Revista Interamericana de Psicología/Interamerican Journal of Psychology - 2011, Vol. 45, Num. 2, pp. 235-242

Most Frequent Errors Made by Psychology Students in the Administration and Scoring of the WISC-III

Luciana de Ávila Quevedo[1]
Vera Lúcia Marques de Figueiredo
Universidade Católica de Pelotas- Pelotas-RS- Brasil
Universidad de Murcia, Spain

Resume

El propósito de este estudio ha sido identificar los errores más comunes cometidos por los examinadores durante la administración y calificacion del WISC-III. Quinientos protocolos de investigaciones anteriores se analizaron. Los resultados mostraron un promedio de 5.11 errores por protocolo, mientras que el mayor porcentaje de errores fureon cometidos durante la calficacion. Durante la administración,el error más común fue la falta de respuestas a las consultas. En cuanto a subtests, el mayor número de errores de la administración y calificación se observó en el vocabulario (29,19%), comprensión (25,9%) y similitudes (17,78%), lo que corrobora los hallazgosliterarios, que sugieren estos como los más subtests de total.
Palabras clave: WISC-III; errores; Normalización; Administración; puntuación.

Los errores más frecuentes de los estudiantes de psicología en la administración e interpretacion del WISC-III

Abstract

The purpose of this study has been to identify the most common errors made by examiners during administration and scoring of the WISC-III. Five hundred test protocols from previous researches were analyzed. The results showed an average of 5.11 errors per protocols, while the larger percentage of errors has been verified in the scoring. During administration, the most common error was Failure to Query responses. Regarding subtests, the largest number of errors of administration and scoring was observed in Vocabulary (29.19%), Comprehension (25.9%) and Similarities (17.78%), thus corroborating literary findings, which suggest out these as the hardest subtests to score.
Keywords: WISC-III; Errors; Standardization; Administration; Scoring.

The Wechsler Scales are internationally recognized as being among the main instruments for psychological evaluation used in the field of neuropsychology, both in clinical and educational contexts. In Brazil, only in 2001 the third edition of the test for school age children (Wechsler Intelligence Scale for Children Third edition – WISC-III) was adapted by Figueiredo (2001), and this is the version still used by psychologists as the new edition is not validated in the country. Although the new WISC-IV was published in the United States in 2003, many professionals are still unwilling to abandon the previous edition due to the familiarity acquired with the techniques in use. This claim can be verified by the relatively small amount of studies identified during a long period of time after the instrument was released.

Whatever the version of the Wechsler Intelligence Scales used and the target population for which it is intended, in order to obtain results that can be interpreted with confidence, one should follow carefully all the instructions for administration and evaluation indicated in the manual. Alterations in phraseology or in the way an item is presented, as well as changes in the amount of time limit or yet any other change in the standardized instructions for the subtests may reduce the validity of the results (Wechsler, 1991). To ensure testing uniformity, every part of the manual provides technical instructions, and these ought to be sufficiently clear, precise and which the examiner should strictly conform to when performing and evaluating the test (Gil, 1994), since getting reliable results in psychological tests is largely dependent on the accuracy in the administration and correction of the instrument (Alchieri & Cruz, 2003; Pasquali, 2003).

The procedures for administration of the Wechsler Intelligence Scales involve the performing of simultaneous tasks such as recording answers, mastering the

[1] Corresponding author:
Luciana de Ávila Quevedo - Praça 20 de Setembro, 904 B, 203 Pelotas –RS, Brazil Cep: 96015-360. Phone: (053) 32253654 e-mail: lu.quevedo@bol.com.br

rules for administration, maintaining rapport with the individual under test and observing the behavior of the examinee, which can as a consequence give occasion to the examiner making mistakes.

The following procedures explain the main rules of administration and scoring, according to the test manual (Wechsler, 2002), which are the same categories analyzed in this study.

