Graduate student WAIS III evaluation reports: Are they thorough?

David B. Moore
Pacific University

Recommended Citation
Graduate student WAIS III evaluation reports: Are they thorough?

Abstract
This study examined 38 WAIS-III interpretative evaluation reports, based on a fictitious client and written by graduate students as part of an assessment course, to determine what information was included in the reports based on the results of the protocol scores. In order to ensure consistent results for the investigation, a computer-based test interpretation (CBTI) program following the steps Kaufman and Lichtenberger (2002) outlined for analysis and interpretation of WAIS-III results was developed for this study. The hypotheses that major findings would be included in the reports was mostly confirmed. The exception was intersubtest scatter measured in 33 reports but included as part of an interpretation in only 6. The hypotheses that indices comparisons and strengths and weaknesses in subtests would be neglected in reports was confirmed. The final hypothesis regarding computational errors was not confirmed. The errors were fewer than expected and below results contained within other research. The findings of this research has implications for instructors of intelligence assessment courses.

Degree Type
Dissertation

Degree Name
Doctor of Psychology (PsyD)

Committee Chair
Lisa Christiansen, Psy.D

Second Advisor
Susan Tinsley Li, Ph.D.

Keywords
Wechsler, psychological assessment, intelligence measures, computer-based test, psychological report, test interpretation

Subject Categories
Psychiatry and Psychology | Social and Behavioral Sciences

Comments
Library Use: LIH

This dissertation is available at CommonKnowledge: https://commons.pacificu.edu/spp/239
Copyright and terms of use

If you have downloaded this document directly from the web or from CommonKnowledge, see the “Rights” section on the previous page for the terms of use.

If you have received this document through an interlibrary loan/document delivery service, the following terms of use apply:

Copyright in this work is held by the author(s). You may download or print any portion of this document for personal use only, or for any use that is allowed by fair use (Title 17, §107 U.S.C.). Except for personal or fair use, you or your borrowing library may not reproduce, remix, republish, post, transmit, or distribute this document, or any portion thereof, without the permission of the copyright owner. [Note: If this document is licensed under a Creative Commons license (see “Rights” on the previous page) which allows broader usage rights, your use is governed by the terms of that license.]

Inquiries regarding further use of these materials should be addressed to: CommonKnowledge Rights, Pacific University Library, 2043 College Way, Forest Grove, OR 97116, (503) 352-7209. Email inquiries may be directed to: copyright@pacificu.edu
ABSTRACT

This study examined 38 WAIS-III interpretative evaluation reports, based on a fictitious client and written by graduate students as part of an assessment course, to determine what information was included in the reports based on the results of the protocol scores. In order to ensure consistent results for the investigation, a computer-based test interpretation (CBTI) program following the steps Kaufman and Lichtenberger (2002) outlined for analysis and interpretation of WAIS-III results was developed for this study. The hypotheses that major findings would be included in the reports was mostly confirmed. The exception was intersubtest scatter measured in 33 reports but included as part of an interpretation in only 6. The hypotheses that indices comparisons and strengths and weaknesses in subtests would be neglected in reports was confirmed. The final hypothesis regarding computational errors was not confirmed. The errors were fewer than expected and below results contained within other research. The findings of this research has implications for instructors of intelligence assessment courses.

KEYWORDS: Wechsler, psychological assessment, intelligence measures, computer-based test, psychological report, test interpretation
ACKNOWLEDGMENTS

First, I would like to thank my dissertation committee, Dr. Christiansen and Dr. Tinsley Li, for all their support, advice, guidance, and encouragement on this study and in general. In my first assessment course Dr. Christiansen initiated the questions and search for answers leading to development of the CBTI and research in this study. I would also like to thank doctor-to-be Brent Ryder for his help and continual encouragement for me to get my dissertation completed. I would also like to thank my family for their constant love and assistance during the years of schooling and changing of careers. My brothers, sister, and mother have been a great support and helped me when I was facing difficult times. My children and grandchildren gave up years of their time, camping trips, vacations, and activities with me to give me quiet time to complete my degrees. Last, I would like to thank my wife for all the love and support she has given me through this long journey.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>REVIEW OF THE LITERATURE</td>
<td></td>
</tr>
<tr>
<td>Errors in Wechsler Scale Administration and Scoring</td>
<td>4</td>
</tr>
<tr>
<td>Accuracy and Completeness of Assessment Reports</td>
<td>9</td>
</tr>
<tr>
<td>Computer Based Test Interpretation</td>
<td>18</td>
</tr>
<tr>
<td>PURPOSE OF THE CURRENT STUDY</td>
<td>28</td>
</tr>
<tr>
<td>METHODS</td>
<td></td>
</tr>
<tr>
<td>Participants</td>
<td>31</td>
</tr>
<tr>
<td>Wechsler Adult Intelligence Scale - Third Edition (WAIS-III)</td>
<td>32</td>
</tr>
<tr>
<td>Computer Based Test Interpretation (CBTI) for the WAIS-III</td>
<td>34</td>
</tr>
<tr>
<td>Procedure</td>
<td>37</td>
</tr>
<tr>
<td>RESULTS</td>
<td>40</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>49</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>62</td>
</tr>
<tr>
<td>APPENDICES</td>
<td></td>
</tr>
<tr>
<td>A. CBTI Example</td>
<td>67</td>
</tr>
<tr>
<td>B. Check Sheet</td>
<td>72</td>
</tr>
<tr>
<td>C. Informed Consent</td>
<td>73</td>
</tr>
</tbody>
</table>
D. Differential Interpretation Criteria
LIST OF TABLES

TABLE 1: Constructs Interpreted .........................................................................................41

TABLE 2: Constructs Interpreted Results .............................................................................44

TABLE 3: IQ And Index Score Review ...............................................................................46

TABLE 4: IQ And Index Score Review With Computational Errors Removed ..................46

TABLE 5: IQ and Indices Comparison Differences...............................................................47

TABLE 6: IQ and Indices Comparison Differences With Computational Errors Removed ....................................................................................................47
INTRODUCTION

The Wechsler intelligence scales have been the one of most widely used assessment tools for all populations in the past decade (Kaufman & Lichtenberger, 2002). The scales have spawned much research and been proven valuable for multiple uses. Research has identified uses for them in diverse settings ranging from school assessment (Alfonso, Johnson, Patinella, & Rader, 1998) to neuropsychology (Sinnet & Holen, 1995). However, errors in the administration and scoring of the Wechsler intelligence scales for adults and the scales for children (all versions) have been found in multiple studies (Alfonso et al., 1998; Kaufman & Lichtenberger, 2002; Loe, Kadlubek, & Marks, 2007; Sattler, Winget, & Roth, 1969; Slate, 1993; Slate & Jones, 1990). The clinician’s use of the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III) and the evaluation report that is written based on that assessment can have a great impact on the life of a client. Thus it is important to ensure that protocols are being scored and interpreted correctly.

An examiner’s evaluation report is the culmination of an assessment. How the information is compiled and how thorough the evaluation report is creates an impact for the client. Therefore, it seems important that the best practices possible should be employed to ensure that the client’s interests are protected. As data exists regarding errors in administration and scoring, one may question whether data also exists regarding errors in interpretation and thoroughness of evaluation reports. A review of available studies within multiple databases found no studies that specifically addressed the thoroughness of WAIS-III interpretive evaluation reports. With the potential impact that the test could
have on a client, the lack of cited research on thoroughness of evaluation reports was noticeable. Without research illuminating possible errors within WAIS-III evaluation reports there is no standard against which to inform the assessor of what areas of interpretation within evaluative reports may be prone to error.

One method that could be used to inform research would include the use of a computer-based test interpretation (CBTI) program to inspect WAIS-III interpretive evaluation reports for accuracy. In an attempt to address the lack of research on the thoroughness of evaluation reports, the primary investigator of this study constructed a CBTI (Moore & Christiansen, 2007) that followed the steps Kaufman and Lichtenberger (2002) outlined for analysis and interpretation of WAIS-III results. Based on the steps of the analysis, the CBTI generated a series of statements for each condition of the indices and subtests of the WAIS-III. This followed the best practices model laid out by Dawes (2002). His recommendation for the use of statistical prediction rules in order to increase clinical accuracy could be followed in a manner that still allowed for clinician judgment. If CBTIs were found to be equal to or better than expert clinician judgment then it should also hold that a properly constructed and validated CBTI would be an effective method of examining the accuracy and thoroughness of WAIS-III interpretative evaluation reports.

The present study intends to examine the usefulness of a CBTI in determining the accuracy and completeness of student generated WAIS-III reports. The examination of those reports will focus on what information graduate students include within the reports. The author hypothesizes that all possible variables that could be calculated will not be included in the interpretative evaluation reports. However, it is hypothesized the three primary findings of FSIQ, VIQ, and PIQ will be reported as will
major uncommon findings such as an abnormal split between the VIQ and PIQ and scatter in an index. Specific areas considered to be prone to a presentation of incomplete findings in interpretation are the indices comparisons and strengths and weaknesses within the subtests. Finally, it is hypothesized that the students will commit computational errors resulting in incorrect interpretation. Implications for instruction will be addressed.
REVIEW OF THE LITERATURE

Errors in Wechsler Scale Administration and Scoring

The Wechsler scales have proven to be multifaceted and useful instruments. However, several research findings have given cause for concern about the tests. Errors in the administration and scoring of the Wechsler intelligence scales for adults and the scales for children (all versions) have been found in multiple studies (Alfonso et al., 1998; Sattler, Winget, & Roth, 1969; Slate, 1993; Slate & Jones, 1990). Kaufman and Lichtenberger (2002) reviewed and presented a summary of the research findings. They found that practitioners and graduate students were both likely to make some errors in administering and scoring the Wechsler scales. They recommended improved methods of training and concentration on the details of administration and scoring to decrease the amount of errors made.

Kaufman and Lichtenberger (2002) cited multiple research studies and their own experience to establish that multiple administration and scoring errors are often committed on the Wechsler scales. Research revealing a consistent pattern of errors made during the administration and scoring of any test adds to systematic error and decreases the reliability of the test. Graduate students are by definition still in training and thus may be susceptible to a greater error rate than practicing clinicians. As will be evident in the forthcoming review, information revealed within the studies did not uphold that premise.

Slate and Jones (1990) examined both the Wechsler scale for adults and the scale for children. The authors wanted to correct deficiencies that they had identified in previous
studies on errors in administration and scoring of the Wechsler scales. The goals of the study were identified as: a) Determining the most frequent types of examiner errors made by graduate students and the items on which they were most likely to occur and b) To discover how much improvement the students made over 5 to 10 practice administrations. Twenty-six masters level students were assigned to complete a combination of WAIS-R and WISC-R practice protocols, for a total of 15 Wechsler scales. A specially trained graduate assistant (GA) scored the protocols for errors. Two hundred seventeen protocols were examined and all contained errors. The most common errors were: a) Failure to record answers, b) Incorrect point assignment (five times more likely to give too many points for an answer than too few), and c) Failure to query. The most frequent errors occurred on the vocabulary, comprehension, and similarities subscales.

The authors stated that accuracy in administration and scoring improved only slightly over 5 to 10 administrations. They concluded that traditional methods of teaching were insufficient to resolve the error rate. They recommended that criteria for awarding of points be clarified, stronger instructor emphasis be placed on reward and punishment for mechanical error, and focusing instructor efforts on the most difficult to score areas of the Wechsler scales. The senior author offered to supply, upon request, rules that were developed to clarify response categories.

Alfonso, Johnson, Patinella, and Rader (1998) studied graduate students, who were still in training on the administration of the Wechsler Intelligence Scale for Children-Third Edition (WISC-III), to examine if the problems cited within other studies of Wechsler scale administration and scoring were similar to those in the WISC-III. Fifteen
school psychology graduate students, one doctoral and 14 specialists, who were enrolled in a cognitive assessment course, took part in the investigation. Each student completed four administrations of the WISC-III. The authors used a checklist and examined each protocol for errors. The results showed the most common errors were: a) Failure to query, b) Failure to record responses verbatim, and c) Giving a wrong FSIQ (similar to the errors reported by Slate & Jones, 1990). The error rate declined from the first administration to the last. The authors concluded that the improvement was a result of the method of teaching and that similar methods may show promise in increasing accuracy. They failed to identify what the method was but did state that information on the method would be supplied on request.

The findings by Slate and Jones (1990) and Alfonso et al. (1998) were in line with deficiencies found in reports by practitioners from the first editions of the Wechsler scales to the most recent. Sattler, Winget, and Roth (1969) first investigated the difficulties of scoring the subscales that would continue to be found by future researchers (e.g., Slate & Jones, 1990) as being most prone to error: comprehension, vocabulary, and similarities. They proposed that due to the lack of consistency in judgment used to score specific WAIS and WISC subtests (comprehension, vocabulary, and similarities) responses would not receive uniform agreement as to scoring. They chose to use experienced examiners to test their hypothesis.

