clinical validity for use of this instrument for individuals with depression, This may be accepted considering the principle that the concepts of reliability and validity are intrinsically associated, and that reliability is an indispensable condition for evidence of validity (Urbina, 2007).

The most accurate coefficients were full scale IQ, verbal IQ, verbal comprehension scale and information scale, both in the clinical sample (all 0.99) and in the Brazilian standardization sample (0.97, 0.97, 0.96 and 0.95). The same was seen in the American sample for full scale IQ (0.98) and verbal comprehension scale (0.96). Moreover, verbal IQ (0.97) and information (0.91) also had high coefficients in this sample. Zhu, Tulsky, Price and Chen (2001) pointed out that, in general, the reliability coefficients associated with the WAIS-III verbal scale are higher than those in the performance scale. Similarly, Urbina (2007) found that the constructs that measure verbal ability are more stable in any adult than the capacity of attention and memory, for example. This was actually found in our study. Apparently, the results of tests that evaluate attention and working memory are more susceptible to the influence of transient conditions or emotional states.

According to Iverson (2001), the WAIS-III test-retest correlations are considerably higher in normal populations. Our results did not confirm it. The correlations for the clinical sample were higher in 15 of the 21 (71.4%) comparisons with the other samples. The lowest correlations (<.80) were found in digit span (0.72), letter-number sequencing (0.73) and matrix reasoning (0.78), which means that those were the scores that varied the most from the first to the second test. In the case of a sample of patients with depression who are receiving treatment, this variation seems plausible because the subtests evaluate specific neuropsychological functions. In this case, they are working memory, attention, concentration and abstract reasoning (Banhato & Nascimento, 2007). They are exactly some of the functions that studies in the literature often describe as being impaired in patients with depression (Teng, Nakata, Rocca, & Yano, 2009). The lower coefficients and greater score variation, therefore, may be explained by a cognitive improvement of these patients, maybe due to the treatment received. Cognitive improvement due to specific treatments of adult patients with depression evaluated using the WAIS-III has already been suggested by other studies (Bastos & Trentini, 2009; 2010).

Reliability coefficients seem to vary in different clinical groups. However, this variation is usually associated with and is directly proportional to the symptoms expected for that specific clinical situation (Zhu, Tulsky, Price, & Chen, 2001; Zhu, Tulsky, & Rolfhus, 1999). This means that the difficulties associated with a certain clinical group will definitely affect score consistency. Therefore, traditional reliability coefficients will automatically increase or decrease, as if they were part or reflex of the disease that affects the individual. As an example, when they entered the study, the patients with depression had low mean values for digit span (9.80) and letter-number sequencing (9.64), two subtests that evaluate working memory, attention and concentration. The scores for these functions are usually lower in individuals with depressions (McDermott & Ebmeier, 2009). Also, because of the improvement probably resulting from the treatment received, these functions recovered and mean scores of the two subtests increased substantially (digit span=10.57; LNS=10.61), which lowered the reliability coefficient. The third subtest with a reliability coefficient below 0.80 was matrix reasoning (0.78), which also evaluates cognitive functions that are usually impaired by depression, such as cognitive flexibility, logical thinking, problem solving and decision making (Ganguli, Snitz, Bilt, & Chang, 2009). Therefore, the interpretation of mean score variation may also indicate that individuals with depression may have scores that lead to changes in reliability in specific cognitive areas, which seems to be an effect of depression itself, of the treatment against depression, or of both.

Several subtests had significantly different results for the clinical sample when compared with the standardization samples ($p<0,001$). Several reasons may explain such variation, such as the educational level of individuals in the three samples, the test-retest time interval, the men-to-women ratio, and even the specific characteristics of depression or the treatment that individuals in the clinical sample were receiving. However, statistical significance does not always reflect clinical importance.

Tulsky et al. (2003) drew attention to the fact that there are important differences between statistical and clinical significance. In other words, the difference between test and retest scores for the same individual may be statistically but not clinically significant, because the difference in scores may be found in the clinical population.

