THE EFFECTS OF EXPERIENCE AND TEST REVISION ON THE RATE OF PRACTITIONERS' CLERICAL ERRORS ON THE WISC-R AND WISC-III

by

ROBERT MARK KLASSEN

B.Ed., University of British Columbia, 1987

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES
Department of Educational Psychology and Special Education

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

APRIL, 1994

© Robert Mark Klassen, 1994
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of **Educational Psychology and Special Education**

The University of British Columbia
Vancouver, Canada

Date **April 19, 1994**

DE-6 (2/88)
The purpose of this study was to investigate the effect of test revision and examiner experience on the rate of examiner clerical error on the Wechsler Intelligence Scale for Children - Revised (WISC-R) and the Wechsler Intelligence Scale for Children - Third Edition (WISC-III). A total of seven school psychologists provided a sample of 252 protocols: 18 from each psychologist for the WISC-R, and 18 from each psychologist for the WISC-III. The errors tabulated consisted of clerical errors only, that is, addition of subtest raw and scaled scores, computation of chronological age, transformation from tables of raw score to scaled scores, and transformation of scaled scores to IQ scores. Clerical errors were found on 38% of WISC-R protocols, and 42% of WISC-III protocols. Errors caused by incorrect addition of raw scores were the most common error type. No significant difference was found between the two tests for any of the types of clerical errors. A positive but statistically non-significant correlation was found between examiners' years of experience on the WISC-R and number of errors made.

On the WISC-III only, a comparison was made between the clerical error rate at the beginning of usage of the test, and the clerical error rate in two time periods in the 12 months following the beginning of usage of the test. With overall errors, no significant difference was found between the three time periods. However, examiners made significantly more Full Scale IQ-changing
errors at the beginning of WISC-III administration than they did in
the third six-month time period. It was suggested that test
publishers provide clear directions on the actual protocol to test-
users to help lower the rate of clerical errors. It was also
suggested that because of the high rate of clerical errors at the
beginning of the use of the WISC-III, practitioners need formal
retraining on a test when it is revised.
# Table of Contents

ABSTRACT ......................................................................................................................... ii

LIST OF TABLES .................................................................................................................. vi

LIST OF FIGURES .............................................................................................................. vii

ACKNOWLEDGEMENTS ....................................................................................................... viii

CHAPTER

I INTRODUCTION ................................................................................................................. 1

  Purpose of the Study ........................................................................................................... 2
  Error of Measurement ....................................................................................................... 2
  Previous Studies ................................................................................................................ 5
  Examiner Differences ....................................................................................................... 8
  Test Revisions .................................................................................................................. 9
  Protocol Changes ............................................................................................................. 10
  Questions and Hypotheses ............................................................................................... 11
  Scope of Study .................................................................................................................. 12

II LITERATURE REVIEW ..................................................................................................... 14

  Introduction ..................................................................................................................... 14
  Studies with Graduate Students ....................................................................................... 14
  Training Programs .......................................................................................................... 16
  Unscored Protocols ......................................................................................................... 17
  Completed Protocols ....................................................................................................... 22
  Definitions of Error .......................................................................................................... 24
  Test Revisions .................................................................................................................. 26
  Examiner Characteristics ............................................................................................... 27
  Assessment Practices ..................................................................................................... 30
  Summary .......................................................................................................................... 30

III METHODOLOGY .......................................................................................................... 34

  Sample ............................................................................................................................. 34
  Characteristics of Examiners ......................................................................................... 36
  Characteristics of Examinees ......................................................................................... 37
  Assumptions .................................................................................................................... 38
  Procedure ......................................................................................................................... 40
### LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Examiners' Years of Experience</td>
<td>37</td>
</tr>
<tr>
<td>2</td>
<td>Characteristics of Examinees</td>
<td>38</td>
</tr>
<tr>
<td>3</td>
<td>Number and Proportion of Protocols with Errors</td>
<td>43</td>
</tr>
<tr>
<td>4</td>
<td>Number of Total Errors by Test per Examiner</td>
<td>46</td>
</tr>
<tr>
<td>5</td>
<td>Location of Clerical Errors by Test</td>
<td>49</td>
</tr>
<tr>
<td>6</td>
<td>Error Categories Across Test</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>Comparison of Error-Types</td>
<td>53</td>
</tr>
<tr>
<td>8</td>
<td>WISC-III Overall Errors Over Time</td>
<td>55</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

FIGURE

1 Number of Protocols with Errors by Test .................. 44
2 Errors by Examiners ............................................. 47
3 Clerical Errors by Category ...................................... 51
4 WISC-III Mean Errors ............................................. 56
ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to Dr. Nand Kishor, my research supervisor, for his help and direction throughout the course of this research. Much guidance was given to help focus this paper on each successive draft.

Thanks also to Dr. Harold Ratzlaff who helped enlighten some murky research questions, and who provided the original impetus to begin the program and to persevere to the end.

I also am appreciative of the feedback of Dr. Kim Schonert-Reichl, who provided valuable critical analysis on several drafts.

Finally, I wish to acknowledge the support, patience and love of my family: Andrea, Danielle, and especially Lenore, who tolerated my 4:30 a.m. awakenings, and who encouraged me to continue to the end.
Chapter 1

Introduction

The Wechsler Intelligence Scales are among the most commonly used individual intelligence tests within educational and clinical psychology settings (Slate & Chick, 1989). In the school setting, critical educational placement and planning decisions are based on the data derived from these tests, which are assumed to be administered in a standardized manner, and scored with minimal error. In fact, the American Psychological Association's Standards for Educational and Psychological Testing (1985) suggests that the scoring of intelligence tests should be completely free of errors.

Numerous studies have shown, however, that test protocols are far from error-free, and that virtually all examiners make errors (e.g., Sherrets, Gard, & Langner, 1979; Slate, Jones, Coulter, & Covert, 1992). Unavoidably, the conclusion must be drawn that critical educational placement and planning decisions are sometimes based on the results from inaccurately-scored tests. These test results, consequently, are not as true a reflection of the child as they might seem to be. Any information, then, that helps professionals working in the field avoid errors in the scoring of these tests, is potentially of great assistance in achieving the ideal of error-free testing.
Purpose of the Study

The Wechsler Intelligence Scale for Children - Revised (WISC-R) has recently (1991) been revised as the Wechsler Intelligence Scale for Children - Third Edition (WISC-III). These tests are among the most widely used tests of mental ability for children aged 6 to 16. This study investigated how the nature and rate of examiner error change from one test, the Wechsler Intelligence Scale for Children - Revised (WISC-R), to its recently published revision, the Wechsler Intelligence Scale for Children - Third Edition (WISC-III).

Error of Measurement

Standardized tests - such as the WISC-R and its revision, the WISC-III, generally report a standard error of measurement ($SE_M$) which is an estimate of the amount of error involved in an examinee's obtained test score. This $SE_M$ accounts only for content sampling errors, and, in the majority of tests, including the WISC-R and WISC-III, is measured through the use of the Spearman-Brown correction of split-half correlations. In this approach to reliability, the test is divided into two approximately parallel forms or components which are then correlated. The resulting correlation, when adjusted for the greater length of the composite through the use of the Spearman-Brown correction, represents the reliability of the test.