Ceiling and basal subtest criteria

At the start of the administration of each subtest, the examiner should be aware of the basal level, i.e., which item she should begin with.

Teaching the task. Several subtests provide some type of help or answer correction for the initial items. The goal is to ensure that the child understands the question or provide for additional instructions in case the examinee does not get such items right. Help should be given only when indicated by the subtest itself, and should be reported in the protocol by '&'. In such items, if the child makes a mistake, the examiner should give the right answer.

Querying Responses. Ambiguous or incomplete answers produced in Verbal Subtests should be queried. What is intended when querying is to clarify the answer initially given by the child.

Recording Answers. The answer given by the examinee in response to each item should be written down so that they can be further scrutinized after the testing session, allowing for a more detailed evaluation. This record also provides for qualitative assessment of the answers, which in turn may yield useful clinical information. Failure to record the answers, execution time, the queries and warnings in the protocol, may lead to doubtful scores. According to Cunha et al. (2000), the psychologist who does not textually record the answer, recording directly the scoring, or the one who takes only short reminders to keep only the relevant and appropriate data for awarding the score afterwards, often lose important clinical evidence, which may sometimes be subtle, but nonetheless essential to helping achieve a better understanding of the underlying emotional problems. For Kaufman (1979), recording clinical observations of behavior and verbalizations is just as important as the scoring.

Scoring. Most of the items that make up the WISC-III subtests are scored in a direct and objective way; however for Similarities, Vocabulary and Comprehension it is required that scoring be performed in a more careful manner. For such subtests, reliability between examiners was evaluated in the standardizing research for the original American test. The agreement between examiners and the reliability was found to be 0.94 for Similarities, 0.92 for Vocabulary, and 0.90 for Comprehension (Wechsler 1991). According to the author, although these subtests require a more elaborate judgment from the examiner, the scores can be consistent; however, according to Sattler (1992), scoring such subtests requires a more skilled examiner. Sattler, Andres, Squire, Wisely and Maloy (1978) developed a survey with 110 psychologists and undergraduate students who gave scores to 187 ambiguous answers in the Similarities, Vocabulary and Comprehension subtests. An 80% agreement in scoring level was reached in only 51% of the ambiguous answers for Similarities, 49% for Comprehension and 38% for Vocabulary making evident the difficulties involved in scoring such answers.

Studies have shown that examiners who use the Wechsler Scales very often incur a variety of errors in the administration and scoring of these tests. According to a review by Groth-Marnat (1999), the most frequent errors include failure to record the answers and the time of performance of the subject; failure to query when indicated, as well as inadequate querying.

Slate and Jones (1990), when studying the mistakes made during the administration of the WAIS-R, found an average of 8.8 errors per protocol in a sample of 180 protocols administered by 26 students. When these protocols were corrected, the results for the IQs changed in 81% of the protocols (62.2% was lower than that claimed by the students whereas 17.8% were higher). The most frequent type of error was the absence of records for verbal responses, with scoring being the second. Another error highlighted was inappropriate querying, which happened more often in items of the Vocabulary, Comprehension and Similarities subtests.

Slate, Jones, Coulter, and Covert, (1992) in another study in which they analyzed 56 WISC-R protocols administered by 9 experienced professionals who had been trained in the administration and scoring of the test, observed that examiners committed errors in all protocols. When the absence of answer records, not reporting the scores, and not recording time of performance were considered as errors, the average was 38.4 errors per protocol (SD = 29.3), while the average number of errors made by examiner ranged from 16.2 to 90.1. The subtests presenting the largest number of errors, including absence of records were: Picture Completion, Picture Arrangement and Vocabulary. When absence of records for the answers was not computed as an error, the average fell to 8.7 errors per protocol (SD = 4.6), with the average number of errors per examiner ranging from 4.7 to 15.1. The subtests with the largest number of errors were Vocabulary, Comprehension and Similarities. Also, according to the authors, the most often found error is absence of records for the answers, an error committed by all examiners (29.5 errors per protocol). The second error most often

found was in the scoring of the answers. The examiners gave 5.8 times higher scores than lower scores when compared to the appropriate score. Another common mistake was failure to query, which happened 4 times more often than inappropriate querying.