According to Kelley (2009), although statistical significance reveals data reliability, it does not say anything about the probability of study replication, effect size, and clinical importance of findings. This question has already been discussed by other authors (Cohen, 1990; 1994; Rosnow & Rosenthal, 1989). The differences found in this study suggest that there is a very small probability that these differences were the result of chance ($p<0.001$).

A better understanding of the logics that underlies the null hypothesis in statistical calculations is crucial to avoid misinterpretations of calculations and importance for clinical practice (Kelley, 2009). This problem may occur when too much attention is given to the statistical significance of data and little importance is assigned to the effect size and clinical significance. However, it is very difficult to define clinical significance (Cohen, 1990).

Clinical significance depends necessarily on the frequency (number of times) that the entity appears in a certain population (Tulsky et al., 2003). In addition, the comparison of scores for a clinical sample with scores of a standardization sample only reveals how typical the score is, that is, the greater the discrepancy, the rarer is the score, and higher is the likelihood that it lies outside the normal curve. Tulsky et al. (2003) called attention to the fact that a certain score pattern becomes relevant if it is rare among the normal population but frequent in the clinical sample. Therefore, score patterns may be useful to distinguish an individual from the clinical group or from a normal group.

According to Groth-Marnat (2003), clinical decisions about WAIS-III variability should be made cautiously. The interpretation of each subtest demands a theoretical basis, observation of each individual, and the integration of the specificities of each case. As a rule, inferences will gain support in direct proportion to the number of subtest scores used to draw such inferences. Several steps should be taken to interpret score variation in a clinical population, such as the determination of the significance of this change, the development of hypotheses about the meaning of these fluctuations, and the integration of these hypotheses into relevant additional information about the individual being examined. According to Groth-Marnat (2003), relevant data are age, educational level, ethnic group, sex, and the existence or not of brain damage, including type and site of lesion.

The source of variation seems to be fundamental to justify the need for a neuropsychological evaluation of patients with depression. If the cognitive functions with the greatest impairment are known, it is possible to seek more efficient and better indicated treatments. Moreover, the cognitive profile of an individual, as defined by WAIS-III, may be a neuropsychological marker that may enable the establishment of a differential diagnosis.

Final Considerations

Changes in neuropsychological tests should be interpreted carefully, and the specific characteristics of the population under study should be taken into consideration. Data in this study may be valuable for psychologists, neuropsychologists, neurologists and psychiatrists that treat or evaluate patients with depression, as well as for those who work with individuals with several neuropsychiatric and psychological disorders that are usually associated with cognitive impairments.

Moreover, further studies about WAIS reliability should be conducted with other clinical groups. Test-retest studies should include a large variety of clinical samples to evaluate the impact of random variability of answers, of different time intervals, and of educational, medical and psychological changes in the individuals under study.

The ability to make diagnostic predictions about score probabilities in retests should not be overestimated, not even in clinical samples that have a good level of reliability, as it does not replace good clinical judgment. The usefulness of statistical data is to provide quantitative evidence so that the clinician may consider them during the process of making decisions about a patient. Therefore, we should attempt to define how clinically significant our findings are, that is, how the changes in test-retest scores reflect actual changes in individuals.

References

Allen, M., & Yen, W. (1979). *Introduction to measurement theory. Monterrey:* Brooks-Cole.

American Psychiatry Association. (2002). DSM-IV – *Manual diagnóstico e estatístico de transtornos mentais*. Edição texto revisado. Porto Alegre: Artmed.

Atkinson, L., Bowman, T., Dickens, S., Blackwell, J., Vasarhelyi, J., Szep, P., Dunleavy, B., MacIntyre, R., & Bury, A. (1990). Stability of Wechsler Adult Intelligence Scale - Revised factor scores across time. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2*(4), 447-450.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletim, 66*, 1-29.