Two investigations were done within the study. One was completed on the WAIS and a year later another was done on the WISC. Ambiguous responses to 98 items for the WAIS and 79 for the WISC, either selected from real protocols or constructed by the
authors, were collected for evaluation. Eight experienced doctoral level clinical psychologists were recruited to score the items. Interrater agreement on scoring was then tabulated. Only three of eight psychologists agreed on 100% of the scoring on the WAIS and again on the WISC. Complete agreement was reached on 12% of the WAIS and 8% of the WISC scoring by the examiners. The authors contended that personal judgment was a large factor in the scoring and concluded that interrater disagreement appeared to contribute to test unreliability. They recommended that better scoring standards be developed to increase agreement among scorers.

More recently, Slate (1993) investigated errors made by practitioners. The author selected 50 protocols at random from a metropolitan school district in a Southern state. Eight licensed and certified professionals (minimum training was a masters degree) with an average of eight years experience in intelligence testing completed all of the protocols. A specially trained GA examined the protocols. There were errors on all 50 protocols examined. Failure to record an answer was the most common error followed by failure to query, and assigning too many points for an answer. The Full Scale IQ (FSIQ) was changed on 54% of the protocols. Of the changed FSIQs 88% were higher than reported and 12% lower.

Slate (1993) concluded that the study indicated even practitioners with years of experience made frequent errors. The author noted that there was an unequal distribution of protocols to examiners in the study, ranging from 2 to 14 protocols per examiner. However, the same errors that were cited by Slate and Jones (1990) and Alfonso et al. (1998) were again noted within this study.
Although the cited studies may have some flaws, the agreed conclusion is that the Wechsler scales are subject to human error. The most common errors in the administration and scoring of the scales are: a) Failure to record answers verbatim, b) Failure to query, and c) Incorrect point assignment. Experience is not a guarantee of exclusion from those errors. An emphasis on correct administration and scoring of the Wechsler scales should lead to better interpretation of the scales.

**Accuracy and Completeness of Assessment Reports**

Despite the wealth of information on administration and scoring errors for the Wechsler scales, there is no published information about the accuracy of interpretation or thoroughness of evaluation report content. The American Psychological Association (APA) *Ethical Principles of Psychologists and Code Of Conduct* (2002) section 9.10 states, “Regardless of whether the scoring and interpretation are done by psychologists, by employees or assistants, or by automated or other outside services, psychologists take reasonable steps to ensure that explanations of results are given to the individual or designated representative unless the nature of the relationship precludes provision of an explanation of results” (p14). A thorough WAIS-III report of test results would include an explanation of the Full Scale IQ, Verbal IQ, Performance IQ, the interrelationships between them, confidence levels, and if they are unitary constructs. It would also include an explanation of any significant findings and relative strengths or weaknesses of the indices, their interrelationships, and the subtests. Although no information specific to the WAIS exists, research regarding psychological testing reports in general indicates several areas that may be in need of attention and may not be meeting the standards indicated by the ethics code.
The content of neuropsychological reports was examined by Donders (2001a). He discovered that no previous surveys had investigated which variables affected neuropsychological report organization and the material contained in the reports. He surveyed members of the APA Division 40 (clinical neuropsychology). The survey consisted of a three page questionnaire on the content and format of neuropsychological assessment reports. Of the 750 surveys sent out 414 usable surveys were returned (55.2%). The author limited reporting to items garnering at least 10% of ‘never’ responses and 10% of the ‘always’ responses (choices were: never, occasionally, routinely, always) in order to examine items of greatest variability. Population, practice, and types of patients seen (i.e. neurological, psychiatric, forensic) and demographics of responders were discussed. Participants averaged over 15 years post graduation, over 19% held an American Board of Professional Psychology (ABPP) diplomate in Clinical Neuropsychology, and 56% were involved in the supervision of predoctoral interns or postdoctoral residents.

The study began by examining the frequency of specific elements of background information included in reports. Items ranked high for ‘always’ included items such as age (99.76%), gender (96.86%), education (92.51%), and prior neurological history (90.34%). Items ranked low for ‘always’ were those such as prior physical/sexual abuse (15.7%), prior criminal conviction (13.53%), and prior financial compensation seeking (12.32%). Reporting prior financial compensation seeking was the area of greatest variance among responders. Whereas over 12% always reported this information, almost 18% never did. This information was noted to be of greatest interest in forensic cases. Regarding specific report content about assessment results, items ranked high for
‘always’ included those such as which tests were administered (87.44%), behavioral observations (86.96%), and descriptive terms for results (82.37%). Items ranked low for ‘always’ included those such as summary data (44.2%) and nature of norms used (23.19%). Of special interest to the author was the finding that ABPP responders were over twice as likely as other neuropsychologists to never include normative information about the tests. In addition, neuropsychologists whose practice was primarily serving pediatric clients were over twice as likely to include normative information about the tests compared to those whose practice was primarily serving adult or geriatric clients. The author noted research indicating controversy regarding the need for age-based norms for children and concerns about the current age-corrections for some tests. He used this to underscore the necessity for pediatric specialists to specify norms used as a possible reason for the difference.

Donders (2001a) concluded by noting some consistency in reporting background information and wide disparities in inclusion of specific content. He found no difference in report writing for those who provided supervision of predoctoral interns or postdoctoral residents as compared to those who didn’t. He indicated this was good because the supervisees were “exposed to procedures that are similar to those most prevalent in the field at large” (p 146). He reviewed the major findings and noted pediatric specialists tended to write different reports. He laid causation for this on a poor state of available norms for children and inconsistency in best practices for norms based on age versus education level or other comparative group. In doing so he appeared to excuse the exclusion of normative information from adult reports rather than praising the greater information placed in the pediatric reports.
In a follow-up article, Donders (2001b) reported on other aspects of the survey including test data, narrative organization, recommendations, and report distribution. He found that ABPP members (20%) were twice as likely as other neuropsychologists to never include numeric scores in reports. Percentile scores were routinely or always included in 74.66% of reports and standard scores were routinely or always included in 75.75%. Age equivalent scores were never included in 12.81% of reports with pediatric specialists never including age equivalent scores 31.48% of the time. He explained this finding as consistent with research in developmental literature reporting problems with that type of score. A discussion of every test administered was routinely or always included in 60.87% of reports and narrative descriptions of tasks in lay person terms were routinely or always included in 59.91%. The author hypothesized this result may be due to perception that discussing every test given might lengthen reports without adding clarity or clinical utility. Report length varied by specialty, reimbursement source, and employment setting. The median number of pages was six, although the range extended from 1 to in excess of 35 pages. Forensic reports and those reimbursed through medico-legal sources evidenced the greatest length with a median of 10 pages (range 4 – 35). This was in contrast to the medians for neurological reports of 6 pages (range 1 - 30), psychiatric reports of 7.5 pages (range 2 - 20), reimbursed by Medicaid/Medicare of 6 pages (range 1 - 15), and reimbursed by commercial insurance of 5 pages (range 1 - 30).

Donders’s (2001a, 2001b) overall conclusion after conducting this survey was that the field of neuropsychology agreed on many areas of content and format of written reports. He added the results also suggested there was no one correct way to write reports in which the needs of all practitioners would be satisfied. He continued by stating,
“Individual neuropsychologists should tailor their reports to the needs of their clients and the most likely target audience” (Donders, 2001b, p 160). He ended by explaining that the information contained within the study was produced with the hope of establishing a baseline for training. He made no recommendation as to the direction of training or about what information should and should not be generally included in neuropsychological reports beyond reporting the survey results.

One of Donders’s (2001a, 2001b) most interesting findings was the greater amount of information included in the reports of those whose practice was primarily devoted to pediatrics. He explained this by noting inconsistencies with the norms associated with pediatric neuropsychological tests and claimed adding this information to adult reports might add length to the report without adding clarity. This finding was at odds with Harvey’s (2006) contention that readability was a primary concern rather than length. Another possible reason for including normative group information was highlighted by Wong (2006), who cited research indicating a lack of consensus by neuropsychologists regarding test performance criterion level for defining impairment. He claimed an ethical imperative is to be clear about the basis for describing performance levels and stated the paucity of defining terms could cause confusion for members of the public. Further, Donders (2001b) found that only about 75% of neuropsychologists routinely or always included percentile scores and standard scores in reports. The lack of norms and numeric representation of scores for the consumer of the reports raises the question of whether or not this practice sufficiently meets a standard of explanation as required by the APA ethics code section 9.10 (APA, 2002).
Research on neuropsychological assessments was reviewed by Brooks, Strauss, Sherman, Iverson, and Slick (2009). They presented an overview of psychometric issues and advances that have led to better methods of interpretation of assessments. They began the review with an examination of normative samples, using the California Verbal Learning Test (CVLT) and the second edition (CVLT-II) to illustrate how updating norms or using different norming samples can lead to interpretive differences. They contended the CVLT was normed on a college educated population and argued this could have caused examinees to be identified as impaired on the CVLT who would be identified as scoring in the average to low average range using the more representative norms of the CVLT-II. They went on to discuss distribution and skew. Due to the possibility of misunderstanding of percentiles when the distribution for a population was skewed, they asserted attempting to use percentiles for $z$ scores would give erroneous impressions. The authors proposed that when norms were based in small sample size populations, caution should be exercised and rank based percentile scores would be more appropriate.

While reviewing findings regarding score magnitude and rank in score distribution the authors referred to a study by Crawford and Garthwaite (2009) on confidence interval and percentile rank reporting within neuropsychological reports to explain recommendations. Brooks et al. (2009) acknowledged the work of Bowman (2002) on the disadvantage of reporting scores as percentiles due to misunderstanding but explained how to make the information more easily understood by non-neuropsychologists. Of special note was inclusion of information on misinterpreting results when interpreting multiple tests. They cited multiple studies showing low scores were common in batteries
of tests and warned it is important to ensure demographic variables were also included in consideration. The authors advanced interpretative practice by discussing reliable change in test scores over time, how to measure it, adjustment for practice effects, and using this information to improve interpretation of test results for a client. They reviewed use of regression to evaluate the difference between predicted and observed scores, limitations for use of regression methods, and the importance for clinicians to use test-retest data in appropriate population samples to finalize the evaluation for interpretation.

The major theme of this comprehensive review was a presentation of psychometric issues necessary for proper interpretation of test scores. However, the authors did not make any reference to the need for the end user of a report to be aware of the information sources used for interpretation. As with Donders (2001b), it could be presumed written reports may be missing information due to normal clinician practice. However, Crawford and Garthwaite (2009) pointed out how to make information about norms and percentiles, etc., understandable and Brooks et al. (2009) noted the importance of using correct information for the basis of test interpretation. If this information is missing from reports, then future users or educated readers of the reports may not have all the information they need to draw correct conclusions from the test results at the present time and in the future.

Harvey (2006) noted the findings of multiple authors and researchers spanning more than thirty years stressing the importance of clear communication in psychological reports. In her earlier research (Harvey, 1997), she found school psychological reports were written at a Flesch-Kincaid grade level in excess of 15 and graduate student practice
reports were written at a grade level in excess of 13.6 even after revisions with instructions to simplify wording. The Flesch–Kincaid grade-level readability score was a formula calculated using a combination of sentence length and syllables per word, developed to gauge the readability of a text based on school grade level (Wempen et al., 2007). In her 2006 study, Harvey explored the possible reasons for a lack of clarity in psychological reports. She conducted exit interviews with recent psychologist graduates and found perceived problems in assessment training. This included teaching methods, an over-emphasis on test results and scores, lack of teaching of incorporation of clear psychological term definitions, the use of jargon, an emphasis on the information and not the client as a person, the time required to write reports, and confusion about writing for multiple audiences.

Harvey (2006) then reviewed 60 example reports from 20 text books and manuals for writing reports and conducted an analysis of the reports for readability. The grade level readability of those reports ranged from a mean of 18.49 for psychoeducational reports to a mean of 20.26 for clinical, neuropsychological, and forensic reports. All but one of the example reports were above a grade level of 14. As multiple sources recommend reduction of jargon, she then investigated the language put into the reports to determine if psychologists were writing to a consistent definition of “average intelligence.” She mailed a survey to psychologists associated with psychological organizations. The response rate to the mailed survey was 35% (n=208) with 96% (N=185) of the psychologists having a minimum of 5 years experience. The numeric definition of average intelligence was defined as between 90 to 110 by 67% of respondents. One standard deviation (85 to 115) was used by 20%, 7% defined it as above 100, while 3
respondents reported it as any score between 70 to 130. Comments about the definition of average intelligence included “not retarded” and not requiring “major modifications in the classroom” (p 10).

The author then conducted a random survey of 500 members of the National Association of School Psychologists on stress and time management. There was a response rate of 54% (n=272), with an average age of 42.99, and an average of 11.7 years of experience as a psychologist. The respondents completed a mean of 66.94 psychological evaluations per year and were mainly female (80%) practitioners who worked in schools (74%). Respondents indicated they wanted to decrease the amount of time writing reports (49%), although 1% reported a desire to increase the amount of time, and 50% failed to respond to the question. However, the findings also indicated few methods to reduce the time required to write reports were used. Secretarial support was not used/available for 75.5% of respondents, CBTIs were not used/available for 70.3%, templates for reports were not used/available for 51.9%, and 53.2% were unable to free themselves for a day to write reports. The author concluded that a lack of resources and time negatively impacted psychologists’ ability to write intelligible reports.