On the WISC-R and WISC-III, the reliability coefficients were first calculated for each individual subtest through the split-half method and Spearman-Brown correction formula (except for Coding,
and on the WISC-III, Symbol Search, which are speeded tests, for which split-half methods are inappropriate). The individual subtest reliability coefficients were then used to compute the overall test reliability, through the use of Nunnally's formula for the reliability of a composite of several tests (Nunnally, 1978, p. 246).

On most standardized tests, the actual scoring of the test by the examiner is assumed to be almost completely objective and error-free. In the case of many tests, no measure of examiner error or of interrater reliability is reported, although this has been recommended: "There is as much a need for a measure of scorer reliability as there is for the more usual reliability coefficients" (Anastasi, 1988, p. 108).

The WISC-R manual, published in 1974 does not include any information on interscorer/interrater reliability or on how examiner error might affect test scores. The WISC-III manual, published in 1991 does offer some improvement over the WISC-R in the area of agreement between scorers. In the new manual, results are reported from a test-development standardization study on interscorer agreement for four of the thirteen subtests, with coefficients generally quite high, ranging from .90 for the Comprehension subtest, to .98 for the Similarities and Vocabulary subtests.

The data from the WISC-III standardization study on interscorer agreement were drawn from four examiners who each scored 4 subtests from 60 protocols. The examiners, however, were
presumably chosen from among the specialists working on the standardization study, whose sole duty was accurate administration of the WISC-III. The WISC-III manual states that in the standardization study “all protocols were double-scored by trained scorers... The computer flagged any discrepancy between scorers ... questions were answered by in-house experts. Scorers received individual feedback on a daily basis and group feedback weekly” (1991, p. 28-29).

These highly-specialized standardization examiners can hardly be considered representative of practitioners working in the field. The practitioner uses a number of measurement instruments, focuses less on the test than on the subject or client, and has a great number of accompanying responsibilities. Moreover, in most cases, practitioners are not formally re-trained on revisions of tests. The protocols used in the standardization study - double-checked and closely scrutinized - also cannot be considered representative of the protocols in the field where protocols are typically scored once only.

In the WISC-III test manual (1991) only a brief mention is made of research that suggests examiner error may contribute significantly to making test scores less reliable. The admission is made that "errors are common ... and may lead to misleading scores" (p. 57) but, surprisingly, apart from a few brief suggestions on how to avoid clerical errors (record scores legibly, and check the
addition for raw scores and scaled scores) more information on the nature or degree of these errors is not given.

Although the reported internal reliability and interscorer reliability coefficients in the WISC-III manual are quite high, the sample of examiners used in the standardization study does not accurately reflect the population of practicing professionals. Also, the reported interscorer agreement coefficients do not account for errors directly attributable to the examiner giving the test, but only for scoring four of the subtests. The end result of not attempting to remedy the fact that examiner errors affect test scores is that "important decisions will continue to be made on the basis of inappropriately scored IQ measures" (Miller & Chansky, 1973, p. 344).

Previous Studies

The question of examiner errors on standardized intelligence tests has been the subject of considerable research. Many researchers have circumvented the difficulty of gaining access to confidential data from practitioners (e.g., school and clinical psychologists, psychiatrists) by using data drawn from graduate students enrolled in training courses for administration of standardized tests (e.g., Slate & Chick, 1989; Stewart, 1987). This information is particularly useful in shedding light on training issues and on errors made by beginning testers. Training programs in graduate schools have been designed as a result of these studies (Conrad, 1991).
The generalizability of the data on errors made by graduate students to errors made by practicing professionals, however, is suspect - the nature of testing in the field is vastly different from that of testing by students in graduate school. Those practicing in the field are quite likely to view their own administration of any particular test as incidental; that is, as a means of gaining more information about a student or a client, and not as an end in itself, which is quite frequently the case in graduate training courses. At the same time, the greater experience and familiarity a professional has with a test may result in increased skill in avoiding error and working with an instrument. In any case, the testing practices of graduate students are not likely representative of the testing practices of examiners working in the field. Although studies investigating student errors are potentially useful, "investigation of the examiner errors made by practicing professionals is sorely needed" (Slate, Jones, Coulter, & Covert, 1992, p. 78).

Some of the earlier studies did investigate the error rate of practicing professional psychologists by having the examiners score completed (i.e., filled-in) but unscored protocols (e.g., Miller & Chansky, 1972; Oakland, Lee, & Axelrad, 1975). These studies usually depended on a volunteer sample of psychologists who responded to a general mail-out to a particular professional body (e.g., National Association of School Psychologists). A number of sampling problems arise in studies of this nature, however, because the data are collected from volunteer subjects.
Several limitations exist when data are collected in this manner. Firstly, the two studies cited above did not have a high rate of return: 23.5 and 32%, respectively. With these disappointingly low return rates, the sample loses its randomness may be less representative of the population. The results from studies of this nature which place relatively heavy demands on the subjects inevitably suffer from the bias inherent in volunteer samples.

Secondly, little is known regarding the degree to which attention to scoring details changes for examiners anonymously scoring protocols on their own time rather than on the job in a clinical situation. Relatedly, the scoring of an already completed protocol is a contrived situation, with the subject quite likely aware of the nature of the study. Because this approach may not produce the same results as having an examiner filling out and scoring an actual protocol while administering the test, the generalizability of data obtained in this manner is limited.

Finally, because the researcher is practically limited to having only one or two protocols scored by the volunteer scorers, the generalizability of the results is questionable. Test protocols vary a great deal in their ease or difficulty of scoring, and one or two protocols may not well represent all protocols. It seems then, that in order to gain the most generalizable and accurate information, actual completed protocols from those practicing in the field should be analyzed. And because the rate and nature of examiner error on
one test may not be generalizable to errors on another test, it is imperative that each widely-used test (and its revisions) should be investigated.

**Examiner Differences**

The differences among examiners may result in differences in their error rate on standardized tests. Studies have been completed investigating the relationship between rate of examiner error and gender (Hanley, 1977; Levenson, Golden-Scaduto, Aiosa-Karpas, & Ward, 1988), education (Hanley, 1977; Levenson et al., 1988; Sherrets, Gard, & Langner, 1979), expectancy (Egeland, 1969) and experience (Conner & Woodall, 1983). Of these three areas, one might expect that an examiner's experience and practice on a test would have the strongest impact on the rate of error. As expected, Conner and Woodall (1983) found a significant decrease in the number of errors on the WISC-R made by examiners as they gained experience. However, the examiners in their study were student-examiners enrolled in a graduate training program, and cannot be considered representative of practitioners in the field.