Banhato, E., & Nascimento, E. (2007). Função executiva em idosos: Um estudo utilizando subtestes da escala WAIS-III. *Psico-USF, 12*(1), 65-73.

Bastos, A. G., & Trentini, C. M. (2009). *Análise longitudinal das mudanças cognitivas em pacientes deprimidos em psicoterapia psicanalítica*. Pôster apresentado no IV Congresso de Avaliação Psicológica, Campinas, SP, Brasil.

Bastos, A. G., & Trentini, C. M. (2010). *Resposta cognitiva de pacientes deprimidos ao tratamento biológico com fluoxetina.* Pôster apresentado no 6º Congresso Brasileiro de Cérebro, Comportamento e Emoções, Gramado, RS, Brasil.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 461-571.

Branco, B., Fernandes, F., Powell, V., Quarantini, L., Miranda-Scippa, A., & Lacerda, A. (2009). Depressão: Considerações diagnósticas e epidemiológicas. In: A., Lacerda, L., Quarantini, A., Miranda-Scippa & J., Del Porto. *Depressão: Do neurônio ao funcionamento social.* (pp.13-26). Porto Alegre: Artmed.

Bruning, J., & Kintz, B. (1968). *Computational handbook of statistics.* Glenview, IL: Scott, Foresman and Company.

Chepenik, L., Cornew, L., & Farah, M. (2007). The influence of sad mood on cognition. *Emotion, 7*(4), 802-811.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist, 49,* 997-1003.

Cunha, J. A. (2000). *Psicodiagnóstico-V.* Porto Alegre: Artmed.

Cunha, J. A. (2001). *Manual da versão em português das Escalas Beck.* São Paulo: Casa do Psicólogo.

Curtis, K., Greve, K., & Bianchini, K. (2009). The Wechsler Adult Intelligence Scale-III and malingering in traumatic brain injury: Classification accuracy in known groups. *Assessment, 16*(4), pgs. 401-414.

Davidson, R., Pizzagalli, D., Nitschke, J., & Putnam, K. (2002). Depression: Perspectives from affective neuroscience. *Annual Review of Psychology, 53,* 545-574.

De Jonghe, F., Kool, S., Van Aalst, G., Dekker, J., & Peen, V. (2001). Combining psychotherapy and antidepressants in the treatment of depression. *Journal of Affective Disorders, 64,* 217-229.

Del Porto, J. (1999). Conceito e Diagnóstico. *Revista Brasileira de Psiquiatria, 21*(1), 06-11.

Drevets, W. (2003). Neuroimaging abnormalities in the amygdala in mood disorders. *Annals of the New York Academy of Sciences, 985,* 420-444.

Dunkin, J., Leuchter, J., Cook, I., Kasl-Godley, J., Abrams, M., & Rosenberg-Thompson, S. (2000). Executive dysfunction predicts non-response to fluoxetine in major depression. *Journal of Affective Disorders, 60,* 13-23.

Eugene, F., Levesque, J., Mensour, B., Leroux, J., Beaudoin, G., & Bourgouin, P. (2003). The impact of individual differences on the neural circuitry underlying sadness. *NeuroImage, 19, 354-364.*

Fachel, J., & Camey, S. (2000). Avaliação psicométrica: A qualidade das medidas e o entendimento dos dados. In: J. Cunha. *Psicodiagnóstico – V.* Porto Alegre: ArtMed.

Ganguli, M., Snitz, B., Bilt, J., & Chang, C. (2009). How much depressive symptoms affect cognition at the population level? The monongahelayoughiogheny healthy aging team (MYHAT) study. International *Journal of Geriatric Psychiatry, 10.1002/*gps.2257.

Greenberg, R., & Goldman, E. (2009). Antidepressants, psychotherapy or their combination: Weighing options for depression treatments. *Journal of Contemporary Psychotherapy, 39,* 83-91.