In a study by Alfonso, Johnson, Patinella, and Rader (1998), 60 WISC-III protocols were administered and scored by 15 undergraduate students. Each participant was in charge of administering and scoring 4 protocols. No observations were made during administration, and the errors were found through the records in the WISC-III protocols. The examiners were instructed to read and study the manual before administering the test, and to re-read about the most common mistakes in the Wechsler Scales. After each administration, students received a verbal and written feedback by the instructor, regarding his/her performance, before starting with the next protocol. As a result, the authors found 468 errors, with approximately 7.8 errors per protocol, and 31.2 errors per student. Comparing the average number of errors of the first protocol to the fourth protocol yielded a difference of 14.4 down to 5.4 errors per protocol. The subtests which presented errors more often were Comprehension, Similarities, Symbol Search and Digit Span. All protocols had errors, and the most common were: failure to query, absence of records for the verbal responses and clerical errors. After reviewing the mistakes in the protocols, 70% of the IQs fell in a different range.

Belk, LoBello, Ray, and Zach (2002) reviewed 100 WISC-III protocols, which had been administered and scored by 21 undergraduate students. The researchers found an average of 45.2 errors per protocol when considering the absence of records for the answers as a mistake. When disregarded as an error, the average went down to 10.9 errors per protocol. The errors were classified into the following categories: administration/scoring and clerical errors. Failure to query was the most evident error in administration. The examiners failed to query the answers 7.6 times more than actually querying inappropriately. Concerning scores, students were 2.3 times more likely to overestimate the scores of items that make up the Verbal IQ composite, therefore yielding higher IQs in 46% of the cases. The Comprehension, Vocabulary and Similarities subtests were the ones to present the highest number of errors. The most common clerical error was erroneous transcription of weighted scores from the manual tables and the inappropriate inclusion of optional subtests for IQ calculation. These were low frequency errors, although they are fundamental for the result of the Full Scale IQ (FIQ).

Loe, Kadlubek and Marks (2007) conducted a study with 51 WISC-IV protocols, administered and scored by 17 students in the 1st year of their graduation course. The error was present in 98% of the test protocols, with the average being 25.8 errors per protocol. The most common error happened in administering the test, appearing in 41% of the total. Errors in score computation appeared in 37%, while in 22% recording the answer was detected. The most often found types of errors were failure to query, absence of recording the answer and scoring. The Comprehension, Block Design, Vocabulary and Similarities subtests presented the higher error rate. As for the IQs, 35% had to be changed, 75% of which had to be corrected up to a value greater than what had been indicated.

According to Pasquali (2003), error is an ever present element in any measure, and its presence constitutes a serious threat to scientific decision-making, and that is the reason why it is important to have means to neutralize or diminish its effects, or at least to find out what its magnitude is with the best approximation possible. Within this perspective, the present study sought to identify the most frequent errors with the WISC-III in a different cultural environment than the one in which the original test was standardized. All studies found in the literature show that, regardless the version of the Wechsler Intelligence Scales used and whatever the target population is (school age children or adults), errors occur both in the administration and in the scoring, with students or professionals using the scales.

Method

This study has been undertaken on a documental basis, where 500 WISC-III protocols, drawn from the research database developed by the second author of this article, were analyzed. The instrument was administered by 10 research assistants who administered at least 20 protocols each. The undergraduates were female participants, all enrolled in a Psychology graduation course, having finished five semesters of studying.

Specific training was given to the students three times a week during one month, with both theoretical and practical content, the latter consisting of simulations of test administration. In order to successfully complete the training the student should administer the WISC-III test to two children in a supervised environment and to be approved in a written test before starting collecting data.