Although little research has been done in the area of examiner experience and error, and more is clearly needed, it is likely that the error rate of practitioners changes as they gain experience working with any particular test. And though one would predict a lowering of the rate of examiner error on a graduate student's test protocols as the result of instruction and experience, this pattern may not be the same with practitioners, who may be less concerned with the actual
administration and scoring of a test, and more concerned with the results and their interpretation.

**Test Revisions**

There is no question that test revisions bring about changes in subjects' test results. For example, subjects usually score lower on newer standardized tests than on older ones, as was the case in the norm changes in the revision of the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) to the Wechsler Preschool and Primary Scale of Intelligence - Revised (WPPSI-R) as has been detailed (Kaufman, 1990; Sattler, 1991). There is also little question that the format of a test or questionnaire influences a subject's or interviewee's response pattern: “Good design is crucial for producing a reliable and valid questionnaire. Respondents feel less intimidated by a questionnaire which has a clear layout and is easy to understand, and take their task of completing the questionnaire more seriously” (Rust & Golambok, 1989, p. 155).

In the same way that testing subjects or interviewees are influenced by the format of a questionnaire, it is also likely that the format and layout of a test protocol influences the response pattern and scoring accuracy of an examiner. A test revision, then, with its corresponding revision of the test protocol, will see changes in the response pattern of examiners, which may include a change in the rate and pattern of examiners' clerical errors.
Protocol Changes

Changes to the WISC-R protocol had been strongly recommended in order to reduce the rate of examiner's clerical error:

The present investigators believe that a revision in the WISC-R Record Form must be considered by the test's publisher, the Psychological Corporation. Such a revised record form might include directions to the psychologist to verify and recheck all response scoring, mathematical, administrative, IQ, and birth computations (Levenson et al., 1988, p. 663).

Although an alternate and expanded protocol was subsequently published for the WISC-R, it did not follow any of the suggestions made by Levenson et al., but simply provided more space for recording clinical observations.

The publishers of a test undergoing revision, such as the WISC-R to the WISC-III, have an opportunity to improve the format of the test's protocol, an opportunity to make it less prone to examiner errors. Although the test's publisher does not explicitly claim that the Record Form or test protocol has been revised in order to lower the rate of examiner error, it does claim in the manual (1991) that "changes from the WISC-R ... have made it easier for the examiner to use" (p.iv), and that "A final goal of WISC-III development was improvement of ... administration " (p. 12). Unfortunately, no empirical evidence is cited to back these claims. Although perhaps true for administration of the test itself, the WISC-III test protocol
does not appear to be easier to use and less prone to examiner error. In fact, on the surface, the protocol appears to be more complicated and less clear, especially in the section where raw scores are to be converted to scaled scores.

Questions and Hypotheses

What then, are the changes brought about in examiner error by the revision of the WISC-R and its protocol? To what extent is the WISC-III test and protocol less prone to examiner clerical error, and is this shown by a lowering of the rate of examiner error? How does the examiner's experience, or lack of experience, on a new test - the WISC-III - affect the rate of clerical errors made?

Most of the changes on the WISC-III test protocol appear to be changes in layout only, and are not seen to have been made for the purpose of lowering examiners' clerical errors.

*Hypothesis #1:* The rate of practitioners' clerical error on completed and scored WISC-III protocols will be equal to their rate of error on the WISC-R.

Little research has been completed in the area of the effects of a practitioner's experience on error rate. Past studies show that graduate students' error rates decrease with experience. This may not be true, however, for practitioners who approach testing in a different manner. And although one would intuitively expect an examiner's errors to decrease with experience, it is possible that as
an examiner becomes more comfortable with using a test, the clerical error rate of that examiner increases, as carelessness sets in. Psychologists with PhDs have been shown to have a higher rate of examiner error than Masters-level psychologists (Levenson et al., 1988; Ryan, Prifitera & Powers, 1983).

_Hypothesis #2:_

*Practitioners' clerical error rates will not change significantly with increasing experience on the WISC-III.*

**Scope of Study**

This study investigated the nature and rate of examiners' clerical errors on the WISC-R and WISC-III. Comparisons of error patterns between the two tests were made, and the relationship between examiner experience on the WISC-III and rate of clerical error was studied.

Examiner error, for the purpose of this study, consisted of clerical errors as defined by Sherrets et al. (1979): addition of subtest raw and scaled scores, computation of chronological age, and transformation from tables. The data for this study were drawn from analysis of actual completed test protocols from school psychologists in a large suburban school district. Because the data were drawn from practitioners' completed protocols, the results are likely to be representative of present practice in the field.
Information derived from this study will enable users of psychoeducational tests to have a better understanding of their own potential clerical errors on the WISC-III, and may result in improved practice in the area of re-training of professionals for test revisions, which in turn would result in the increased reliability of the tests. Improvements in testing practices in educational settings should result in the improvement of the accuracy and quality of planning and placement decisions.
Introduction

Although a look through the manuals of standardized tests might suggest that examiner error is not a prevalent problem, or at least that it does not have any serious impact on a test's reliability, many researchers have found substantial evidence that suggests that errors made by examiners are an important source of error, not covered by merely reporting the $SEM$ of a test. These studies regarding the nature and rate of examiner error have utilized many different designs, used both graduate students and practitioners of various stripes, investigated errors made on actual completed protocols and on fabricated protocols, and have looked at a variety of error types on a collection of different tests.

Studies with Graduate Students

The majority of the previous studies have investigated the errors made by graduate students enrolled in test-administration courses. Obtaining raw data from graduate students is frequently easier than from practitioners who may not volunteer as readily to take part in a study which reflects on their professional skills. Many of the earliest studies followed this route. Miller, Chansky and Gredler (1970) had 32 Clinical and School Psychology graduate students score a single completed WISC protocol and found the resulting Verbal and Performance scale scores to vary by as much as 21 points and the Full Scale IQs by as much as 17 points.
Warren and Brown (1972) examined the first three and last three completed protocols of 40 student examiners in a graduate-level psychology program. Of the 240 protocols, there were 89 cases in which Full Scale IQs were changed (according to the authors, the number would have been much higher had not some errors which added points not cancelled out others which deducted points). On the WISC protocols, IQ changes ranged from 1 to 16 points, with only 2 of the 240 protocols error-free. The mean IQ change after correction was 3.32 points for the WISC, and 4.77 points for the Binet. There was no significant change in error rate from the beginning to the end of the training period. From the above studies, it is clear that scoring and administration errors are common among graduate students.

Without question, data are more easily obtained from graduate students than from practicing professionals. However, the generalizability of the results from studies involving graduate students to practitioners working in the field is questionable. In training programs, at least in the initial phases, the focus of students and instructors is, naturally, on the actual administration and scoring of the test. In the field, however, the focus is more likely to be on the results of the test, and competent administration and scoring of the test is assumed. This is not to say that studies which look at novice examiners are not valuable - several studies using student examiners have had important implications for the training of test users in graduate psychology programs, training
which, currently, is widely discrepant from one institution to another.