Harvey (2006) gave suggestions intended to improve the readability of reports written for multiple audiences. This included defining all terms, using percentile ranks, not making global or negative predictions, and writing to the level of the recipient’s education. In addition, she noted the need for psychologist students to have model reports that were clear, had less jargon, and were easier for non-psychologists to understand. She gave examples of writing for a report recipient in an easily understood manner. Harvey
suggested using report templates to lessen the time burden on psychologists and to add clarity to definitions. She also recommended using software for scoring and interpretation of tests, but added the evaluator should rewrite the reports to eliminate overstatements and redundancies.

Interpretative evaluation reports have been researched for some areas of content. Findings indicate areas in which improvements could be made such as readability, inclusion of numeric data, and inclusion of norms (Brooks et al, 2009; Crawford & Garthwaite, 2009; Harvey, 2006). Sattler (2001) identified these areas as necessary elements for report writing in his instructional guide as did Kaufman and Lichtenberger (2002). Donders’s (2001b) contention that adding this material could add length without clarity and that the practices profiled within his research were good because it showed consistency in training went against the recommendations of leaders in intelligence testing (Kaufman & Lichtenberger, 2002; Sattler, 2001; Tulsky, 2003). Therefore research into what is actually placed into evaluative reports is a necessity to understand if the principles articulated by leaders in the field have been followed. Investigation of the content could be accomplished through re-interpreting each report and comparing it with the original report for interpretive accuracy and completeness of content. This would be a time-consuming task and liable to errors if conducted by individual investigators. In order to ensure accuracy and completeness of the comparative reports an alternative method would be the use of a CBTI, the method investigated by this study.
Computer Based Test Interpretation

Research on CBTIs has been focused primarily on personality tests and the interpretive statements produced. Grove, Zald, Lebow, Snitz, and Nelson (2000) performed a meta-analysis on clinical versus mechanical (CBTI) predictions. Their work included an exhaustive review and analysis of research from multiple fields (psychology, medical, education, and financial) utilizing CBTI and clinical prediction. Each study was encoded for effect size (ES) and compared. Interrater reliability was checked and found to be satisfactory (\( R = .97 \)).

The findings indicated that the training or experience of the clinicians “did not significantly predict the degree of superiority of mechanical over clinical prediction” (Grove et al., 2000, p 463). The authors also found that mechanical prediction was superior to or as accurate as clinician prediction in all but a few studies (prediction notably favored by CBTI in 47% of studies and notably favored by clinical prediction in 6%). The studies which showed superior clinician judgment included clinical interviews and were found in studies that dated from 1958 to 1984 and consisted of diagnosing schizophrenia, homosexuality, and intellectual deterioration. Although mechanical prediction was shown to be superior overall the results were not striking, with an ES that was only 10% greater than for clinician prediction. For many studies the ES for clinical prediction and mechanical prediction were roughly equal. These findings were in line with previous research that also found equivalent results for clinical and mechanical prediction or a modest advantage for mechanical prediction (Endres, Guastello, & Rieke, 1992; Garb, 2000; Tsemberis, Miller, & Gartner, 1996).
Lichtenberger (2006) noted the work of Grove et al. (2000) in her report on the greatly expanded role of computers in psychological assessment since their introduction and the controversy still surrounding their use. She addressed the benefits of computer use in assessments: clerical efficiency, data storage, and processing and translating large amounts of data into interpretive statements. The author continued discussion of the benefits by exploring administration of tests via computers, computerized adaptive tests, and the ability of computerized tests to include multi-media formats. She then presented the controversy associated with CBTI. The first, and possibly most problematic, area is that many CBTIs have not been validated through research, leading to the potential that the programs could be based on biased information produced by the authors of the program. The concern over validity has been echoed by others (Groth-Marnat & Horvath, 2006; Grove et al., 2000; Meyer et al., 2001; Michaels, 2006; Snyder, 2000). The MMPI-2 CBTIs are one area in which research has been completed to prove validity with a comparison to expert clinician interpretation (Butcher, Perry, & Atlis, 2000; Williams & Weed, 2004). Lichtenberger expressed the concern that CBTIs use generalized rather than targeted statements in narratives and could: a) Cultivate passivity toward evaluations; b) Produce a perception of empirical accuracy (when the validity of the CBTI has not been researched); c) Foster reliance on generalized statements rather than targeted statements; d) Encourage use of CBTI statements as a clinician’s own without acknowledgment; e) Reduce clinician participation in the decision-making process; f) Promote inadequately trained examiners to use tests and reach beyond his/her area of expertise; g) Lead to the use of obsolete software.
Lichtenberger (2006) further set goals for the future use of CBTIs and improving the clinical usefulness of them. She observed that, while well trained and knowledgeable clinicians were crucial to the assessment process, CBTIs rooted in actuarial prediction models, coupled with narrative statements designed by experts, and combined with clinician judgment may produce an excellent report. She then laid out specific steps and recommendations for clinicians, instructors, and test developers to enhance their abilities to produce the best possible final report. For clinicians she recommended that they: a) Carefully evaluate all hypotheses generated by a CBTI for congruency with known information about the client before inclusion in the report; b) Investigate alternative hypotheses for support in the conclusions; c) Reduce dependence on memory by use of written notes taken during assessments; d) Use other statistical prediction models to gauge data that was not included in the CBTI narrative. For instructors she recommended: a) Formal training in assessment that included use of CBTI; b) Ensuring that CBTI narrative statements are not used verbatim or injudiciously within assessment reports; c) Teaching trainees to assimilate CBTI narratives within evaluative reports. She also recommends that test developers: a) Modify programs to include decision trees to incorporate clinician/CBTI interface; b) Utilize research to enhance clinical judgment by including questions on examinee characteristics and concerns.

A Wechsler Intelligence Scale for Children–Revised (WISC-R) CBTI was investigated for validity and efficacy by Tsemberis, Miller, and Gartner (1995). The CBTI was based on actuarial data and population norms. The authors recruited 33 clinicians that had between 3 and 32 years experience in testing and who had no exposure to the WISC-R CBTI. The clinicians were not informed of the purpose of the study. A
psychologist with 10 years experience submitted 2 WISC-R evaluative reports of a female child for use in the study. The reports had descriptive information removed. CBTI reports were produced based on scores contained in the WISC-R evaluative reports. The CBTI reports were modified to change gender-neutral terms to specify a female child. A CBTI report and the psychologist written report for each child were given to the clinicians. The clinicians did not know who authored the reports. They were then asked to compare the reports that were written for each child and rate them on a nine-point Likert scale. The scale consisted of items distilled from a review of literature and included interpretations, recommendations, organization, and comprehensiveness. At the end of each survey the clinicians were asked to identify if the report they rated was created by a computer or written by a psychologist. Results indicated that the CBTI reports were consistently rated as superior in every category. The clinicians correctly identified 68% of the CBTI reports as being computer generated and misidentified 29% of the psychologist written reports as being generated by a CBTI.

The authors posited one reason for superior ratings of the CBTI may have been inclusion of information on the child’s performance on all scales while the psychologist reports focused on scores that deviated from the mean. They noted several weaknesses in the CBTI. First, the CBTI didn’t differentiate between the meanings of different score configurations but generated the same comparisons regardless of the meaning of the difference. Second, it couldn’t answer specific referral questions. They recommended that the CBTI be used as a baseline to augment clinician observations rather than an end product to address these weaknesses. They also noted that the study would have benefited from using a greater number of reports and expert raters. They concluded that the study
was an initial finding for the validity and efficacy of the WISC-R CBTI. Discussion about the validity of any CBTI should recognize both strengths and weaknesses, exemplified by the Tsemberis et al. (1995) study.

Computer programs used for interpretation differ between primarily consultative CBTIs and CBTIs based on actuarial data that “intended to supplant rather than compliment clinical judgment” (Snyder, 2000, p 52). Snyder restated the arguments about the generalized descriptive nature of and validity of CBTIs, but also addressed the issue of customer satisfaction. This was also the contention of Brenner (2003), who observed that research on CBTIs should consider the clinical utility of the programs, that researchers should “focus on the needs of the consumers of psychological assessments” (p 240), and identified the consumers as the referral sources to whom the final report was directed.

The consumers of the reports want specific statements and recommendations, reduced jargon, and a focus on strengths and many of the current CBTIs do not fulfill those needs (Brenner, 2003; Harvey & Carlson, 2003; Michaels, 2006; Snyder, 2000). Thus it becomes essential that users of CBTIs understand the test that the CBTI is based on and ensure the CBTI is valid for a specific use, but should also use the CBTI report as a portion of data to be integrated with other information about the client after examining it to ensure accuracy rather than as an end report (Berger, 2006; Harvey & Carlson, 2003; McMinn, Ellens & Soref, 1999). Research has shown that some CBTI reports are considered better than or equal to clinician written reports (Andrews & Gutkin, 1991; Tsemberis, Miller, & Gartner, 1995). Because of the perceived superiority of CBTI
reports over clinician written reports it could be tempting to cut and paste the generalized statements generated by a CBTI into an interpretive report. This leads to a concern about the Barnum effect, the use of broad statements that could refer to anyone (Guastello, Guastello, & Craft, 1989), and recognition that competence in interpretation of tests is an ethical imperative.

The ethical issues regarding competence in consideration of advances being made in assessment practices and CBTIs have been addressed by the ethics codes that govern psychologists. The APA Ethics Code was updated in 2002 and section 9.09(c) addresses the above concern stating that the psychologist rendering services is responsible for the product of an assessment whether he or she scores and interprets it or uses a CBTI to do so (APA, 2002). The National Association of School Psychologists (NASP) goes further in stating that CBTI reports that are not edited must be identified as such and that a school psychologist should not use any test for which they have not been trained (NASP, 2000). In 1999 the APA joined with the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) to establish guidelines specifically governing use of tests and CBTIs in the fields governed by their respective organizations (AERA, APA, & NCME, 1999). The result was a clarified set of standards that crossed the boundaries of related fields.

The guidelines lay the foundation for test construction, validation, and documentation of the processes. The first section sets forth the primary principles in establishing the validity of a test. It states that the rationale for interpretation should be documented and included with information provided to assessors along with a summary of the evidence
used to reach conclusions and the theory that the interpretation is based on. It also 
addresses the initial creation of a test. It states that the qualifications of the creators of a 
test, a description of the training of raters used to establish the validity of the test, detailed 
information on data collection methods, effectiveness for the proposed use, and the 
population the test was normed on should be included within test documentation so that 
assessors have the ability to make informed judgments about whether it should be used 
for the purpose and the individual planned. Testing conditions, applications, and the 
rights and responsibilities of test users and those taking the tests are also addressed. The 
guidelines are enumerated within the volume and expounded upon to explain proposed 
best practices. The guidelines specifically address CBTIs in several sections. In addition 
to the above principles an assessor is charged with ensuring the quality of the 
interpretation of a CBTI and that the interpretation is based on correct norms for the use. 
Every concern that has been expressed in studies on CBTIs is approached and 
unambiguously answered. Unlike some guidelines that have generalized statements 
whose meanings are explained by others, the *Standards for Educational and 
Psychological Testing* is comprehensive and well articulated (AERA, APA, & NCME, 
1999).

The primary author of this research study constructed a CBTI (Moore & Christiansen, 
2007) that followed the steps Kaufman and Lichtenberger (2002) outlined for analysis 
and interpretation of WAIS-III results. Although the use of CBTIs to analyze assessments 
and formulate interpretations in clinical settings has not been without controversy (Groth-
Marnat & Horvath, 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Lichtenberger, 
2006; Meyer et al., 2001; Michaels, 2006; Snyder, 2000) the CBTI was designed for
research purposes only. It followed the guidelines outlined in the *Standards for Educational and Psychological Testing* (American Educational Research Association, APA, National Council on Measurement in Education, 1999) for creation and validation of a CBTI. The validation of the CBTI was predicated upon the premise of the WAIS-III validity. As such the validation of the CBTI was established by demonstrating the integrity of the program in returning consistent reports per WAIS-III criteria. The validation study also followed the *Standards for Educational and Psychological Testing* guidelines (AERA, APA, & NCME, 1999). No errors were found within the research study, meaning the CBTI returned numerical results that were 100% consistent with WAIS-III results (Moore & Christiansen, 2007).

The CBTI formulated interpretive statements based on analysis of various score comparisons. For example, the CBTI calculated a cognitive strength by comparing a subtest score to the other subtest scores, determining the significance of any difference via documented WAIS and statistical guidelines, and offered a series of statements about what the strength may indicate in the functioning of the individual being tested. The CBTI was designed to be overly inclusive by listing all possible interpretive statements and did not make any indication about which potential interpretive statement best fit the individual client. The use of score differences listed in the WAIS manual within the program allowed an incorporation of the findings contained in the manual (Wechsler, 1997) to substantiate reliability for this CBTI and fulfill the guidelines established by the AERA, APA, and NCME (1999). The consistency of the calculations with those within the manual also worked to establish reliability of the CBTI in accordance with the manual and the guidelines.
There were some weaknesses in the study and the CBTI. The study included only a small sample of WAIS-III protocols (n = 20). Testing with a larger sample may have revealed weaknesses in the program undiscovered by the small sample. The CBTI also had some areas of weakness. It generated the same data for individual protocols that had high or low findings if those findings had similar comparative ratings rather than generating different hypotheses based on the levels within the protocol. This could give a false impression of the significance of the findings as was pointed out in the *WAIS-III/WMS-III Technical Manual* (Wechsler, 1997a). For example, a low Working Memory Index (WMI) score in an individual with a learning disability is a common finding but the same finding is not applied to individuals with a high IQ. The CBTI also did not return any recommendations. This may be acceptable for use within research settings investigating how thoroughly WAIS-III interpretive reports are written but would require greater clinician judgment if used in a clinical setting. The statements used to generate the analysis produced by the program were written by the primary author and crosschecked by the secondary author for accuracy. Submission of the statements to a greater number of experts may have improved the quality of those statements. The statements also contained jargon that may not be properly understood by those who are not thoroughly familiar with intelligence testing. This could be considered a weakness if the statements were not edited for inclusion within an interpretive report that was to be given to a client.