Gualtieri, C., & Morgan, D. (2008). The frequency of cognitive impairment in patients with anxiety, depression, and bipolar disorder: Unaccounted source of variance in clinical trials. *Journal of Clinical Psychiatry, 69*(7),1122-1130.

Harvey, P., Le Bastard, G., Pochon, J., Levy, R., Allilaire, J., & Dubois, B. (2004). Executive functions and updating of the contents of working memory in unipolar depression. *Journal of Psychiatric Research, 38,* 567-576.

Heijden, P. van der., Donders, J. (2003). WAIS-III factor index score patterns after traumatic brain injury. *Assessment, 10*(2), 115-122.

Ingram, R. (2009). *The international encyclopedia of depression.* New York: Springer Publishing Company.

Iverson, G. (2001). Interpreting change on the WAIS-III/WMS-III in clinical samples. *Archives of Clinical Neuropsychology, 16*(2),183-191.

Kaplan, B., & Sadock, V. (2008). *Kaplan & Sadock's concise textbook of psychiatry.* Philadelphia: Lippincott, Willians & Wilkins.

Kelley, J. (2009). The perils of p-values: Why tests of statistical significance impede the progress of research – an open letter to psychotherapy researchers. In: R. Levy, & J. Ablon. *Handbook of Evidence-Based Psychodynamic Psychotherapy.* MA: Humana Press.

Lange, R., Schoenberg, M., Woodward, T., & Brickell, T. (2005). A development of the WAIS-III estimate of premorbid ability for Canadians (EPAC). Archives of clinical neuropsychology: *The official journal of the National Academy of Neuropsychologists, 20*(8), 1009-1024.

Lemay, S., Bédard, M., Rouleau, I., & Tremblay, P. (2004). Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *Clinical Neuropsychology,18*(2), 284-302.

Lezak, M. (2004). *Neuropsychological Assessment.* New York: Oxford University Press.

Lykken, D. (1968). Statistical significance in psychological research. *Psychological Bulletim, 70,* 151-159.

McDermott, L., & Ebmeier, K. (2009). A meta-analysis of depression severity and cognitive function. *Journal of Affective Disorders,119*(1), 1-8.

Magnusson, D. (1967). Test theory. Reading: Addison-Wesley.

Marvel, C., & Paradiso, S. (2004). Cognitive and neurological impairments in mood disorders. *Psychiatric Clinics of North America, 27,* 19-36.

Matarazzo, J., & Herman, D. (2010) Base rate data for the WAIS-R: Test-retest stability and VIQ-PIQ differences. *Schizophrenia Bulletim, 36*(2), 400-409.

Miller, W. (1975). Psychological deficit in depression. *Psychological Bulletin, 82*(2), 238-260.

Murphy, F., Nimmo-Smith, I., & Lawrence, A. (2003). Functional neuroanatomy of emotions: A meta-analysis. *Cognitive, Affective, and Behavioral Neuroscience, 3,* 207-233.

Nascimento, E. (2000). WAIS-III. In: J. Cunha. *Psicodiagnóstico – V.* Porto Alegre: ArtMed.

Nascimento, E. (2004). *Manual da Escala de Inteligência Wechsler para Adultos.* São Paulo: Casa do Psicólogo.

Parker, K., Hanson, R., & Hunsley, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin, 103*(3), 367-373.

Phan, K., Wager, T., Taylor, S., & Liberzon, I. (2002). Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage, 16,* .331-348.

Philips, M., Drevets, W., Rauch, S., & Lane, R. (2003a). Neurobiology of emotional perception I: The neural basis of normal emotion perception. *Biological Psychiatry, 54,* 504-514.

Philips, M., Drevets, W., Rauch, S., & Lane, R. (2003b). Neurobiology of emotional perception II: Implications for major psychiatric disorders. *Biological Psychiatry, 54,* 515-528.