**Training Programs**

Conner and Woodall (1983) examined students' errors in relation to a training program that provided the students with written and verbal feedback on types and frequencies of errors. They concluded that:

Response Scoring, Mathematical, and IQ Errors remain a source of variance, affecting the reliability and validity of the WISC-R. Further research is needed in this area to determine if there are any effective teaching methodologies that can significantly decrease the number of Response Scoring errors (p. 378).

Slate and Jones (1989) challenged the conventionally-held wisdom that testing courses should consist mostly of opportunities to practice test administrations, and that instructional time should be secondary: "In this study, students' performance improved only as a result of detailed classroom instruction. ... Neither group improved during practice administrations" (p. 409). In a subsequent study the investigators stated that "Five practice administrations were not sufficient to improve student accuracy in scoring and 10 practice administrations improved student accuracy only slightly. Even on the 10th administration error rates were high" (Slate & Jones, 1990c, p. 139).
It seems, then, that traditional training programs often fail to prepare graduate students to administer and score standardized tests in a competent manner. Russ (1978) outlined the causes for the gap between academic training in psychological assessment and the comparatively higher standards established by supervisors in internship settings. Other studies have offered more detailed and structured training programs or methods for systematic observations of graduate student assessment behavior (Strein, 1984). Blakey, Fantuzzo, Gorsuch and Moon (1987) found that a peer-mediated, competency-based training package, which uses student peers to give detailed feedback to examiners, significantly lowered the error rate for student examiners. Conrad (1991) constructed a comprehensive performance checklist, the Criteria for Competent WISC-III Administration - Revised (CCW3A-R) to be used to aid students in locating and avoiding their most frequent errors.

It is clear that useful information - information that may result in improved training programs - may be derived from studying the errors of student-examiners. However, the population which is of most concern is not graduate students, but professionals who use standardized tests in order to gain greater understanding and to make decisions.

Unscored Protocols

A number of studies have looked at the issue of examiner error by having psychologists and graduate students score test protocols which have been filled in but not scored. The intent of this type of
study is that the examiner will display typical behavior in scoring the protocol. In a study by Perot (1992) investigating the cognitive processes behind scoring decisions, eight school psychologists scored the Verbal Scale subtests from a completed but unscored WISC-R protocol. It was found that the psychologists produced Verbal IQ scores which varied by as much as 11 points.

Ryan, Prifitera and Powers (1983) used practicing psychologists as well as graduate students in their study of examiner error on the Wechsler Adult Intelligence Scale - Revised (WAIS-R). They had the 39 examiners score two protocols and found that the standard deviations of the Full Scale IQ scores averaged about 1.95. In comparison, the WAIS-R manual reports a $SEM$ of 2.53 for Full Scale IQ (On the WISC-R, the reported Full Scale $SEM$ is 3.19; on the WISC-III, 3.2). In an earlier study, Miller and Chansky (1972) sent a single completed WISC protocol to 200 doctoral-level school or clinical psychologists. From the 64 returned protocols it was found that Verbal and Performance IQ's varied by as much as 23 points and Full Scale IQs by as much as 17 points.

It must be questioned, however, just how seriously the responding sample treated their task in the study. In a follow-up phase of the study, Miller and Chansky sent a letter in which they offered to discuss with the the responding psychologists the types of errors that they had made. Out of the 64 volunteers, only 1 psychologist responded to the offer. Miller and Chansky concluded that "If the apathy discovered in this brief follow-up is
representative of the field at large, we can assume that important decisions will continue to be made on the basis of inappropriately scored IQ measures" (p. 345). The difficulty in knowing the representativeness of the sample to the population is raised by Miller and Chansky themselves: it is unknown if the results are representative of the field at large when a volunteer sample of 32% (64 out of 200) responded to the study, and 1.6% (1 out of 64) responded to the follow-up offer. Another difficulty with this type of study is the representativeness of the to-be-scored protocol. When only one protocol is used, it is not necessarily a typical protocol, but may be much easier or much more difficult to score than most.

In order to overcome the problem of non-representative single protocols, some researchers have several scorers score two or more protocols either varying in examinee ability level or in the expected ease and difficulty of scoring the protocol. Whereas one protocol might not well-represent the population of protocols, two or three can better cover the range of ease or difficulty of scoring, or examinee ability. Behind this type of study is the idea that examinee differences reflected in the protocol may result in differences in error rate for examiners.

Oakland, Lee, and Axelrad (1975) sent three completed but unscored actual (as opposed to fabricated) WISC protocols reflecting three different levels of intellectual abilities (below average, average and above average) to 400 APA-registered school
psychologists. Of the 400 randomly selected psychologists, 94 (23.5%) volunteered to score the tests and send them back to the investigators. From the 94 responding psychologists, data were collected in the form of means and standard deviations for subtest scale scores and Verbal, Performance, and Full Scale IQs for the three protocols. Oakland et al. found no significant correlation between examiner error rate and examinee's level of intellectual ability.

As in the Oakland et al. (1975) study, the standard deviation of scores is commonly used as an index of interrater agreement in studies comparing examiners scoring on one or two protocols: the higher the standard deviation, the lower the rate of interrater reliability. However, it must be remembered that, when comparing the $SD$ to the $SE_M$, two different sources of error are being compared. The $SE_M$ when measured through split-half reliability coefficients reflects error attributable only to item or content sampling, whereas the $SD$ reflects error variance attributable to examiner differences. In other words, the $SD$ of the differences between examiners should be viewed as an additional source of error.

In another study involving multiple protocols sent to a random sample of psychologists, Bradley, Hanna, and Lucas (1980) attempted to investigate the differences in rate of examiner error between "hard" (which responses were deliberately ambiguous) and "easy" (which responses were deliberately clear) WISC-R protocols. The
widely differing protocols were used to ensure that the results of the study would cover the range of the population of protocols. Of the 280 randomly-sampled psychologists, 63 volunteered to participate in the study, for a 27% return rate. The measure of interrater agreement used - the standard deviation - varied from 2.9 for the easier protocol and 4.3 for the more difficult one (compared to the published $SEM$ of 3.19 in the test manual). They concluded that the lack of scoring accuracy was due to a shortage of scoring criteria for the more objective subtests and was cause for increased attention and further study.

This genre of study does provide some beneficial information for users of standardized tests, but weaknesses inherent in the research design limit the generalizability of the results. First, the task of scoring a completed protocol is contrived. In actual practice the administration and scoring are inextricably linked. Second, it is difficult to determine how much attention the practitioner pays to the task of scoring (usually anonymously) a protocol for a research study. Third, the representativeness of the sample to the population is impossible to establish when using a volunteer sample. None of the studies using volunteer samples had a very high return rate of protocols. Finally, because this type of study is limited in the number of protocols used, the generalizability of one or two (usually fabricated) protocols to the population of actual protocols is again impossible to establish.
Completed Protocols

One way to avoid the problems inherent in having volunteers score already filled-in protocols is to investigate the frequency and nature of examiner error from a sample of completed (and scored) protocols. In a review of the literature on examiner error, Slate and Hunnicutt (1988) state: "Future research needs to address the frequency, type, and specificity of errors on actual protocols. Are practicing professionals making scoring errors on test protocols in the field?" (p. 286). Few studies to date, however, have been able to answer this question. The problem of gaining access to completed and scored protocols has prevented most researchers from investigating the error rate of practicing professionals. Only a small number of investigations of this type have been conducted.