All protocols were corrected by the examiner herself and later reviewed by two other examiners. Following the correction, the examiner provided feedback on all the errors identified, both administration errors and scoring errors. The latter were corrected before performing the conversion of the weighted scores yielding the IQs.

For the purpose of this study, the selected protocols were sorted in different categories according to the following types of error:

Administration Errors.

a) *Basal error:* the examiner did not start the administration by the most appropriate item for the subject's age, as well as not returning to the previous items.

b) *Ceiling error:* the examiner stopped the administration of the subtest before or after the criterion established in the manual.

c) *Help error:* the examiner did not record help in items indicated by the '&' symbol; intervention in inappropriate items; wrong warnings in items for the Picture Completion subtest and incorrect aid recordings.

d) *Query errors:* the examiner did not query the answer or did it inadequately; or the examiner did not query Picture Completion items when answers were doubtful; the examiner did not ask for a second argument on items marked with an asterisk in the Comprehension subtest.

e) *Record error:* the examiner did not record on the test protocol, data regarding identification and answers given by the examinee, items' execution time, scoring, as well as any other relevant in formation.

Correction Errors. These refer to incorrect scores for items and miscalculations when performing the summation of subtests scores.

Results

The average number of errors both in administration as well as in scoring was 5.14 per protocol (SD=3.81). Scoring was responsible for 59.5% of the errors identified in the protocols.

The errors in the correction of the answers are presented in Table 1, including both the scoring of the items (the most frequent error) and the summation of the scores in the subtests. The Vocabulary, Comprehension and Similarities subtests presented, respectively, the highest number of scoring errors, while Symbol Search and Digit Span the lowest. Taking the total of the scoring errors, 85.94% occurred in the verbal subtests.

Table 1
Correction Errors per Subtest

Verbal subtests	Errors	Order	Performance subtests	Errors	Order
Information	67 (4,38%)	4	Picture Completion	44 (2,88%)	5
Similarities	218 (14,24%)	3	Coding	40 (2,61%)	6
Arithmetic	34 (2,22%)	8	Picture Arrangement	38 (2,48%)	7
Vocabulary	504 (32,94%)	1	Block Design	32 (2,1%)	9
Comprehension	479 (31,31%)	2	Object Assembly	34 (2,22%)	8
Digit Span	13 (0,85%)	11	Symbol Search	27 (1,76%)	10
Verbal Total	1315 (85,94%)		Performance Total	215 (14,05%)	

According to Tables 2 and 3, when considering only the observed errors in the administration of the Verbal Scale (n = 794), the majority referred to the Vocabulary subtest, mainly regarding Querying - category in which were included the errors "Failure to Query" (n = 473) , "Not asking for a 2nd argument in items reported in Comprehension" (n = 41) and "Inadequate Querying" (n = 28). The second most frequent error was Help, either doing it inadequately (n = 65) or omitting it (n = 111). The least amount of errors appeared in basal errors.

Taking the administration errors in the Performance Scale (n = 248), the most frequent error was of the "Record error" type, while Ceiling was the least frequent. When considering the administration errors observed in the test as a whole (n = 1042), 76.2% occurred in the Verbal Subtests. Among the Performance subtests, Picture Completion was particularly noteworthy for the presence of query errors or warning errors. Some categories of errors were not considered to be appropriate in some subtests and are reported in Tables 2 and 3 with "does not apply" note. In Vocabulary, for example, there is no basal error, since regardless the age of the subject, the subtest should always start with item 1.