The CBTI has not been compared to or recalibrated for use with the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV). Further research in this area is required before any recommendation could be made on what changes in the program may be
required. This should limit its utility to its original purpose, research, at this time. For the field of CBTIs in general, an adherence to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) would result in an improvement in validity and could also enhance trust in CBTIs.
PURPOSE OF THE CURRENT STUDY

The current study seeks to examine thoroughness of student generated WAIS-III reports by comparing the content of reports done as a final class assignment to the content generated by the CBTI. It is conceptualized as a foundational review seeking to establish a base of knowledge as to what essential information is placed within a graduate student WAIS-III interpretative evaluation report. The purpose is to initiate questions that will spur future researchers to seek areas of knowledge and understanding about the conveyance of information that could benefit from further development in emerging clinicians and practitioners.

The premise of the study follows the line of reasoning that errors found in the administration of the WAIS would also be reflected in a similar type of errors within the interpretation of the WAIS. Each report will be considered separately and the scores reported in each protocol considered on their own merit. No attempt will be made to match the interpretation of the scores obtained to an interpretation of the baseline scores found by the instructors. All administration and scoring errors will be considered to be valid findings in order that the focus be placed on the interpretation of those findings in their own right as the primary consideration of this study is on what information is included within graduate student WAIS-III interpretative evaluation reports.

In line with the findings on errors in the administration and scoring of the Wechsler intelligence scales we expect to find that all possible variables that could be calculated will not be included in the interpretative evaluation reports. However, it is hypothesized
the three primary findings of FSIQ, VIQ, and PIQ will be reported as will major uncommon findings such as an abnormal split between the VIQ and PIQ and scatter in an index. Specific areas considered to be prone to a presentation of incomplete findings in interpretation are the indices comparisons and strengths and weaknesses within the subtests. Finally, it is hypothesized that the students will commit computational errors resulting in incorrect interpretation.

The definition of what constitutes a finding is taken from the method proposed by Kaufman and Lichtenberger (2002) in addition to the WAIS-III manuals (Wechsler, 1997; Wechsler, 1997a). One area that may be considered controversial is the issue of reporting and interpreting scatter. Although investigations into scatter have raised questions as to interpretative meaning and have shown a wider range of variability as the FSIQ increases (Grégoire, 2005; Matarazzo, Daniel, Prifitera, & Herman, 1988; Mclean, Kaufman, & Reynolds, 1989; Ryan, Kreiner, & Burton, 2002), in order to maintain consistency within the study the Kaufman-Lichtenberger definition was selected as a base finding. They defined scatter as an inconsistency in subtest scores within either an IQ domain or within an index that was significant and could affect the unitary construct of the domain or index (Kaufman & Lichtenberger, 2002). They then went on to note that the meaning of the discrepant finding should be interpreted before proceeding further. This concept was placed into practical terms by Brooks et al. (2009). They discussed change over time as an element of methodology and cautioned against relying on statistically reliable change as it did not always convert into meaningful change, but rather proposed using clinical judgment to identify abnormal differences and to provide clinical confidence in the meaning of those differences when inferring impairments. To
this end and to allow clinical judgment for necessary items included to be properly measured, intersubtest scatter is identified as an exclusionary finding (e.g., if scatter is found within an index then comparisons involving that index will not be identified as necessary elements to be included within the interpretative evaluation report).

It is hoped that this study will be used to spur further research by beginning the discussion of why information is or is not placed within the reports rather than to simply answer the question of what is placed within the reports. Graduate students within psychology programs writing interpretative evaluation reports need to know and understand the significance of information they include within those reports as it relates to the current and future functioning of the client. Without an initial finding on what information is placed within interpretative reports then patterns of information included in those reports related to referral questions will not be identified.

As such the goal of this preliminary study is to identify areas within WAIS-III interpretative evaluation reports which could benefit from increased attention by graduate students. An analysis of the findings hopes to highlight portions of the interpretative reports that are being missed. Areas of weakness within the student reports could provide valuable information to instructors regarding areas of interpretation needing to be better elucidated and emphasized within an instructional base.
METHOD

Participants

Participants in the study were first year Pacific University psychology graduate students who were enrolled in an intelligence assessment course at the School of Professional Psychology (SPP) during the time of the study. The class curriculum included WAIS-III protocols based on fictitious clients. Students were notified of the study and provided an opportunity to voluntarily participate in the study during class time. No payment or compensation was offered for participation in the study. They were notified the decision whether or not to participate would not affect their current or future relations with Pacific University, grade, or standing in the class. They were also informed a decision to participate included the freedom to withdraw participation at any time without prejudice or negative consequences. An oral presentation about the study, accompanied by a short-form consent document (stating the necessary elements and a written summary of what was presented orally) was provided to members of the class. All students were given the opportunity to request the findings of the study regardless of participation through a sign-up sheet.

If individuals wished to participate they were required to sign an informed consent release. Each participating student submitted a copy of the final WAIS-III interpretative report they wrote for the class, which was based on a protocol of a factitious client. A total of 38 of 46 students signed consent forms and presented completed WAIS-III interpretative reports for evaluation. All student identifying information was blacked out
and a copy made of the properly redacted report. The original pages were then returned to
the instructor. No further contact with participants occurred.

The WAIS-III interpretative reports completed as the final class assignment were
compared to the CBTI to investigate the level of thoroughness in the reports. The
students were expected to score the protocols and write a report based on the findings of
the assessment in relation to the referral question. In the case of the assignment used for
this study, the referral question asked for assessment of cognitive functioning with
implications for treatment and referred to a 51 year-old married female, who was college
educated, diagnosed with Major Depressive Disorder two years ago, and whose clinical
status included depression and forgetfulness. Students were expected to interpret all the
indices, configurations of them, and strengths and weaknesses of each subtest. However,
the main objective of the training exercise was for them to do a critical analysis of the
information and present a clear hypothesis in answer to the referral question. They were
also to include an explanatory reference to the test data they used to formulate any
hypotheses proffered in the report. The instructors directed the students to describe the
results, hypotheses, and recommendations in layman’s language. To this end they
provided examples of reports written in layman’s language and written feedback on the
previous practice reports. There was no expectation placed upon formatting the data in a
specific manner or that specific items of information were to be included in the report.

Measures

*Wechsler Adult Intelligence Scale - Third Edition (WAIS-III).*

The *Standards for Educational and Psychological Testing* guidelines lay the
foundation for test construction, validation, and documentation (AERA, APA, & NCME,
Those standards were upheld in the validation of the WAIS-III. All appropriate information (as per the guidelines) was documented (Hess, 2001; Kaufman & Lichtenberger, 2002; Rogers, 2001) and placed in the technical manual (Wechsler, 1997a). The Wechsler Adult Intelligence Scale - Third Edition (WAIS-III) was built upon previous editions and included updated norms in consideration of the Flynn effect, an increase in age range (upper age from 74 to 89 years of age) due to increased life expectancy, addition of a subtest (Matrix Reasoning) to increase efficacy for older adults, and modified subtest items to improve cross cultural effectiveness, make items more contemporary, and to enhance clinical utility of the subtests. The WAIS-III was standardized using 13 age range bands on a sample of 2450 individuals (Wechsler, D., 1997a).

Evidence presented within the *WAIS-III/WMS-III Technical Manual* (Wechsler, D., 1997a) indicated the subtests were reliable. Test-retest reliabilities were examined within 2 to 12-week spans and ranged from the .70s to the .90s. The Verbal IQ (VIQ), Performance IQ (PIQ), and FSIQ exhibited stability coefficients in the .90s and index stability coefficients ranged from the high .80s to the .90s. There was a gain of about 3 points for the VIQ and FSIQ and of about 5 points for the PIQ upon retesting. Interrater reliability coefficient scores during test evaluation were exemplified by verbal domain subtest scores ranging from .91 to .95. The criterion related validity of the WAIS-III included correlation with the WAIS-R (VIQ .94; PIQ .86; FSIQ .93) and with the WISC-III (VIQ .88; PIQ .78; FSIQ .88). The WAIS-III also showed a .88 correlation with the Stanford-Binet Fourth Edition. Content validity was established with the assistance of an advisory panel, outside experts, and literature reviews conducted to investigate possible
problems with the WAIS-R. Construct validity was substantiated through factorial analyses which supported the four factor index model.

The technical manual presented a section delineating different constellations of scores for multiple populations as identified within special group studies. This was done in keeping with the stated goals of the WAIS-III to assist in assessment of psychoeducational disability, giftedness, and organic and neuropsychiatric dysfunction. The categories included neurological disorders (as identified by Alzheimer's disease, Huntington's disease, Parkinson's disease, multiple sclerosis, temporal lobe epilepsy, and traumatic brain injury), alcohol related disorders (as identified by chronic alcohol abuse and Korsakoff's syndrome), a schizophrenic group (identified as a neuropsychiatric disorder), psychoeducational and developmental disorders (as identified by mental retardation, attention-deficit-hyperactivity, and learning disorders), and deaf and hearing impaired. This information was included under the auspices of construct validity. Tables were also included to assist in identifying predicted scores of the WAIS-III from WIAT scores and to identify index score differences within the WAIS-III by age groups. Exclusionary information for all studies was also included (Wechsler, D., 1997a).

Computer Based Test Interpretation (CBTI) for the WAIS-III.

In an attempt to address the lack of research on the thoroughness of evaluation reports the principal investigator designed and constructed an Excel based computerized analytical program (Moore & Christiansen, 2007), a CBTI, following the steps Kaufman and Lichtenberger (2002) outlined for analysis of the WAIS-III. The primary author codified those steps, created formulae to reproduce them, and wrote interpretative
statements for each condition. The program used the formulae to generate the interpretative statements both unique to the individual and standardized.

The CBTI provided individual interpretations through a series of analytical calculations of scores obtained from each protocol (test booklet) that had been completed. The obtained analysis generated a standardized series of interpretive statements that corresponded to different configurations of scores and calculated each variable of the WAIS-III according to the Kaufman-Lichtenberger method. The program was designed to be used in research and was not designed to be a ‘stand-alone’ instrument. It was written to be overly inclusive of information in order that clinician judgment could be used to parse out the specific details required for each examinee.

Although the Kaufman-Lichtenberger method was used, the calculations for strengths and weaknesses within the subtests were based on data in the WAIS-III examiner’s manual (Wechsler, 1997). Kaufman and Lichtenberger (2002) used cutoff points that were smaller than those found in the WAIS-III manual (Wechsler, 1997). Research indicated use of less stringent criteria could result in inclusion of scores within the confidence interval of the subtest (Charter, 2003). This could result in providing inaccurate feedback to a client. The APA ethical guidelines (APA, 2002) charge psychologists with notifying clients of limitations in the accuracy of the interpretations. Use of confidence intervals outside of those that have been established through research may violate the spirit of the assessor’s ethical duties. For those reasons the more conservative values provided in the WAIS-III manual (Wechsler, 1997) were used.
The program was written to include all possible permutations of the calculations outlined by Kaufman and Lichtenberger (2002). For example, the value for each of the indices, if scatter was found within them, and how the indices related to each other was assigned a statement. The statements were derived from analysis of the Kaufman and Lichtenberger (2002) text and the WAIS-III manual (Wechsler, 1997). The Kaufman and Lichtenberger text (2002) was selected based on its common use as a text for instruction and Kaufman’s reputation for excellence in intellectual assessment.

For each finding within the protocol several statements were provided by the CBTI. One statement was an explanation of each calculation that was used to formulate the interpretation in the evaluation report (i.e. VIQ/PIQ No Significant Difference; 0 – 8 points). The second statement was an example of how the calculation may appear in narrative form in the evaluation report (i.e. The VIQ/PIQ difference is not significant. This means that the verbal reasoning abilities and the nonverbal, visual-spatial reasoning abilities are similar. No significant difference is 8 or fewer points difference.). In addition, where warranted, there was an explanation as to what any differences in scores may mean (i.e. The discrepancy between the indices indicates that there is a stronger ability in nonverbal, visual-spatial reasoning abilities, processing information, manipulating visual-spatial information [integration of visual-motor skills and mental manipulation to solve problems], and the application of visual-motor skills than verbal reasoning abilities), what the difference may appear as in client functioning (i.e. There is an inclination to be relatively slower to understand, mentally manipulate information, and express ideas. This could result in a preferred method of learning by being shown rather than being told instructions), and what hypotheses may account for a significant
difference within each subtest (i.e. A low score may reflect problems with attention span, state anxiety, inflexibility, learning disabilities, negativism, impulsiveness, auditory impairment, or less than ideal testing conditions). Subtest strengths and weaknesses were accompanied by the amount of variance from the mean.