In one study, impressive because of the size of the sample, Sherrets et al. (1979), obtained two random samples of 100 WISC and WISC-R protocols drawn from the files of a psychiatric facility and from the files of the student services department of a large school system. They found that nearly 89% of the 39 examiners had made at least one error and that 46.5% of the protocols contained at least one error. Although most of the errors found did not change the diagnostic classification of the child, the largest errors found would have raised the Full Scale IQ by 9 points or lowered it by 7 points. No significant difference between the error rates on protocols from the two facilities or difference in levels of training was found.
Although Sherrets et al. (1979) used a sample considerably larger than in most other studies of its kind, they neglected to investigate two key areas. No comparisons were made of error rates between the two tests (WISC and WISC-R), something which might have shed some light on the contribution of the protocol layout to examiner error. The researchers also failed to investigate the relationship between examiner error rate and experience.

Slate et al. (1992) used data from nine school psychologists and found that these practitioners made errors on all 56 of the protocols drawn from a metropolitan school district. In this particular study, though, the term "error" was used rather liberally, and included failure to record response, failure to circle scores, and failure to record times. These may be technically considered errors because they depart from the prescribed manner of test administration, but they cannot change the score of an examinee.

Peterson, Steger, Slate, Jones, and Coulter (1991) investigated examiner error on the Wide Range Achievement Test - Revised (WRAT-R). They found that 95% of their sample of 55 protocols contained errors, with an average of three errors per protocol. They concluded that "Practitioners make a significant number of errors that adversely affect standard scores" (p. 208). Wagoner (1988) examined a similar number of protocols - 50 - in her study of examiner error on the WISC-R. She found a total of 8.4 total errors per protocol, with 34% of the Full Scale IQs altered as a result of these errors.
Definitions of Error

Most investigators use their own definition of error depending on the purpose of their study. On a standardized test, three broad areas of examiner error can be categorized. First, administration error, that is, error that occurs when the examiner strays from the prescribed standardized procedure in giving the test. The importance of adhering to standardized procedure is well-stated by Buckhalt (1990):

One of our explicit objectives in testing is to standardize the test situation so that variations in test performance can be assumed to be a function of true differences in the examinees' abilities rather than of any extraneous contextual variables (p. 478).

Studies which investigate administration errors are generally restricted to using small samples because of the difficulty involved in actually observing test administrations. Buckhalt's study (1990) used the results derived from just 14 test administrations from practitioners, and 12 from graduate students. Although the information drawn from this type of study may help direct future research, its generalizability is questionable due to the small sample size.

The second category of examiner error is scoring error: error that occurs when the examiner scores the responses of the examinee. Three WISC subtests (i.e., Vocabulary, Comprehension,
Similarities) frequently result in ambiguous responses and have been the focus of some study. Brannigan, (1975) suggested revisions for the WISC and WAIS: "more thorough scoring standards are needed for the Wechsler Intelligence Scales" and, "It would also be desirable to revise those test items which lend themselves to ambiguous replies" (p. 314). Many other studies have also investigated the problem of scoring errors and found them to be one factor contributing to examiner error on standardized tests (e.g., Ryan et al., 1983; Slate & Jones, 1990a; Warren & Brown, 1972).

A major difficulty, though, with studies investigating response-scoring errors is that psychologists and other examiners frequently do not record verbatim the responses of the examinee. Slate et al. (1992) found that school psychologists frequently departed from the prescribed standardized administration procedure. In their study they found that the school psychologists failed to record responses verbatim 30 times per protocol! Without the benefit of direct observation of the test administration, it is clearly difficult to determine if the examiner is scoring the response recorded on the test protocol or the actual (but not recorded in its entirety) response of the examinee. Studies which purport to examine the rate of scoring errors by looking at completed protocols may end up with misleading conclusions.

A final category of examiner error on standardized tests is clerical error; that is, errors caused by examiner carelessness, and perhaps exacerbated by complicated or poorly laid-out test record
forms. Sherret et al. (1979) found that nearly 89% of the examiners in their study made clerical errors, and that 46.5% of the protocols had errors. They recommended "Greater caution is called for in the clerical computations, but especially in the simple addition of the scaled scores, where the majority of errors appeared in the present samples" (p. 496). Similarly, Levenson et al. (1988) found in their study that 57% of school psychologists and interns made clerical errors, leading the authors to call for improved training in graduate programs and a revision of the WISC-R protocol in order to reduce this type of error.

Slate and Hunnicutt (1988) claimed that "Mechanical and clerical errors seem to occur frequently and detract from the accuracy of Wechsler scores" (p. 284). The problem is of great enough magnitude for them to recommend that clerks could be employed to check the areas on the protocol where clerical errors are most prevalent. Clearly, clerical errors are a major problem on standardized tests, and more information about them is called for in order to minimize their impact in weakening test reliability.

**Test Revisions**

The revision of a psychological test may bring about changes in the examiner’s response pattern and also changes in the rate of examiner’s clerical errors. The Wechsler Intelligence Scale for Children - Third Edition (WISC-III) is a revision of the Wechsler Intelligence Scale for Children - Revised (WISC-R). Because it is relatively new (1991), few studies have been published which
examine differences between the two tests, especially in the area of susceptibility to examiner errors. Little (1992) compared the two tests and found few major changes apart from the appearance of actual test materials. The record form or protocol has been re-designed with some changes, for example, there is now more space for writing behavioral observations.

Of greatest concern for Little (1992) is the changed process for converting raw scores to standard scores. The subtests on the record form are listed in their order of administration, while on the conversion table, they are split with the Verbal subtests in one location, and the Performance subtests in another. Little predicts that “This different format of presentation is sure to be confusing and cause recording errors” (p. 151).

Examiner Characteristics

The interaction between examiner characteristics and examiner error has been the subject of several studies. In one of the earlier studies, Cieutat (1965) evaluated the effect of examiner differences on Stanford-Binet scores. Although he found no significant effect for sex of examinee, sex of examiner was significant, with the female examiners eliciting significantly higher IQ scores than the male examiners. Interaction between sex of examiner and examinee was "of marginal significance" (p. 318). The results of this study can be called into question, though, because of difficulties with the interpretation of the data analysis. Because of
unequal *ns* in the analysis of variance cells, Cieutat issued the caveat that "tests of significance are only approximate" (p. 318).

Levenson et al. (1988) also looked at the effects of sex of examiner. In their study, sex of examiner was not a significant factor, but education of examiner did account for a significant portion of the differences in error rates. They found that doctoral school psychologists made more errors than both school psychology interns and Master's-level school psychologists.