Table 2
Administration Errors for Verbal Subtests

Subtests	Basal	Ceiling	Help (&)	Query (Q)	Records
Information	26 (2,5%)	10 (0.96%)	1 (0,1%)	16 (1,54%)	3 (0,29%)
Similarities	NA	9 (0,86%)	126 (12,09%)	95 (9,11%)	8 (0,77%)
Arithmetic	9 (0,86%)	12 (1,15%)	0	NA	21 (2,02%)
Vocabulary	NA	23 (2,21)	4 (0,38%)	214 (20,53%)	8 (0,77%)
Comprehension	NA	6 (0,58)	20 (1,93%)	157 (15,07%)	2 (0,2%)
Digit Span	NA	11 (1.06)	NA	NA	13 (1,25%)
Total	35 (3,36%)	71 (6,82%)	151 (14,5%)	482 (46,25%)	55 (5,3%)

Note: NA= Not applicable

Table 3
Administration Errors for Performance Subtests

Subtests	Basal	Ceiling	Help (&)	Query (Q)	Records
Picture Completion	23 (2,2%)	4 (0,38%)	24 (2.3%)	60 (5,76%)	15 (1,43%)
Coding	NA	NA	NA	NA	14 (1,34%)
Picture Arrangement	NA	2 (0,19%)	NA	NA	17 (1,63%)
Block Design	8 (0,78%)	14 (1,34%)	NA	NA	24 (2,3%)
Object Assembly	NA	NA	1 (0,1%)	NA	20 (1,93%)
Symbol Search	NA	NA	NA	NA	22 (2,1%)
Total	31 (2,98%)	20 (1,91%)	25 (2,4%)	60 (5,76%)	112 (10,73%)

Note: NA= Not applicable

Table 4 shows the items that in each subtest appeared with higher error rate. For the Picture Arrangement subtest, two items are shown since they presented the same error rate. In the Picture Completion subtest, item 22, when getting answers for non-key elements in the picture, there was no querying asking which would be the most relevant. In Information, item 6 and in Comprehension, item 10, scoring error was observed. In the Similarities subtest, the most common error was Failure to Help in item 2. In subtests Picture Arrangement and Object Assembly, errors were found relating to scoring. In Arithmetic, item 18, and Block Design, item 3, Failure to Record Answer and/or execution time has been observed. In the Vocabulary subtest, the main problem was noted in item 7, concerning Failure to Query. In subtests Coding, Symbols Search and Digits Span the analysis was not performed.

Table 4
Subtest item that presented the highest number of errors

Subtests	Most Problematic Item	Type of Error
Picture Completion	22	Query
Information	6	Scoring
Similarities	2	Failure to help
Picture Arrangement	9 and 14	Scoring
Arithmetic	18	Inadequate recording
Block Design	3	Inadequate recording
Vocabulary	7	Failure to query
Object Assembly	5	Scoring
Comprehension	10	Scoring

Discussion

Results showed that the number of errors per protocol was lower than what is found in the literature, regardless the edition of the Wechsler Intelligence Scale considered (Alfonso et al., 1998; Belk et al., 2002; Loe et al., 2007; Slate & Jones, 1990; Slate et al., 1992). In spite of the errors made, it seems that the training had a certain impact in minimizing the failures in administration and scoring; the examiners also had familiarity and showed mastery of the material although they were undergraduate students. Similar results were reported by Ramos, Alfonso and Schermerhorn (2009), who found that adequate measures in training programs can minimize the number of errors made by the examiners in the Woodcock-Johnson III and other cognitive test batteries.

Scoring errors occurred more frequently than administration errors, agreeing with the results obtained by Slate and Jones (1990), Slate et al. (1992) and Alfonso et al. (1998). However, results differ from those found by Belk et al. (2002) and Loe et al. (2007) that reported administration errors as the most frequent, probably because the authors had included scoring errors in the category "administration".