The multiple permutations possible ensured each WAIS-III evaluation report analyzed by the CBTI was both unique to the individual and standardized. The Excel program was unique to each individual through a series of analytical calculations of individual scores obtained from completed protocols. The obtained analysis generated a standardized series of interpretive statements that corresponded to different configurations of scores and the comparative strength of those scores. The CBTI did not incorporate interpretations beyond those given within the cited sources. Construct and criterion related validity for the CBTI generated statements were previously established by the same sources. The research study used to validate the CBTI found 100% consistency between the CBTI and calculated scoring of protocols. This level of reliability was deemed necessary to maintain validity of the CBTI with the validity established for the WAIS-III and to allow for consistent findings of fact.

Procedure

Each WAIS-III interpretative report was assigned a tracking number. Scores and the tracking number from each protocol were entered into the CBTI and a printout made of the CBTI report (sample copy of a CBTI report included in Appendix A). A template was created for scoring the student reports in comparison to the CBTI reports (sample copy included in Appendix B). Each template was identified as to the specific tracking number and CBTI generated report compared. The CBTI report was used to assist in identifying
the configuration of elements that should have been included in the student interpretative evaluation reports according to the protocol dataset scored by the student and to assist in identifying any information presented that could be interpreted as identifying key elements sought in this study.

The rules of identification for this study were designed to allow for variance based on individual writing styles. If any element of the student WAIS-III interpretative report contained data identifying specific elements of interest (domains, index scores, consistency, comparisons, or subtests) or could be seen as including an element of the corresponding statement of the CBTI report it was marked as included. For example, if the interpretative information on the VIQ noted it was based on two sources of data, verbal comprehension and memory and provided some type of definition of those abilities, then it could be surmised the author defined the VIQ, Verbal Comprehension Index (VCI), and the Working Memory Index (WMI) and those items were marked as included on the score sheet. However, a generalized statement of the comparison of the VIQ and PIQ was not deemed to be sufficient to account for comparison of specific indexes (i.e. VCI/Performance Speed Index [PSI]). The exception to the rule was if essential elements of the index scores were included, note was made of multiple sources of information, and/or there was a statement about consistency/inconsistency within a domain while noting the combination of elements (i.e. Performance Organization Index [POI]/PSI within the PIQ and including nonverbal, visual-spatial reasoning abilities, processing information, manipulating visual-spatial information, integration of visual-motor skills, mental manipulation to solve problems, and the application of visual-motor skills and abilities of response speed, speed of thinking, and motor speed). If a discussion
of strengths and weaknesses referred to any elements of a subtest identified as such or if the
discussion used an example of abilities which could be interpreted as specifying the
skill set measured by the subtest it was marked as included. The accuracy of the
computations used to evaluate the relationship of the indexes was also checked. However,
no attempt was made to gauge the accuracy of the scoring of subtests within the protocol.

The reports were initially checked by the primary researcher. A second check of the
reports was done separately by a specially trained research assistant. The assistant was a
doctoral level intern who had successfully completed coursework in intelligence testing
using the Kaufman-Lichtenberger method. The assistant then checked the results of the
first check against his own. Any variance between the results was examined by both
parties and an agreement reached as to the meaning of terms in relationship to the
findings. Interrater reliability for initial degree of agreement was 92.06%. Material used
as a guideline for interpretative evaluation included Assessing adolescent and adult
Intelligence Second Edition (Kaufman & Lichtenberger, 2002), WAIS-III: Administration
WAIS-III/WMS-III Technical Manual (Wechsler, 1997a), and Clinical Interpretation of
the WAIS III and WMS III (Tulsky, 2003).
RESULTS

The findings were initially tallied as to whether or not they were measured per previously noted criteria according to the protocol scores included in the report and recorded as $n$ in Table 1. Any finding of subtest scatter within an index that precluded that index from indices comparative measurement and all associated indices comparisons were not scored as measured results. The interpretations of constructs measured in each report were then checked against expectations of constructs which should have been found and interpreted as defined by the rules of identification. Constructs measured and expected to be interpreted were recorded in Table 1 as Finding interpreted. To determine whether or not the specific hypotheses proposed by this study were supported, each area of interpretation and inclusion was examined for the sample as a whole. Constructs which were not measured but for which an interpretation was provided within the evaluation report were included in Table 2 as Interpreted without results. Table 2 also included the total number of all interpretations for each construct. This was to allow the findings to be presented as totals of specific measurements and as an overall view. An interpretive element was considered to be lacking in the total sample of assessment reports if it was present in less than 80% of the sample.

The hypothesis that FSIQ, VIQ, PIQ and any VIQ/PIQ difference would be consistently interpreted and included in the evaluative reports was generally validated. The student interpretative reports included some rendering of interpretation for the major findings of FSIQ, VIQ, and PIQ and an abnormal split in the VIQ/PIQ (defined as
exceeding 19 points difference). Of 38 protocols investigated only 1 failed to report a finding in these areas (VIQ/PIQ split) although the abnormal difference was measured according to the protocol scores included in the report.

Table 1

*Constructs Interpreted*

<table>
<thead>
<tr>
<th>Construct</th>
<th>n</th>
<th>Finding interpreted</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSIQ</td>
<td>38</td>
<td>38</td>
<td>100.00%</td>
</tr>
<tr>
<td>VIQ</td>
<td>38</td>
<td>38</td>
<td>100.00%</td>
</tr>
<tr>
<td>PIQ</td>
<td>38</td>
<td>38</td>
<td>100.00%</td>
</tr>
<tr>
<td>VIQ/PIQ split</td>
<td>38</td>
<td>37</td>
<td>97.37%</td>
</tr>
<tr>
<td>VCI</td>
<td>38</td>
<td>37</td>
<td>97.37%</td>
</tr>
<tr>
<td>WMI</td>
<td>38</td>
<td>36</td>
<td>94.74%</td>
</tr>
<tr>
<td>PSI</td>
<td>38</td>
<td>36</td>
<td>94.74%</td>
</tr>
<tr>
<td>POI</td>
<td>38</td>
<td>36</td>
<td>94.74%</td>
</tr>
<tr>
<td>Scatter in VCI</td>
<td>33</td>
<td>6</td>
<td>18.18%</td>
</tr>
<tr>
<td>VCI/POI</td>
<td>5</td>
<td>4</td>
<td>80.00%</td>
</tr>
<tr>
<td>VCI/WMI</td>
<td>5</td>
<td>2</td>
<td>40.00%</td>
</tr>
<tr>
<td>POI/PSI</td>
<td>38</td>
<td>17</td>
<td>44.74%</td>
</tr>
<tr>
<td>VCI/PSI</td>
<td>5</td>
<td>3</td>
<td>60.00%</td>
</tr>
<tr>
<td>POI/WMI</td>
<td>38</td>
<td>16</td>
<td>42.11%</td>
</tr>
<tr>
<td>WMI/PSI</td>
<td>38</td>
<td>16</td>
<td>42.11%</td>
</tr>
<tr>
<td>Verbal Strengths</td>
<td>22</td>
<td>16</td>
<td>72.73%</td>
</tr>
<tr>
<td>Verbal Weaknesses</td>
<td>38</td>
<td>24</td>
<td>63.16%</td>
</tr>
<tr>
<td>Performance Strengths</td>
<td>3</td>
<td>1</td>
<td>33.33%</td>
</tr>
<tr>
<td>Performance Weaknesses</td>
<td>0</td>
<td>0</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

*Scatter in VCI precluded comparison for purposes of study.*

*Finding interpreted was the number of each construct interpreted in agreement with those measured according to the protocol scores included in the report.*

The individual indexes also showed a high number of interpretations with only 7 of 152 (4 indices in each of 38 protocols) possible interpretations being neglected. The finding was clouded by the rules of interpretation. If any element of the index was
included in the interpretation of the domain, the specific index was considered to have been interpreted (i.e. if the PIQ interpretation included a statement that incorporated motor speed and perceptual organization as part of the interpretation then both the PSI and POI were considered to have been interpreted and were included within the results). Almost all of the interpretations of individual indexes were included within the domain interpretation rather than individually. The overly broad rules for inclusion were designed to ensure that the elements of each construct were included without penalizing the writer for either attempting to write in a style easily understood by a layman or for providing material that could be considered overly dense for ease of reading.

However, the hypothesis that scatter as a major finding would be consistently interpreted and included in the reports was not sustained. In 33 protocols the student reports had positive measurements for scatter within the VCI but only 18% of those measured for scatter provided an interpretation of the finding (as minimally defined by stating there was scatter or inconsistency within the index or domain). It should be noted that interpretation of scatter was not considered to be a necessary element of the report by the instructors although it was included in training.

The VCI/POI indices comparison interpretation was provided as an alternative to the FSIQ score (due to the abnormal difference in the VIQ/PIQ affecting the FSIQ) in only 66% of the reports, and this was the most frequently included of all the possible indices comparisons. This confirmed the hypothesis that the indices comparison would be an area in which an incomplete presentation of findings would occur. The guidelines for inclusion of findings indicated that a statement of comparison could include sentences about each index next to each other within the same paragraph. In the case of domain
related indices a statement of consistency/inconsistency or differing levels of generalized ability within the domain sufficed to trigger inclusion. The results included in Table 1 reveal that interpretation of the indices comparisons other than the VCI/POI fell below 50% of the total when scatter was not included in consideration and only reached the level of 60% (3 of 5) in one comparison (VCI/PSI) when scatter excluded the majority of interpretations.

Interpretation of strengths and weakness of subtests within the reports fared little better. Only 58% of the protocols measured a verbal strength and 73% of those were interpreted within the body of the report. All of the protocols measured a verbal weakness and 63% of the reports included an interpretation of this finding. Within the performance domain only 3 protocols had strengths measured and none had weaknesses measured. Performance domain strengths were interpreted by only 33% of those students who scored them within the protocols and a total of 16% interpreted them overall.
Inclusion of the constructs including the VCI, which were excluded from consideration due to scatter, improved the percentage of interpretations. The VCI/POI interpretation, an alternative to the FSIQ when a VIQ/PIQ split exists as in this case, reached 66% but the other constructs still fell below 50%. Although there were no

Table 2

<table>
<thead>
<tr>
<th>Constructs Interpreted</th>
<th>Interpreted without results</th>
<th>%</th>
<th>Total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSIQ</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>VIQ</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>PIQ</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>VIQ/PIQ split</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>VCI</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>WMI</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>PSI</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>POI</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>Scatter in VCI</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>VCI/POI a</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>VCI/WMI a</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>POI/PSI a</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>VCI/PSI a</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>POI/WMI</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>WMI/PSI</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>Verbal Strengths</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>Verbal Weaknesses</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>Performance Strengths</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
<tr>
<td>Performance Weaknesses</td>
<td>0</td>
<td>0.00%</td>
<td>5</td>
<td>15.79%</td>
</tr>
</tbody>
</table>

aScatter in VCI precluded comparison for purposes of study. Interpreted without results was the number of each construct interpreted without being measured according to the protocol scores included in the report. Total is the total number of each construct interpreted within all protocols.
measured performance domain or verbal domain weaknesses found in the protocols, Table 2 revealed that 34% of the reports included an interpretation of a performance domain weakness and 63% included an interpretation of a verbal domain weakness. The findings on inclusion within the interpretative reports of strengths and weaknesses confirmed the hypothesis that the subtest strength and weakness would be an area in which an incomplete presentation of findings would occur. In spite of the low percentage of measured strengths and weaknesses, the inclusion of those constructs within the reports wherein they were not measured indicated an understanding of the importance of the constructs albeit without cause.