In a study comparing PhD psychologists and graduate students, Ryan et al. (1983) found that the PhD psychologists committed a significantly higher rate of scoring errors than the graduate students on the Performance scale of the WAIS although there were no significant differences between the two groups on any of the subtest or IQ score means. Unfortunately, the results from this study are weakened from the earlier-mentioned basic design flaw in which examiners find themselves in the contrived situation of scoring completed but unscored protocols.

Bird (1992) attempted to find which examiner characteristics were most closely linked to rate of clerical error on the WISC-R. In her study, the percentage of protocols with Full Scale IQ changes ranged from 4% to 24% in the three samples of protocols (50 in each sample) she selected. She concluded that the issue of time pressure appeared to be more related to the error rate than did any variability in training.
The effect of examiner experience has been found to be worthy of some study, but most of the research has dealt exclusively with graduate students and not with practitioners. Conner and Woodall (1983) found that student examiners showed a significant decrease in administrative and scoring errors with experience in administering the WISC-R. Clerical errors, however, did not decrease, leading the authors to call for further research in this area. Slate and Jones (1990b) found that error rates decreased after 8 administrations of the WAIS-R. Gilbert (1992) also found student error rates on the WAIS-R decreasing from the beginning to the end of a course on test administration.

Studies involving graduate students are relevant for those organizing testing courses for graduate programs. It does not necessarily follow, however, that the same trend holds true for practitioners who are not typically in the process of learning how to use a test. One of the omissions in studies involving practitioners is the inclusion of examiner experience as an independent variable in the analysis of the data. Whereas sex and education of examiner have been studied, experience has rarely been investigated. The result of this gap in the research is that little is known regarding the effect that experience plays in practitioners' error rates. Research into the impact that experience has on the error rates of practicing professionals is in short supply but plainly needed.
Assessment Practices

A great deal of a clinician's time is spent in assessment. A study by Anderson, Cancelli, and Kratochwill (1984) revealed that 44% of the school psychologist respondents to their questionnaire spent between 41% and 80% in testing. During this time spent testing, a plethora of different tests are used. Studies investigating examiner error have been conducted on many of these tests, including the Stanford-Binet (e.g., Cieutat, 1965; Warren & Brown, 1972), the Kaufman Assessment Battery for Children (Hunnicutt, Slate, Gamble, & Wheeler, 1990) and, the Wide Range Achievement Test - Revised (Petersen et al., 1991) with all finding significant error attributable to examiner error.

But the most widely-used tests among practitioners (Lubin, Larsen, & Matarazzo, 1984), and the tests on which most of the work investigating examiner error has been done, are the Wechsler scales. On these scales, studies of examiner error have largely focussed on the WISC-R and the WAIS and WAIS-R, but few studies have yet to be published on the WISC-R's revision, the WISC-III. Because the WISC-III is a frequently-used instrument, and because little is available at this time regarding rate of examiner error, any study which attempts to investigate this instrument would be valuable.

Summary

Much research has been completed in the area of examiner error and standardized tests. Many of these studies have analyzed the testing practices of graduate students, and have found that their
examiner errors are a significant source of variation in test scores.
Although data from graduate students are more readily available and
accessible than data from practicing professionals, and although the
results may be useful to improve graduate training programs, the
data produce results that are not easily generalizable to
professionals working in the field.

Another way of avoiding the difficult task of obtaining data
from actual, completed test records is to have a number of subjects
score a completed, but unscored test protocol. Researchers that
have utilized this design typically send out one or two fabricated
protocols to a random sample of professionals. The returned and
scored protocols form the data-base for the subsequent analysis.
Unfortunately, the sample loses its randomness, and hence its
representativeness, when a typically low proportion of
professionals choose to participate in the study. The task of
scoring, but not administering a test is an exercise that is
contrived, and not representative of practice in the field.

Investigators have looked at a number of different error-types,
including administrative error, scoring error, and clerical error.
Administrative error is studied infrequently because of the
difficulty and time involved in observing test administrations.
Scoring error is difficult to substantiate, because examiners very
frequently do not record responses verbatim. A look at completed
protocols, therefore, does not always give a clear indication of the
actual responses of examinees. The third category of error, which
can be accurately measured from examinations of test protocols, is clerical error. Clerical error has been found to be a factor in the protocols of most practitioners (e.g., Levenson et al., 1988; Sherrets et al., 1979).

The magnitude of examiner error varies from study to study, and is measured in different ways depending on the design of the study and on the definition of examiner error. Investigations using multiple scorers of single protocols generally report the standard deviation of scores. For example, the study by Bradley et al. (1980) on the WISC-R, reports the standard deviation ranging from 2.9 for the easier-to-score protocol, to 4.3 for the harder protocol.

In studies investigating errors in completed and scored protocols, error rates have been reported as change in Full Scale IQ score (as high as 17 points for Full Scale IQ in Miller et al.'s study [1970]). Proportion of errors among examiners and protocols has also been used, as in the study by Sherrets et al. on the rate of clerical errors on the WISC and WISC-R. The investigators found that 89% of examiners made clerical errors, and that 46.5% of the protocols contained this type of error. If the definition of error is broadened to include scoring and administrative errors, almost all protocols contain errors (e.g., Franklin, Stillman, & Burpeau, 1982; Slate & Jones, 1990a; Wagoner, 1988).

Various examiner characteristics have been investigated in relation to rate of examiner error. Sex of examiner appears to have
little impact on error rate (Levenson et al., 1988). Education of
examiner (Ph.D vs. Master's degree) may make a difference, with
studies showing doctoral-level psychologists making more errors
than master's-level psychologists (Levenson et al., 1988; Ryan et al.,
1983).

Although examiner experience appears to be related to the
errors graduate students make on standardized tests, little research
has been completed using practitioners such as clinical or school
psychologists for whom assessment is a major part of their day-to-
day practices. The value of studies involving practitioners' errors
would be enhanced if the experience of examiner was included as an
independent variable in the analysis of the data. More research is
clearly called for to investigate the relation between practitioners' experience and clerical errors. Test revisions, with accompanying
changes in test protocols, may also result in a change in examiners'
clerical errors. Research comparing practitioners' error rates
between test versions would be widely welcomed.
Chapter III
Methodology

The purpose of this ex post facto study was to investigate how the nature and rate of examiner error changes from the Wechsler Intelligence Scale for Children - Revised, to the Wechsler Intelligence Scale for Children - Third Edition. The two central questions addressed in this study were how the examiner clerical error rate changed as a result of the test's revision, which included major alterations to the format of the test record or protocol; and how the pattern or trend of error changed as a result of the examiner's increasing experience on the new test, the WISC-III.

This chapter will first define the sample and population of test protocols as used in this study. Next, the characteristics of the examiners will be examined, followed by the assumptions that need to be made if the results of this study are to be seen as generalizable to a greater population. Finally, the actual procedures and data analysis methods used will be explained.

Sample

The population of WISC-III protocols was made up of the total protocols (approximately 1200) completed by all of the school psychologists in a suburban school district since the beginning of the test's use, in 1991. The population of WISC-R protocols (approximately 1800) consisted of the total number of protocols completed by the school psychologists in the district in the last
three years of its use. Protocols from before this time period were not available.