Robbins, T., Joyce, E., & Sahakian, B. (1992). Neuropsychology and neuroimaging. In: E. Paykel, (Ed.). *Handbook of affective disorders.* Edimburgh: Churchill Livingstone.

Rogers, R., Kasai, K., Koji, M., Fukuda, R., Iwanami, A., & Nakagome, K. (2004). Executive and prefrontal dysfunction in unipolar depression: A review of neuropsychological and imaging evidence. *Neuroscience Research, 50,* 01-11.

Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and justification of knowledge in psychological science. *American Psychologist, 44,* 1276-1284.

Rozeboom, W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletim, 57,* 416-428.

Ryan, J. J., Tree, H. A.; Morris, J., Gontkovsky, S. T. (2006). Wechsler Adult Intelligence Scale-III inter-subtest scatter: A comparison of brain-damaged patients and normal controls. *Journal of clinical psychology, 62*(10), 1319-1326.

Sattler, J. (1992). *Assessment of children: Revised and updated third edition.* San Diego, CA: Author.

Shavelson, R. (1996). *Statistical reasoning for the behavioral sciences.* Needham Heights, MA: Allyn & Bacon.

Steele, J., & Lawrie, S. (2004). Segregation of cognitive and emotional function in the prefrontal cortex: Stereotactic meta-analysis. *NeuroImage, 21,* 868-875.

Stefanis, C., & Stefanis, N. (1999). Diagnosis of depressive disorders: A review. In: M. Maj, & N. Sartorius (Ed.). *Depressive disorders.* Chichester: John Wiley & Sons.

Teng, C., Nakata, A., Rocca, C., & Yano, Y. (2009). Depressão e cognição. São Paulo: Editora Atheneu.

Tolman, A. (2009). *Depressão em adultos: As mais recentes estratégias de avaliação e tratamento.* Porto Alegre: Artmed.

Tomba, E., & Fava, G. (2009). The sequential combination of pharmacotherapy and psychotherapy in mood disorders. *Journal of Contemporary Psychotherapy, 39,* 101-109.

Urbina, S. (2007). *Fundamentos da testagem psicológica.* Porto Alegre: ArtMed.

Wallerstein, R. (1986). *Forty-Two lives in treatment: A study of psychoanalysis and psychotherapy.* New York: Guilford Press.

Watkins, C. E. Jr., Campbell, V. L. (1984). The test-retest reliability and stability of the WAIS-R in a sample of mentally retarded adults. *Journal of Clinical and Experimental Neuropsychology, 6*(4), 351-366.

Wechsler, D. (1939). *The measurement of adult intelligence.* Baltimore: Williams & Wilkins.

Wechsler, D. (1975). Intelligence defined and undefined: A relativistic appraisal. *American Psychologist, 30,* 135-139.

Wechsler, D. (2004). *WAIS-III – Manual Técnico.* São Paulo: Casa do Psicólogo.

Zhu, J., Tulsky, D., & Rolfhus, E. (1999). WAIS-III internal consistency reliability for special clinical groups. *Archives of Clinical Neuropsychology, 14*(8), 724-725.

Zhu, J., Tulsky, D., Price, L., & Chen, H. (2001). WAIS-III reliability data for clinical groups. *Journal of the International Neuropsychological Society, 7,* 862-866.

Zorzetto Filho, D. (2009). Neuropsicologia da depressão. In: A., Lacerda, L., Quarantini, A., Miranda-Scippa & J., Del Porto. *Depressão: Do neurônio ao funcionamento social.* (pp.253-268). Porto Alegre: Artmed.

Received 08/31/2011
Accepted 11/26/2011

Andre Goettems Bastos. Federal University of Rio Grande do Sul, Brazil
Bruna Monego Gomes. Federal University of Rio Grande do Sul, Brazil
Denise Ruschel Bandeira. Federal University of Rio Grande do Sul, Brazil
Clarissa Marceli Trentini. Federal University of Rio Grande do Sul, Brazil