The final hypothesis to be tested was that the students would commit computational errors resulting in incorrect interpretation. Although this finding was confirmed by the presence of errors, the incidences of errors were committed by only 18% of the students and individual errors comprised only 4.75% (27 of 569 calculations) of those possible. This rose above the established criteria level of 80% correct and thus provided a null hypothesis for the purpose of this study. Of the 27 out of 569 possible errors made, 22% resulted in a different interpretation due to the error. This included 2 changes in direction of the difference (i.e. reporting a higher POI than WMI), an increase in VIQ (i.e. increase in excess of 10 points), a decrease in the WMI (i.e. decrease in excess of 15 points), and several changes in the performance range given to the client (i.e. reduction from Low to Extremely Low range or from Superior to Average range).
Based on the computational errors a change was noted within a review of the scores. This was considered to be a separate issue from whether or not the errors were translated into a change in interpretation within the reports. Table 3 notes a consistency in a comparison of the mode of the constructs and the baseline construct scores although the high and low score differences ranged from 31 points to 4 points difference. When computational errors were removed from consideration (Table 4) the mode did not

| Table 3 |

**IQ and Index Score Review**

<table>
<thead>
<tr>
<th></th>
<th>Baseline*</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>SD</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSIQ</td>
<td>100</td>
<td>102.21</td>
<td>102</td>
<td>100</td>
<td>2.32</td>
<td>110</td>
<td>99</td>
</tr>
<tr>
<td>VIQ</td>
<td>119</td>
<td>122.76</td>
<td>122</td>
<td>122</td>
<td>4.3</td>
<td>140</td>
<td>116</td>
</tr>
<tr>
<td>PIQ</td>
<td>78</td>
<td>77.84</td>
<td>78</td>
<td>78</td>
<td>0.68</td>
<td>79</td>
<td>75</td>
</tr>
<tr>
<td>VCI</td>
<td>114</td>
<td>114.21</td>
<td>114</td>
<td>114</td>
<td>2.7</td>
<td>120</td>
<td>110</td>
</tr>
<tr>
<td>POI</td>
<td>78</td>
<td>77.79</td>
<td>78</td>
<td>78</td>
<td>0.78</td>
<td>78</td>
<td>74</td>
</tr>
<tr>
<td>WMI</td>
<td>126</td>
<td>124.53</td>
<td>126</td>
<td>126</td>
<td>5.94</td>
<td>126</td>
<td>95</td>
</tr>
<tr>
<td>PSI</td>
<td>76</td>
<td>75.82</td>
<td>76</td>
<td>76</td>
<td>2.24</td>
<td>79</td>
<td>63</td>
</tr>
</tbody>
</table>

*Baseline was the score provided as correct for the protocol by instructors of the course

| Table 4 |

**IQ And Index Score Review With Computational Errors Removed**

<table>
<thead>
<tr>
<th></th>
<th>Baseline*</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>SD</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSIQ</td>
<td>100</td>
<td>102.21</td>
<td>102</td>
<td>100</td>
<td>2.32</td>
<td>110</td>
<td>99</td>
</tr>
<tr>
<td>VIQ</td>
<td>119</td>
<td>122.30</td>
<td>122</td>
<td>122</td>
<td>4.3</td>
<td>130</td>
<td>116</td>
</tr>
<tr>
<td>PIQ</td>
<td>78</td>
<td>77.84</td>
<td>78</td>
<td>78</td>
<td>0.68</td>
<td>79</td>
<td>75</td>
</tr>
<tr>
<td>VCI</td>
<td>114</td>
<td>114.21</td>
<td>114</td>
<td>114</td>
<td>2.7</td>
<td>120</td>
<td>110</td>
</tr>
<tr>
<td>POI</td>
<td>78</td>
<td>77.79</td>
<td>78</td>
<td>78</td>
<td>0.78</td>
<td>78</td>
<td>74</td>
</tr>
<tr>
<td>WMI</td>
<td>126</td>
<td>125.32</td>
<td>126</td>
<td>126</td>
<td>3.37</td>
<td>126</td>
<td>106</td>
</tr>
<tr>
<td>PSI</td>
<td>76</td>
<td>76.16</td>
<td>76</td>
<td>76</td>
<td>0.69</td>
<td>79</td>
<td>76</td>
</tr>
</tbody>
</table>

*Baseline was the score provided as correct for the protocol by instructors of the course
change but the maximum range of high and low scores was reduced by 11 points. The changes in Table 4 IQ and index score review resulted from 3 computational errors. This was due to the minor amount of errors generally found. They primarily consisted of single point differences made while calculating differences or in adding up lists of scores.

Table 5

<table>
<thead>
<tr>
<th></th>
<th>Baseline*</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>SD</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIQ/PIQ</td>
<td>41</td>
<td>44.92</td>
<td>45</td>
<td>44</td>
<td>4.25</td>
<td>62</td>
<td>38</td>
</tr>
<tr>
<td>VCI/POI</td>
<td>36</td>
<td>36.42</td>
<td>36</td>
<td>36</td>
<td>2.68</td>
<td>42</td>
<td>32</td>
</tr>
<tr>
<td>VCI/WMI</td>
<td>-12</td>
<td>-10.32</td>
<td>-12</td>
<td>-12</td>
<td>6.80</td>
<td>19</td>
<td>-16</td>
</tr>
<tr>
<td>POI/PSI</td>
<td>2</td>
<td>1.97</td>
<td>2</td>
<td>2</td>
<td>2.39</td>
<td>15</td>
<td>-2</td>
</tr>
<tr>
<td>VCI/PSI</td>
<td>38</td>
<td>38.39</td>
<td>38</td>
<td>38</td>
<td>3.55</td>
<td>51</td>
<td>33</td>
</tr>
<tr>
<td>POI/WMI</td>
<td>-48</td>
<td>-46.74</td>
<td>-48</td>
<td>-48</td>
<td>6.04</td>
<td>-17</td>
<td>-52</td>
</tr>
<tr>
<td>WMI/PSI</td>
<td>50</td>
<td>48.71</td>
<td>50</td>
<td>50</td>
<td>4.36</td>
<td>50</td>
<td>30</td>
</tr>
</tbody>
</table>

*Baseline was the difference as calculated from scores provided as correct for the protocol by instructors of the course

Table 6

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>SD</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIQ/PIQ</td>
<td>41</td>
<td>44.46</td>
<td>45</td>
<td>44</td>
<td>3.19</td>
<td>52</td>
<td>38</td>
</tr>
<tr>
<td>VCI/POI</td>
<td>36</td>
<td>36.42</td>
<td>36</td>
<td>36</td>
<td>2.68</td>
<td>42</td>
<td>32</td>
</tr>
<tr>
<td>VCI/WMI</td>
<td>-12</td>
<td>11.75</td>
<td>-12</td>
<td>-12</td>
<td>2.82</td>
<td>-6</td>
<td>-16</td>
</tr>
<tr>
<td>POI/PSI</td>
<td>2</td>
<td>1.62</td>
<td>2</td>
<td>2</td>
<td>1.01</td>
<td>2</td>
<td>-2</td>
</tr>
<tr>
<td>VCI/PSI</td>
<td>38</td>
<td>38.39</td>
<td>38</td>
<td>38</td>
<td>3.55</td>
<td>51</td>
<td>33</td>
</tr>
<tr>
<td>POI/WMI</td>
<td>-48</td>
<td>-47.54</td>
<td>-48</td>
<td>-48</td>
<td>3.50</td>
<td>-28</td>
<td>-52</td>
</tr>
<tr>
<td>WMI/PSI</td>
<td>50</td>
<td>49.69</td>
<td>50</td>
<td>50</td>
<td>1.06</td>
<td>50</td>
<td>45</td>
</tr>
</tbody>
</table>

*Baseline was the difference as calculated from scores provided as correct for the protocol by instructors of the course
Table 5 collated the differences noted in IQ and indices comparisons. The significance of errors for the difference was established by the use of the *WAIS III Technical Manual* chart of significant differences (i.e. VCI/WMI difference of 10 points at a significance level of .05).

Based on the computational errors a change was noted. This was considered to be a separate issue from whether or not the errors were translated into a change in interpretation within the reports. Table 5 notes a consistency in a comparison of the mode of the constructs and the baseline construct scores although the high and low score differences ranged from 35 points to 10 points difference.

The only changes considered for this section were those caused by computational errors. The findings did not support a generalized pattern of computational errors causing different interpretations of the findings. In fact, Table 3 and Table 5 revealed a consistency within the reports and with the scoring by the instructors. When computational errors were removed from consideration (Table 6) the mode did not change but the maximum range of high and low scores was reduced by 11 points. This was the same level of change noted in the IQ and index score review (Table 4).
DISCUSSION

The primary purpose of this study was to examine what elements of a scored WAIS-III protocol were included within an evaluative report as written by graduate students within an intelligence assessment course. It was designed to be a foundational study to illuminate areas needing possible attention in order to assist instructors in preparing students. Continuing concerns regarding the administration and scoring of the Wechsler scales were reviewed as well as the use of CBTIs for generating reports, some practices in report writing, and ethics involved in assessment. This guided the direction of the study to establish the foundation of knowledge in order to best serve the interests of clients and spur further research into the area.

This study had a number of weaknesses. The participants were obtained from only one school and the same material was used to teach all participants. This limited the generalizability of the study. Inclusion of participants from multiple schools utilizing different teaching methods and material would allow a greater ability to generalize the results to a broader audience. The sample size was small. Further studies need to be completed with an increase in sample size to view a greater spectrum of interpretative report writing styles. The same protocol was used for all participants. This factor combined with students from the same school does not take into account the possibility of collaboration among students to produce similar reports. This could skew the results toward the mean if a consensus approach was used to identify and include elements within the reports. In addition, an examination of reports on multiple different protocols could produce a more consistent finding of elements regularly included or excluded from
reports. The research question was framed with a broad set of guidelines to be inclusive of interpretative data. Minimally sufficient information for a clinician to identify the specific details of an interpretation as existing within the report may not be adequate for a layman consumer of the report to translate or understand what that information means. Writing in layman’s terms was a focus of the instructors of the course.

What material could be considered to be sufficient for inclusion in a report could also vary based on other considerations. Michaels (2006) completed a review on the ethics involved in writing assessment reports. He included IQ score reporting in his treatise. His concerns included the increase in IQ over time, as reported by Flynn (1998), changing the accuracy of the scores reported. He considered this to be ambiguous in regards to ethical considerations surrounding outdated test results. He went on later to claim a decision in reporting the information should include whether the information would harm the client. He then proceeded to state information that could benefit the client should be left out if it was not part of the referral question. This type of rationale clearly indicates alternative reasoning for the lack of inclusion of material within assessment reports while maintaining a presumed high level of ethical attention. An examination of this rationale for a lack of inclusion should be a part of any future research on the thoroughness of reports.

Another area that could have impact on investigating errors in interpretation and thoroughness of evaluation reports was elucidated by Longman (2004). He reported on comparison of the WAIS-III index scores with the overall mean of those scores. His contention was this method could reduce Type I errors caused from the complexity of multiple paired index comparisons. This complexity was the same concern reported by
Brenner (2003). Longman’s method would reduce the amount of information provided within the reports and thus affect any research seeking to reveal what was placed within interpretative evaluations and would also need to be accounted for.

Future research should include investigation of the accuracy and thoroughness of WAIS-III interpretive reports produced by clinicians. As in the research demonstrating that administration errors were shown to be similar in both students and clinicians (Kaufman & Lichtenberger, 2002), it may be expected that findings within this study on student WAIS-III interpretative reports could be reflected in research on clinician WAIS-III interpretative reports. As clinicians may be perceived to have access to seminars and professional organizations providing updates on issues regarding their areas of expertise, it may be prudent to include distinctions between levels of reported expertise and knowledge as part of the research. Assessing the underlying approach taken by the clinician could reveal insight into practices following outlined ethical considerations or recent concepts highlighted by research.

It was hoped that the findings of this study (indices comparison interpretation and subtest strengths and weakness interpretation as areas in which greater attention should be focused on by students) might highlight portions of the interpretive reports that were being missed by graduate students and that use of this data could result in instructors stressing the need for extra attention being paid to those areas shown to be neglected within the student reports. Identifying deficiencies within the reports could provide valuable feedback to instructors regarding areas of interpretation to be better elucidated and emphasized within an instructional base. The foundation for the calculated identification of specifics was taken from Assessing adolescent and adult Intelligence.

The hypotheses generated to investigate the thoroughness of graduate student generated WAIS-III reports completed as a final class assignment were believed to reflect errors found in the administration of the WAIS as reported by Kaufman and Lichtenberger (2002). It was expected that all possible variables that could be calculated were not included in the evaluation reports, areas considered to be primary to the interpretation of results would be included, indices comparisons and strengths and weaknesses within the subtests would be neglected within the interpretations, and that the students would commit computational errors resulting in incorrect interpretation.

With the focus of previous research on errors made by students and professionals in the administration and scoring of protocols, this study presents a different view. While it was noted that the scores of some portions of the protocols did vary, the presentation of the information demonstrated a close proximity to the baseline protocol scores established by the instructors of the course. In this case the mean was the correctly administered and scored protocol provided by the instructors of the course. Consideration should be given to the results indicating consistent overall findings of the students with the scores presented as a baseline (i.e. Table 2). Although the students did score the protocols higher in the verbal domain than the baseline, the difference was not generally sufficient to skew the interpretation. This means the errors presumed to be present in the
student protocols were insufficient to invalidate the interpretative findings. One possible explanation of this finding is that previously reported concerns about student administration and scoring errors (Alfonso et al., 1998; Kaufman & Lichtenberger, 2002; Sattler, Winget, & Roth, 1969; Slate, 1993; Slate & Jones, 1990) were addressed by the instructors of the course on intelligence testing.

It is up to the author of a report to define the meaning of the scatter within the IQ or index scores for the individual client. This could be seen as citing a lack of consistency within the specific index and continuing to interpret it with consideration for the score(s) that reveal a different level of specific functioning or noting the inconsistent findings could result in inconclusive interpretative findings and reporting strengths and weaknesses in that area of functioning only. Use of the Kaufman-Lichtenberger method posits that abnormal scatter within the VIQ, PIQ, or individual indices affect the unitary structure of the precepts to such a degree as to make them invalid for interpretation (Kaufman & Lichtenberger, 2002). Those abnormal findings were reported in their text and used as a basis for exclusion of interpretations as expected within this study. Failure to exclude those interpretations as expected could be deemed to penalize students for following the method they were instructed in. Based on Kaufman-Lichtenberger’s method of analysis, scatter was considered a primary finding within this study. Although it was found in 86% of the reports, only 18% of those reported on the lack of consistency revealed. This could be seen as an opportunity for instructors to emphasize the inclusion of reporting consistency/inconsistency within constructs. This could also be an area in which Longman’s (2004) proposal to format the interpretation of indexes against the overall mean may address the issue and provide an alternative to including comparative
interpretations of inconsistent findings or failing to include any interpretation (as was the considered method engaged in this study).