The sample consisted of a total of 252 test protocols: 18 WISC-R and 18 WISC-III protocols randomly drawn from the files of each of seven school psychologists practicing in a large suburban community near a large Western Canadian city, chosen because of the convenience of access to files for the researcher. The community showed demographic similarity with the province in several key areas. The community's average annual income, educational level of population, and ethnic composition were very similar with the same provincial indicators (Ministry of Education, 1993). Also similar were the student/teacher ratio and teacher experience and certification.

Of the total number of school psychologists in the school district, seven fit the criteria for this study: a) had administered both the WISC-R and WISC-III, and b) still had access to older (i.e., WISC-R) protocols. In some cases, older WISC-R protocols had been destroyed because of lack of storage space.

The sample size - that is, number of protocols from each examiner - was determined through comparison to the number of protocols used in previous studies of a similar nature. For example, Slate et al. (1992) collected 56 WISC-R protocols from nine examiners in a large metropolitan school district. Wagoner (1988) collected 50 protocols from 6 examiners, for her study on the WISC-
In a study of examiner errors on the WRAT-R, Peterson et al. (1991) looked at 55 protocols from 9 examiners. The study found to have the largest sample was conducted by Sherrets et al. (1979), in which 100 protocols from each of two institutions were examined. The present study’s total of 126 protocols for each test was deemed sufficiently large in comparison with the sample size used in previous studies of a similar nature.

**Characteristics of Examiners**

All of the examiners had Master’s degrees in either school or clinical psychology from four different graduate training programs in Canada and the United States. Their years of experience (see Table 1) ranged from 3 to 27 years (Mean = 9.6, S.D. = 8.6). All of the participants were trained to administer the WISC-R, but received no formal training on its revision, the WISC-III. The number of administrations per year varied with each school psychologist and with the particular year, with the mean number of WISC-R or WISC-III administrations ranging from 50 to over 70 per examiner per school year.

The protocols from each school psychologist were drawn in the following manner: the WISC-R protocols were randomly drawn from the last three years of its administration, and the WISC-III protocols from the first year-and-a-half of its administration. The sample, therefore, consisted of 18 protocols for the WISC-R, and 18 protocols for the WISC-III from each examiner, drawn from a pool
ranging from 150 to 210 total protocols for the WISC-R, and 75 to 105 total protocols for the WISC-III for each examiner. In order to

Table 1
Examiners' Years of Experience

<table>
<thead>
<tr>
<th>Examiners</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years of Experience</td>
<td>15</td>
<td>27</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>7</td>
</tr>
</tbody>
</table>

Note. Years of experience consists of number of years employed in a position which requires the administration of standardized tests.

examine the trend or pattern of error over time on the WISC-III, the 18 months of administration were further divided into half-year segments, and six protocols were drawn from each of the three half-year periods.

Characteristics of Examinees

Table 2 displays the characteristics of the examinees from whom the protocols were drawn in this study. Note that the mean and median IQs reflect a special education population, i.e. a population of children who have been referred for assessment or service because of questions regarding their performance in an academic setting.
### Table 2

**Characteristics of Examinees**

<table>
<thead>
<tr>
<th></th>
<th>WISC-R</th>
<th>WISC-III</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Age Range</td>
<td>6-2 to 16-8</td>
<td>6-3 to 16-7</td>
</tr>
<tr>
<td>2. Median Age</td>
<td>9-10</td>
<td>10-6</td>
</tr>
<tr>
<td>3. IQ range</td>
<td>46 to 126</td>
<td>50 to 115</td>
</tr>
<tr>
<td>4. Mean IQ</td>
<td>89</td>
<td>85</td>
</tr>
<tr>
<td>5. Median IQ</td>
<td>89</td>
<td>86</td>
</tr>
</tbody>
</table>

**Note.** Total of 126 protocols per test. Also, "6-2" is read as 6 years and 2 months.

**Assumptions**

Several assumptions need to be made if the results of this study are to be generalizable to the population of protocols in the school district and in the province. It is assumed that the sample protocols drawn from the school district files are representative of the total number of protocols in the files, and of the population of protocols in the province. Because the protocols were drawn randomly (on the WISC-III, the total number of a particular examiner's protocols for each half-year period divided by the number necessary to draw 6 files, and then every 4th or 5th protocol drawn) this assumption can be supported. The characteristics of the examinees and the demographics of the community of the examinees,
as outlined previously, are not atypical for the province, suggesting that the examinees, and therefore, the protocols did not vary demographically from other protocols in the province.

It is assumed that the testing practices of the seven school psychologists who completed the test protocols are similar to the practices of other school psychologists using the tests. The graduate training, and years of experience of the seven are varied enough to support this assumption. Previous studies have shown that Master's-level psychologists have a lower examiner error rate than Doctoral-level psychologists (e.g., Levenson et al., 1988; Ryan et al., 1983). Thus, because only Master's-level psychologists were included in this study, the results may be an under-estimate of the actual rate of error among all practitioners.

It is not assumed that the examiners' training is the same for each of the two tests. It is acknowledged that the examiners received formal graduate-level training on the WISC-R, but no direct formal training on the WISC-III. In a study by Perot (1992) no significant difference in number of scoring errors was found between those with formal training on the WISC-R and those without. Most practitioners in the field are not formally re-trained on revisions of tests, and the conditions in this study are likely representative of the conditions in the population of examiners in the province. In any case, the results of this study, and any conclusions drawn regarding the ease-of-use of the WISC-III protocol, must be seen in the light of the fact that any training on
the second test (the WISC-III) is not as comprehensive as the training on the first test (the WISC-R).

Procedure

In order to protect the identity of the examiners, an associate photocopied the 252 completed protocols with the names of tester and test subject deleted in order to ensure anonymity of examiner and confidentiality of the examinee and actual test results. The examiner was identified by a code number only; the examinee was not identifiable in any way.

The protocols were then analyzed, item by item, for clerical errors using an error checklist and data compilation sheet. Error types tabulated were clerical in nature (Levenson et al., 1988; Sherrets et al., 1979) and consisted of: addition of subtest raw and scaled scores, computation of chronological age, transformation from tables of raw scores to scaled scores, and transformation of scaled scores to IQ scores. Because the WISC-III is relatively new (published 1991) and has been used for a relatively brief time period, no comparison between error rates over an extended time period was possible.

Analysis

The analysis of the data consisted of a tally of overall clerical errors on each test, as well as a tally of "serious" clerical errors, that is, errors causing Full Scale IQ changes (as has previously been done (e.g., Bird, 1992). On each test, this number of errors was
shown as a proportion of total protocols, along with mean, range and standard deviation of errors. The number of errors made was also correlated, on both tests, with the examiners' years of experience in administering standardized tests.