In the Vocabulary, Comprehension and Similarities subtests, scoring presented an added difficulty, wherein the results actually oppose those of Wechsler (1991) which characterized the scoring of items in such subtests as an objective task. However, data corroborate those of Sattler et al. (1978) for the WISC-R; Slate and Jones (1990) for the WAIS-R; Slate et al. (1992) for the WISC-R; Belk et al. (2002) for the WISC-III and Loe et al. (2007) for the WISC-IV. Actually, scoring errors may be related to the lack of good parameters in the manual to guide scoring, since there are only very few examples of standard answers. When a different answer is given which is not in the manual, its acceptability becomes doubtful, and scoring involves subjectivity. According to the results, we conclude that the Verbal subtests comprising the Verbal Comprehension factor, and hence the VIQ, are the most susceptible to inadequate scoring and, probably for this same reason Belk et al. (2002) observed scores in overestimated VIQ.

Considering the total number of administration errors, both in Verbal subtests and Performance subtests, the most common error was query, also found by Slate and Jones (1990), Slate et al. (1992), Alfonso et al. (1998) and Groth-Marnat (1999), Belk et al. (2002) and Loe et al. (2007) with all the different versions of the Wechsler Scales. Failing to query may be related to the examiner's lack of knowledge about the expected answers, taking as correct any verbalization related to the item. Not explaining poor and dubious answers will, as a consequence, lead to imprecise scores and eventually to erroneous diagnoses as well.

Among all Verbal subtests, Failure to Query happened more often in the Vocabulary subtest, which was also the same subtest that examiners had more difficulty scoring, suggesting that this is the most complex test for examiners. Picture Completion was the Performance subtest which presented the highest number of query errors, result also found by Slate et al. (1992). This error is composed of two failures. The first refers to not clarifying ambiguous answers such as in Item, in which case the examiner should always try to clarify the answer by asking "where". The second failure relates to the examiner not giving a warning when receiving non-essential answers or answers which refer to parts that are not present in the picture. Both errors can be seen to stem from a weak command of the test administration techniques.

Failure to help was another frequent error during administration, which could be overestimated, since even if the examiner provided helped to incorrect answers during the initial items, still the procedure was considered omission if the examiner did not record it on the protocol. Not giving help during initial items is more closely related to the examiner's lack of attention than to weak command of the administration techniques, since not only the manual gives instructions, but also there is a "&" symbol in the test protocol indicating the place for such intervention.

As for items with higher error rate, in the Picture Completion subtest "soap and water" were very common verbalizations for item 22. Since these are considered non-essential answers and because no previous warning had been given, the examiner should ask: *but what is the most important part that is missing?* - Intervention was not performed in most of the cases. Despite the objectivity in scoring for the Information subtest (1 or 0 point), many different answers were observed in item 6 which are not given in the manual, such as "currency and paper", giving the examiner opportunity for querying. In the Comprehension subtest the scoring errors in item 10, may be related to the difficulty of matching the answers given by the examinees to the general ideas in the manual. The errors relating to item 7 of the Vocabulary subtest occurred often when scoring answers of the kind "you can ride/one rides on it". In the manual, similar answers appear both in 2-point as well as in 1-point examples, however they differ in content complementation in the first case, which could simply remain unnoticed by the examiners. No reasons were found other than not paying enough attention that could warrant the scoring errors found in subtests Picture Arrangement and Object Assembly, since scores in these subtests are only related to the correct

performance of the item and its execution time. The same reason could be linked to the failure of recording time, scoring and answers in subtests Arithmetic and Block Design. The research works found in the literature, which make reference to the mistakes made by the examiners during WISC-III testing did not present any results concerning analysis of test items, rendering impossible any comparison between the results.

From this work we conclude that no matter how efficient a training for an examiner is, one cannot do without the monitoring from a specialist who can identify the errors, providing feedback during the first few sessions the test is administered, allowing those rules which are not fully automated initially to become gradually mastered. Considering that the answers given by the examinees are often different from those presented in the manual, it is important to have available a more comprehensive list of model answers, especially for subtests Similarities, Vocabulary and Comprehension. According to all studies found in the literature, such subtests are reported as the hardest to score, regardless the edition of the Wechsler Scales (Alfonso et al., 1998; Belk et al., 2002; Loe et al., 2007; Slate & Jones, 1990; Slate et al., 1992). Another suggestion is to consult the other pairs when one is unsure about the scoring.