The student reports were deemed to be reticent when it came to inclusion of the indices comparative interpretations. Analyses to determine inclusion of these interpretative data points were sufficiently broad as to consider separate descriptions of each index in adjacent sentences as a comparison. Although broad, the inclusion of material written in this manner was presumed to be appropriate for layman’s language. However, even with a wide range of acceptable responses included in the comparisons, this area of interpretation reached a high of 80% of expected responses included (4 of 5) and 66% of total responses interpreted for the VCI/POI. According to the Kaufman-Lichtenberger method, the VCI/POI interpretation was presented as an alternative explanation to the FSIQ when there was an abnormal difference between the VIQ and PIQ (Kaufman & Lichtenberger, 2002), but the protocol used for this class assignment contained scatter in the VCI and thus it was not a required interpretation for the purpose of the study. The reason for a lack of inclusion of indices comparison was not sought. Brenner (2003) and Longman (2004) posited the comparisons were complex, subject to possible errors, and difficult to incorporate into the interpretation. This may be an area in which greater emphasis could be placed in instruction. Multiple meanings of the comparisons could be provided to students to enhance their understanding of how the comparison could look to a layman. For example, when describing a comparison of the indices in which a significant difference existed, such as between the WMI and POI, and in which the WMI was greater than the POI, could include describing the difference manifesting as being able to remember the details of what was required to complete a
task but not being able to organize the information and complete the task properly. A further hypothesis might be offered that this discrepancy in abilities could lead to some frustration for the client.

Inclusion of specific strengths and weaknesses was part of the hypothesis generation for explanation of results that was recommended. This was to enable a shift from global interpretation of indices comparison to patterns of abilities and deficits or identification of specific abilities or deficits that could be addressed within the recommendations (Kaufman & Lichtenberger, 2002). Unfortunately, while all of the protocols included a finding of weakness within the verbal domain only 63% of reports interpreted the finding. Reporting on strengths fared somewhat better considering 57% of the protocols included a finding of a verbal strength and 72% of the reports interpreted the finding. A total of 46% of reports included interpretations of verbal strengths where none were reported in the protocol and 34% reported performance domain weaknesses where none were reported in the protocol. Reporting specific strengths and weaknesses where none are found in the protocol may serve to weaken hypothesis generation of consistent patterns found in the profile generated by the scores. It could be more helpful to note a consistency in the scores denoting stability within the domain. It may assist students to consider the approach advocated by Snyder, Ritschel, Rand, and Berg (2006) when analyzing strengths and weaknesses. They reported on a predominance of referral questions framed as negatives and sought to balance assessment reports with an approach seeking to capitalize on the concept that humans are goal driven and focusing efforts on building motivation to change. Interpreting weaknesses where none were identified does not take advantage of this concept.
The final hypothesis tested in the study was the existence of computational errors. Although errors did exist, they fell below the level considered to be relevant for this study. They were found in 18% of reports, comprised less than 5% of calculations possible and resulted in only 6 changes in interpretation. This was not consistent with the findings reported in administration and scoring of protocols (Alfonso, Johnson, Patinella, & Rader, 1998; Belk, LoBello, & Zachar, 2002; Loe, Kadlubek, & Marks, 2007; Slate & Jones, 1990; Slate, Jones, & Murray, 1991) used as a base for the hypothesis. Although there were only a relative few different interpretations, those changes could have been serious for a client who was the recipient of the report. Use of a check sheet with redundancies to look for errors was investigated by Thompson and Hodgins (1994) and shown to have an effect. However, the majority of practitioners investigated indicated they would not use the system because of the additional time required. Instructors should emphasize the number of calculations involved in scoring protocols, seek resolution of problematic errors with any student displaying a pattern of computational errors, and may wish to consider use of a check sheet. Use of a redundant check sheet turned in with each protocol scored could assist students in identifying errors before writing a report.

A further factor in assessing the generalizability of this study may exist within the nature of the psychology program itself. The course evaluated was within a Psy.D. program. Harding (2007) investigated the phenomenon of clinical decision making skills and found that Psy.D. programs provided twice the decision making skill focus as Ph.D. programs. As the inclusion of information within the interpretative evaluation reports is dependent on clinical decision making regarding a diagnosis and how information may answer the referral question, this factor may be of importance. The impact of this should
not be underestimated. Gottfredson and Saklofske (2009) questioned the validity of information provided through testing for treatment consideration. They suggested that the construct measured may not be the ones users want to know or be able to use for treatment and intervening constructively in the lives of those tested. A consideration in future research should include clinical decision making skills as a variable in what information is included within reports.

Previous research indicated errors made in administration and scoring tend to continue after the first 5 administrations (Alfonso, Johnson, Patinella, & Rader, 1998; Slate & Jones, 1990). This was not an issue addressed in this study. However, if this pattern were reflected in interpretation, then providing early intervention and correction may prove beneficial. If deficiencies in interpretation reporting were noted within student reports, it may be of benefit for students to rewrite and resubmit reports in order to practice full interpretation of reports as regards the needs of the consumer. This practice could also be used to identify and improve sections of reports in which jargon makes the report unreadable for a layman. It should be noted that in the case of this study instruction did include feedback on readability of the report to reduce jargon. It may be of interest to future researchers to identify the evolution of changes in student reports.

Loe, Kadlubek, and Marks (2007) and Slate, Jones, and Murray (1991) noted that methods of teaching were not having the desired result in reducing errors in Wechsler scale administration and scoring. Intelligence assessment instructors, in conjunction with assistants, in APA accredited programs average a total of 28 hours per week on lesson activities and few instructors use the methods believed to improve administrative skills
(Cody & Prieto, 2000). Future research should look at programs in which fewer errors are being made and center research on techniques that are working rather than focus efforts on decades old methods that continue to evidence problems. Future researchers may wish to consider surveying instructors in intelligence testing utilizing the Wechsler scales to obtain some information on the rate of errors within administration and scoring to delineate programs evidencing a greater purported success rate for further investigation. This has the potential to provide insight into instructional techniques that may enhance the accuracy of testing, capitalize on methods that have been actually used, and shorten the time needed to effect change. This could maximize the effort put forth by highlighting what is working rather than make suggestions. This refocus of efforts may reduce the amount of time necessary to move methods from the trial phase to implementation.

Harvey (2006) reported on reasons assessment reports tended to be short. She stated a major reason was the lack of time. If areas of interpretation are not included within the reports of students who are training to know and understand the significance of information to be included within reports, could it be expected of them to give greater due diligence when they are under time constraints to write a report? This could be viewed as poor initial report writing habits being exacerbated by the pressure of time management when writing for consumers. Therefore, it is recommended that greater weight be given to inclusiveness of material within student reports. This could include organization of the material in a manner in which the body of the report includes all reportable findings and with the summary targeting the referral question in relation to the test results. This would serve the interests of the referring party while maintaining fiduciary responsibility to the client’s future interests.
A focus on the referral question was emphasized by Brenner (2003). Brenner’s contention that the likelihood of a Type I error by presenting a significant finding where none existed would be increased due to the comparison of multiple pieces of information through the inclusion and interpretation of too much information could easily be addressed within the presentation of the information. Comparison of multiple scales and separating that information from the body into the summary would provide shortened targeted information. In the case of the reports investigated within this study, the information was divided into initial observations (provided by the instructor), results of the testing (the area of interest targeted within this study), a summary, and recommendations. In this manner the referral question was to be answered in layman’s language, individualized, and specific to the need of the referral source as suggested by Brenner. The length of the reports was not a consideration. This was consistent with Harvey (2006) emphasizing the need for clarity over length. In reaching for this goal she recommended that the report contain enough information that it could be sufficient unto itself.

A concern then becomes that the tests given can create a baseline of functioning. If we neglect to include information, clinicians reviewing the reports at a later date will have incomplete data. We do not know how the information we collect today may be viewed in the light of unknown changes in the client. Therefore, in a referral for testing it may be of greater benefit for the long term planning of a client to be overly inclusive of information rather than providing a limited amount targeted to a singular question. This factor was noted by Wong (2006) when he addressed ethical controversies stemming from a lack of agreement on aspects of neuropsychological testing. He referred back to
research by Donders (2001a) revealing a deficiency of consistent reporting of results as evidence for a need to emphasize accurate interpretation in reports and disclosure of the information obtained. This could become more serious depending on the reason for the referral. Specifically, the WAIS-III has been evaluated for multiple impairments to provide accurate and reliable demographics for comparison (Blake, Fichtenberg, & Abeare, 2009; Taylor & Heaton, 2001; Kaufman & Lichtenberger, 2002; Wechsler, 1997a; Zhu, Tulsky, Price, & Chen, 2001). Failure to utilize uniform standards in interpretation could be considered a lack of adherence to ethical responsibilities.

Although the APA established divisions to advance scientific knowledge, guidelines for standards in testing (AERA, APA, & NCME, 1999), and boards to establish high levels of expertise. This presents the appearance of uniform standards within each area of expertise. However, it may behoove the APA to follow the lead of the British Psychological Society (BPS) in order to provide tangible evidence of clinicians meeting those standards. The BPS established the Psychological Testing Centre (PTC) as a separate organization in March 2002 (BPS, 2007). The PTC listed tests passing a review and considered to be valid for specific uses. It also had a list of assessors who passed a competency exam for each type of test. Current tests and assessors included those for occupational and educational testing. A new database was being developed for mental health tests and assessors competent to administer the tests. The public had access to this information in order to make competent decisions on which tests they may consent to and if the assessor was competent to use the tests (PTC, 2007). The establishment of a similar system by the APA could assist psychologists and the general public by establishing lists of types of tests (i.e. personality, depression, intelligence) that reach a minimum standard
of validity within each type and delineating core competency standards for assessors who
would be considered able to use those tests. This would allow clinicians and the general
public to verify competencies and ensure adherence to the APA Ethics Code standards
concerning boundaries of competence (APA, 2002) that have been obtained through
training, independent study, seminars, and supervised experience outside of formal
training.

Research on the Wechsler scales has covered the constructs, administration, scoring,
neuropsychological uses, diagnostic uses, and vocational/educational uses among others.
The reports written based on these tests direct treatment, guide rehabilitation, assist in
school placement, and can be used to determine eligibility for support. However, until
this date there has been a gap in the research. No prior investigations have been
completed on what is actually written related to the constructs measured. Competence in
interpretation of tests is an ethical imperative but without research indicating what is
actually placed into the test reports there is no basis for defining what level of
interpretation meets ethical responsibilities.
REFERENCES


Appendix A

On a standard administration of the WAIS-III the following age adjusted scores were obtained, in which a score of 100 is average:

<table>
<thead>
<tr>
<th>IQ/Index</th>
<th>Standard score</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal IQ</td>
<td>104</td>
<td>Average</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>93</td>
<td>Average</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>100</td>
<td>Average</td>
</tr>
<tr>
<td>Verbal Comprehension Index</td>
<td>108</td>
<td>Average</td>
</tr>
<tr>
<td>Perceptual Organization Index</td>
<td>97</td>
<td>Average</td>
</tr>
<tr>
<td>Working Memory Index</td>
<td>94</td>
<td>Average</td>
</tr>
<tr>
<td>Processing Speed Index</td>
<td>81</td>
<td>Low Average</td>
</tr>
</tbody>
</table>

The Full Scale IQ (FSIQ) is in the Average Range.

The FSIQ is a measure of the general level of cognitive functioning.

Because the VIQ/PIQ difference is 16 or fewer points, the mean (used to calculate scatter within the indices and for the differences among the subtests) is calculated jointly for the Verbal domain and the Performance domain.

There is abnormal scatter in the subtests of the Verbal IQ (VIQ) that should be explained. The VIQ may not be a unitary construct. Abnormal scatter is 8 or more points difference between the high score and the low score of the subtests that comprise the VIQ.

The VIQ is a measure of acquired knowledge, verbal reasoning, attention to verbal materials, and includes short-term memory abilities. The mean scaled score used for the VIQ is 10 on the scale in which an average score is 10. The highest subtest scaled score is 14. The Comprehension subtest scaled score matches the highest score. It measures projective qualities of moral reasoning, general social knowledge, normative social behavior, and experiences. The lowest subtest scaled score is 6. The Arithmetic subtest scaled score matches the lowest score. It is comprised of the ability to attend to verbal information, remember, and manipulate numbers mentally, and output that information verbally.

There is no abnormal scatter in the subtests of the PIQ. The PIQ is a unitary construct. Abnormal scatter is 8 or more points difference between the high score and the low score of the subtests that comprise the PIQ.