One of the strengths of this study is the larger number of protocols analyzed, in contrast to those reported in similar studies found in the literature. Also, some limitations may be mentioned. The first limitation was the lack of a one-way mirror room to use for observation purposes during test administration. Because of this limitation, errors analyzed according to the records of the examiners could not be verified. Secondly, in the case of ceiling error, it was not recorded whether administration was stopped before or after the established criterion, something that could lead to different implications.

In this study we identified the most common errors in the administration and scoring of the WISC-III. The results presented here will enable users of the Wechsler Intelligence Scales, both psychology students and professionals, to place more emphasis on such criteria at the time of their preparation to use the scales. Furthermore, these results will also prove helpful to those who teach and give training on these instruments. Although we based this study on the WISC-III, these results can certainly be applied to the newer versions of the scales. The results found in Brazil agree with those reported by different investigators.

References

Alchieri, J.C. & Cruz, R.M. (2003). *Avaliação Psicológica: Conceitos, Métodos e Instrumentos* (pp.57-68). São Paulo: Casa do Psicólogo.

Alfonso, V.C., Johnson, A., Patinella, L., & Rader, D.E. (1998). Common WISC-III Examiner Errors: Evidence from Graduate Students in Training. *Psychology in the Schools, 35*(2), 119-125.

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 Assessment, 20,* 290-300.

Cunha, J.A. (2000). Escalas Wechsler. In Cunha, J.A. & Colaboradores. Psicodiagnóstico V. Porto Alegre: *Artes Médicas* (pp.529-615).

Figueiredo, V. L. M. (2001). *A Brazilian Adaptation of the WISC-III Intelligence Test*. 230p.

Doctoral Thesis. Institute of Psychology, University of Brasília.

Gil, A.C. (1994). Os testes In Gil, A.C. *Métodos e Técnicas de Pesquisa Social*. (pp.146-156). São Paulo: Atlas (pp.146-156).

Groth – Marnat, G. (1999). *Handbook of Psychological Assessment*. New York: Wiley & Sons.

Kaufman, A. (1979). *Intelligent Testing with the WISC-R*. New York: John Wiley & Sons, Inc.

Loe, S.A.; Kadlubek, R.M.; & Marks, W.J. (2007). Administration and Scoring Errors on the WISC-IV Among Graduate Student Examiners. *Journal of Psychoeducational Assessment, 25*(3), 237-247.

Pasquali, L.D. (2003). *Psicometria: Teoria dos Testes na Psicologia e na Educação* (pp.23-51). Petrópolis: Vozes.

Ramos, E.; Alfonso, V. & Schermerhorn, S.M. B (2009). Graduate students' administrationnd scoring errors on the Woodcock-Johnson III Tests of Cognitive Abilities. *Psychology in the Schools, Vol. 46*(7), 651-657.

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, 15*(4), 486 – 489.

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

Slate,J., & Jones,C. (1990). Identifying Students' errors in administering the WAIS-R. *Psychology in the School, 27,* 83-87.

Slate, J. R., Jones, C. H., Coulter, C., & Covert, T.L. (1992). Practitioner's Administration and Scoring of the WISC-R: Evidence that we do err. *Journal of Psychology, 30,* 77-82.

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

Wechsler, D. (2002) *Escalas de Inteligência Wechsler para Crianças Terceira Edição: Manual*. Adaptação e Padronização Brasileira, Vera Lúcia Marques de Figueiredo. São Paulo: Casa do Psicólogo.

Received 20/07/2010
Accepted 03/12/2011

Luciana de Ávila Quevedo. Universidade Católica de Pelotas- Pelotas-RS- Brasil
Vera Lúcia Marques de Figueiredo. Universidade Católica de Pelotas- Pelotas-RS- Brasil