VIQ/PIQ Mean

<table>
<thead>
<tr>
<th>VIQ Scatter</th>
<th>Abnormal &gt;= 8 points</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

VIQ Scatter explanation and VIQ explained

<table>
<thead>
<tr>
<th>PIQ Scatter</th>
<th>0-7 points</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
The PIQ is a measure of visual-spatial ability, attention to detail, visual-motor integration, and fluid reasoning.

The VIQ is greater than the PIQ. This means that the verbal reasoning abilities are stronger than the nonverbal, visual-spatial reasoning abilities. The difference is 9 - 11 points, significance p < .05. The abnormal scatter may affect the unitary construct of the FSIQ. Therefore the FSIQ may not be an accurate representation of the true score, the overall level of intellectual functioning. All abnormal scatter should be explained. It may not be possible for the FSIQ to give a true representation of the overall level of intellectual functioning.

The VCI is a measure of verbal reasoning and conceptualization, acquired knowledge, and the ability to express ideas verbally. There is no abnormal scatter in the subtests of the VCI. It is a unitary construct. Abnormal scatter is 5 or more points difference between the high score and the low score of the subtests that comprise the VCI.

The POI is a measure of nonverbal reasoning, visual-spatial processing, visual-motor integration, and attention to detail. There is abnormal scatter in the subtests of the POI that should be explained. The POI may not be a unitary construct. Abnormal scatter is 6 or more points difference between the high score and the low score of the subtests that comprise the POI. The highest subtest scaled score of the POI is 13. The Matrix Reasoning subtest scaled score matches the highest score in the POI. It measures visual information processing, visual-spatial reasoning, abstract reasoning, and visual organization. The lowest subtest scaled score of the POI is 6. The Picture Completion subtest scaled score matches the lowest score in the POI. It measures visual processing, long-term visual memory, and processing speed.

Although there is a significant difference between the VIQ and PIQ, it alone is not sufficient to affect the unitary construct of the FSIQ, therefore the FSIQ may be a unitary construct and likely to be an accurate representation of the true score, the overall level of intellectual functioning. A significant difference is considered to be 9 - 11 points, p < .05. Although there is a significant difference between the Verbal Comprehension Index (VCI) and Perceptual Organization Index (POI), it alone is not sufficient to affect the unitary construct of the FSIQ, therefore the FSIQ may be a unitary construct and likely to be an accurate representation of the true score, the overall level of intellectual functioning. A significant difference is considered to be 10 - 12 points, p < .05.
Due to the abnormal scatter in the subtests of the Working Memory Index (WMI), an interpretation of the Verbal Comprehension Index (VCI)/WMI relationship (verbal reasoning abilities and short-term memory abilities) cannot be made. The abnormal scatter may also affect the interpretation of the VIQ. The common subtests between the VIQ and the WMI are the Arithmetic and Digit Span subtests.

Due to the abnormal scatter in the subtests of the Performance Organization Index (POI), an interpretation of the POI/Performance Speed Index (PSI) relationship cannot be made. The abnormal scatter may also affect the interpretation of the PIQ. The common subtests between the PIQ and the POI are the Picture Completion, Block Design, and Matrix Reasoning subtests.

The VCI is greater than the PSI. The discrepancy between the indices indicates that there are stronger verbal reasoning abilities, to understand, mentally manipulate information, and express ideas than abilities of response speed, speed of thinking, and motor speed. This could manifest as a relatively greater ability to understand and explain a task than to perform one. A significant difference is 12 or more points, p < .05.

The Vocabulary subtest is a relative strength. A strength is 3 or more points above the mean. This subtest is 3 points above the mean. It is considered to tap crystallized intelligence. It measures language development, word knowledge, and previously input knowledge. A high score may reflect school learning, reading ability, a rich early environment, cultural opportunities, foreign language background, or intellectual ambitiousness and striving.

The Comprehension subtest is a relative strength. A strength is 3 or more points above the mean. This subtest is 4 points above the mean. It is considered to tap crystallized intelligence. It measures projective qualities of moral reasoning, general social knowledge, normative social behavior, and experiences. A high score may reflect social maturity and judgment, moral sense, concrete/conventional thinking, or coping abilities.
The Arithmetic subtest is a relative weakness. A weakness is 3 or more points below the mean. This subtest is 4 points below the mean. It is considered to tap crystallized intelligence. It is comprised of the ability to attend to verbal information, remember, and manipulate numbers mentally, and output that information verbally. A low score may reflect problems with anxiety, attention span, distractibility, concentration, school achievement, learning disabilities, problems working under time constraints, mathematical achievement, quantitative reasoning, or problems with short-term memory.

The Matrix Reasoning subtest is a relative strength. A strength is 3 or more points above the mean. This subtest is 3 points above the mean. It is considered to tap fluid intelligence. It measures visual information processing, integration and sequencing ability, nonverbal reasoning, visual-spatial reasoning, abstract reasoning, and visual organization. A high score may reflect field independent cognitive style, flexibility, motivation level, persistence, or decisiveness.

The Picture Completion subtest is a relative weakness. A weakness is 4 or more points below the mean. This subtest is 4 points below the mean. It is considered to tap both crystallized and fluid intelligence. It measures visual processing (accurate visual perception/scanning to recognize and distinguish essential from non-essential details, and understand the context), long-term visual memory, output information either verbal or motor, and processing speed. A low score may reflect concentration deficits, lack of attention to environmental details, problems working under time constraints, indecision, negativism, or visual agnosia.

The Digit Symbol-Coding subtest is a relative weakness. A weakness is 4 or more points below the mean. This subtest is 4 points below the mean. It is considered to tap fluid intelligence. It measures applied non-verbal learning, psychomotor speed, visual short-term memory, visual perception of stimuli, visual sequencing, the ability to input information visually by using a key, manipulate and transfer the information with short-term memory, and output that through fine motor skills. A low score may reflect visual impairment, learning disabilities, illiteracy, distractibility, anxiety, lower ability to work under time constraints, infrequent use of paper and pencil skills, perfectionism, or problems with clerical speed and accuracy, persistence, following directions, or short-term memory ability.
All other subtest scores are consistent within the expected ability level. Age adjusted subtest scaled scores, in which a score of 10 is average, appear in the table below:

<table>
<thead>
<tr>
<th>Verbal Subtests</th>
<th>Scaled Score</th>
<th>Performance Subtests</th>
<th>Scaled Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>13</td>
<td>Picture Completion</td>
<td>6</td>
</tr>
<tr>
<td>Similarities</td>
<td>12</td>
<td>Digit Symbol-Coding</td>
<td>6</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>6</td>
<td>Block Design</td>
<td>8</td>
</tr>
<tr>
<td>Digit Span</td>
<td>12</td>
<td>Matrix Reasoning</td>
<td>13</td>
</tr>
<tr>
<td>Information</td>
<td>10</td>
<td>Picture Arrangement</td>
<td>10</td>
</tr>
<tr>
<td>Comprehension</td>
<td>14</td>
<td>Symbol Search</td>
<td>8</td>
</tr>
<tr>
<td>Letter-Number</td>
<td>9</td>
<td>Object Assembly</td>
<td>6</td>
</tr>
<tr>
<td>Sequencing</td>
<td>9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix B

<table>
<thead>
<tr>
<th>ID #</th>
<th>209</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✔</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Was interpretive information reported on:</th>
<th>Correct math for interpretation?</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSIQ</td>
<td>VIQ/PIQ 0</td>
</tr>
<tr>
<td>VIQ</td>
<td>VCI/POI 0</td>
</tr>
<tr>
<td>PIQ</td>
<td>VCI/WMI 0</td>
</tr>
<tr>
<td>VIQ/PIQ split</td>
<td>POI/PSI 0</td>
</tr>
<tr>
<td>Scatter</td>
<td>VCI/PSI 0</td>
</tr>
<tr>
<td>VCI</td>
<td>POI/WMI 0</td>
</tr>
<tr>
<td>WMI</td>
<td>WMI/PSI 0</td>
</tr>
<tr>
<td>PSI</td>
<td></td>
</tr>
<tr>
<td>POI</td>
<td></td>
</tr>
<tr>
<td>VCI/POI</td>
<td></td>
</tr>
<tr>
<td>VCI/WMI</td>
<td></td>
</tr>
<tr>
<td>POI/PSI</td>
<td></td>
</tr>
<tr>
<td>VCI/PSI</td>
<td></td>
</tr>
<tr>
<td>POI/WMI</td>
<td></td>
</tr>
<tr>
<td>WMI/PSI</td>
<td></td>
</tr>
</tbody>
</table>

Verbal Strengths

Verbal Weaknesses

Performance Strengths

Performance Weaknesses
Appendix C

PACIFIC UNIVERSITY
INFORMED CONSENT TO ACT AS A RESEARCH PARTICIPANT

Graduate Student WAIS-III Interpretative Evaluation Reports: How Thorough Are They?

Investigators Contact Information

Principal Investigator:
David B. Moore
Pacific University
School of Professional Psychology
(503) 261-9396
moor1699@pacificu.edu

Faculty Advisor:
Lisa Christiansen, Psy.D
Pacific University
School of Professional Psychology
(503) 352-2627
lisac@pacificu.edu

1. Introduction and Background Information

You are invited to be in a research study of graduate student WAIS-III interpretative evaluation reports because you are enrolled in the GPSY 821 course. Please read this form carefully and ask any questions you may have before agreeing to be in this study.

David Moore M.S. and Lisa Christiansen Psy.D are conducting this study. The purpose of this study is to investigate graduate student Wechsler Adult Intelligence Scale-Third Edition (WAIS-III) reports. Student practice reports will be compared to a computer-based test interpretation (CBTI) of the same WAIS-III scores. This will be used to identify if the data has been interpreted correctly and completely. Any differences between the CBTI reports and the graduate student reports will be recorded. Any areas of difference noted within the reports will be gathered and analyzed. This information will not be used in grading of assignments.

2. Study Location and Dates

The study is expected to begin April, 2008, and to be completed by April, 2008. The location of the study will be at the College of Health Professions Campus, Hillsboro.
3. Procedures

If you agree to be in this study, we will ask you to provide a copy of a practice WAIS-III evaluation report that you have written as a class assignment.

4. Participants and Exclusion

Only participants who meet the following conditions will be included in the study: Pacific University psychology graduate students who are currently enrolled in an intelligence assessment course at the School of Professional Psychology (SPP). Participants who do not meet the above criteria will be excluded from the study. You will not receive payment or compensation for your participation.

5. Risks and Benefits

There are risks to participating in this research. There is a minor risk that a research assistant may know the author of a report. The research assistant will black out the identifying information before forwarding the report for data analysis. The research assistant(s) will be trained in the rules governing confidentiality. The risk to any author of a report is low. There are no direct benefits to the study group.

6. Alternatives Advantageous to Participants

Not Applicable

7. Participant Payment

You will not receive payment or compensation for your participation.

8. Promise of Privacy

The records of this study will be kept private. The records of this study will be kept private. Data will be kept in a locked file cabinet in the home office of the primary investigator. If the results of this study are to be presented or published, we will not include any information that will make it possible to identify you as an individual. This informed consent form will be kept separately from any data we collect. If the results of this study are to be presented or published, we will not include any information that will make it possible to identify you as an individual.

9. Voluntary Nature of the Study

Your decision whether or not to participate will not affect your current or future relations with Pacific University, your grade, or your standing in the class. If you decide to participate, you are free to withdraw at any time without prejudice or negative consequences.
10. Compensation and Medical Care

Not Applicable

11. Contacts and Questions

The researcher will be happy to answer any questions you may have at any time during the course of the study. The researcher can be reached at (503) 261-9396 or by email at moor1699@pacificu.edu. If you are not satisfied with the answers you receive, please call Pacific University’s Institutional Review Board, at (503) 352 – 2112 to discuss your questions or concerns further. All concerns and questions will be kept in confidence.

12. Statement of Consent

I have read and understand the above. All my questions have been answered. *I am 18 years of age or over and agree to participate in the study.* I have been given a copy of this form to keep for my records.

<table>
<thead>
<tr>
<th>Participant’s Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant contact information:</td>
<td></td>
</tr>
<tr>
<td>Street address:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Telephone:</td>
<td></td>
</tr>
<tr>
<td>Email:</td>
<td></td>
</tr>
</tbody>
</table>

This contact information is required in case any issues arise with the study and participants need to be notified and/or to provide participants with the results of the study if they wish.

Would you like to have a summary of the results after the study is completed?  
___Yes  ____No

<table>
<thead>
<tr>
<th>Investigator’s Signature</th>
<th>Date</th>
</tr>
</thead>
</table>
Appendix D

<table>
<thead>
<tr>
<th>Points difference required for differential interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIQ/PIQ 9 @ .05 Split @ 17</td>
</tr>
<tr>
<td>VCI/POI 10 @ .05 Abnormal @ 19</td>
</tr>
<tr>
<td>VCI/WMI 10 @ .05</td>
</tr>
<tr>
<td>POI/PSI 13 @ .05</td>
</tr>
<tr>
<td>VCI/PSI 12 @ .05</td>
</tr>
<tr>
<td>POI/WMI 11 @ .05</td>
</tr>
<tr>
<td>WMI/PSI 13 @ .05</td>
</tr>
</tbody>
</table>