The errors made were divided into their respective categories: addition of raw scores, transformation of raw to scaled scores, addition of scaled scores, transformation of scaled to IQ scores, and chronological-age computation. Comparisons between the two tests were made using t-tests for paired or dependent observations for error type, or category, as well as for overall and "serious", or Full Scale IQ-changing errors. Location of error (type #1 only, the other error categories cannot logically be given a subtest location) on each test was examined.

On the WISC-III, the rate of clerical errors was examined over the three six-month time periods. Both overall and serious, i.e. IQ-changing errors were examined. A repeated-measures 2-factor ANOVA was used to compare the number of errors over the three time periods. When significant differences were found, Tukey's multiple comparison technique was used to locate these significant differences.
Chapter IV

Results

This chapter will describe the results of the study outlined in Chapter III. The nature and range of clerical errors on the WISC-R and WISC-III will be described, with frequency of error per examiner given, as well as error location and category. Error frequencies between the two tests will be compared, and the correlation between examiner experience and number of errors will be examined. On the WISC-III only, the number of overall and FSIQ-changing errors will be plotted over three six-month time periods, with comparisons made between the number of errors in the three time periods.

Errors and IQ Changes

Table 3 (see also Figure 1) displays the proportion of clerical errors on both tests and the resulting IQ changes. The WISC-III had a higher number of total errors, Verbal IQ (VIQ), Performance IQ (PIQ), and Full Scale IQ (FSIQ) changes than did the WISC-R. Although 38% and 42% of the WISC-R and WISC-III protocols, respectively, contained clerical errors, not all of these clerical errors resulted in a VIQ, PIQ, or FSIQ change. Nevertheless, almost one-quarter (23%) of the WISC-III protocols completed reported incorrect Full Scale IQ scores.

On the WISC-R protocols with FSIQ changes, the mean FSIQ change was 2.3 points, with a range of 1-8 points. The mean FSIQ change on the WISC-III protocols with errors was 1.8 points, with a
Table 3

**Number and (Proportion) of Protocols with Errors by Test**

<table>
<thead>
<tr>
<th></th>
<th>WISC-R</th>
<th>WISC-III</th>
<th>DIFFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protocols</td>
<td>126</td>
<td>126</td>
<td></td>
</tr>
<tr>
<td>Overall Errors</td>
<td>48 (.38)</td>
<td>53 (.42)</td>
<td>5 (.04)</td>
</tr>
<tr>
<td>VIQ changes</td>
<td>9 (.07)</td>
<td>12 (.09)</td>
<td>3 (.02)</td>
</tr>
<tr>
<td>PIQ changes</td>
<td>15 (.12)</td>
<td>23 (.18)</td>
<td>8 (.06)</td>
</tr>
<tr>
<td>FSIQ changes</td>
<td>21 (.17)</td>
<td>29 (.23)</td>
<td>8 (.06)</td>
</tr>
</tbody>
</table>

**Note.** Proportions of error-type for each test are presented within parentheses. Error proportions do not sum to 1.0: errors may cause a VIQ or a PIQ change, and a FSIQ change, or only a VIQ or PIQ change, or no change in scoring at all.
<table>
<thead>
<tr>
<th>Error Effect</th>
<th>WISC-R</th>
<th>WISC-III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIQ Changes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIQ Changes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSIQ Changes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1.** Number of Protocols with Errors by Test (per 126 protocols)
range of 1-7 points. In comparison, the $SE_M$ on the WISC-III is 3.20, but it can be argued that this does not account for clerical errors. Of the errors which led to a change in Full Scale IQs, one error on the WISC-III led to a descriptive category change, from Mildly Mentally Handicapped to Borderline. Both tests showed numerous other changes (e.g. 70-73, on the WISC-R; 71-75, on the WISC-III) which were close enough to category changes to possibly influence placement decisions.

**Examiners**

Six of the seven examiners made clerical errors (see Table 4; also shown graphically in Figure 2), with one examiner (#5) completing error-free protocols on both tests. On the WISC-R, examiners committed a mean of 6.86 clerical errors per 18 protocols (38%, Range 0-11, S.D. 4.41), while on the WISC-III, a mean of 7.57 errors were found per 18 protocols (42%, Range 0-12, S.D. 3.78).

The relationship between the examiners' years of experience as school psychologists and the number of clerical errors made on each of the tests was examined using a Pearson product-moment correlation coefficient. On the WISC-R, no strong relationship was found between years of experience and errors. On the WISC-III, however, a positive relationship was found between overall errors and experience ($r = .54$). When years of experience was correlated with Full Scale IQ-changing errors, the correlation coefficient was even higher ($r = .60$). These correlation
Table 4

Number of Total Errors by Test per Examiner

<table>
<thead>
<tr>
<th>Examiner</th>
<th>WISC-R</th>
<th>WISC-III</th>
<th>DIFFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>9</td>
<td>-2</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>8</td>
<td>-2</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>10</td>
<td>-1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

Mean Errors per test 6.9 7.6 .70

Note. There are 18 protocols per test for each examiner.
Figure 2. Errors by Examiner (per 18 protocols)
coefficients were not statistically significant, but, based on this sample alone, the pattern suggests that the correlation likely would have been significant if the $n$ had been larger.

**Location of Errors**

The location of errors on each test (see Table 5) showed a definite pattern, with more errors made on the Performance subtests than on the Verbal subtests. The one test most prone to error on both tests was Coding (errors on this test could be defined as either scoring or clerical errors) with 54% and 38% of total errors. The errors on this subtest, though numerous, frequently did not have a great impact on the subtest scaled score, and the resulting Performance IQ and Full Scale IQ. On the WISC-R, three subtests were without any clerical errors: Information, Similarities and Object Assembly. On the WISC-III, only Information and Arithmetic were error-free.

Far more category-#1 (addition of raw score) errors were found on the Performance subtests than on the Verbal subtests for both tests. On the WISC-R, 67% of category-#1 errors were found on the Performance subtests; on the WISC-III, 82% of these errors were found on the Performance subtests.

**Error Category**

Table 6 (shown graphically in Figure 3) displays the breakdown of clerical error by category for each test. The most common error,
Table 5

Location of Clerical Errors by Test

<table>
<thead>
<tr>
<th>Protocols</th>
<th>WISC-R</th>
<th>WISC-III</th>
<th>DIFFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal Subtests</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Similarities</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>4</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Comprehension</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Digit Span</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Performance Subtests</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Picture Comp.</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Coding</td>
<td>18</td>
<td>16</td>
<td>-2</td>
</tr>
<tr>
<td>Picture Arr.</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Block Des.</td>
<td>6</td>
<td>1</td>
<td>-5</td>
</tr>
<tr>
<td>Object Ass.</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Mazes</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Symbol Search</td>
<td>n/a</td>
<td>1</td>
<td>n/a</td>
</tr>
</tbody>
</table>
Table 6

Error Categories Across Test

<table>
<thead>
<tr>
<th>Protocols</th>
<th>WISC-R</th>
<th>WISC-III</th>
<th>DIFFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Addition of Raw Scores</td>
<td>33 (.69)</td>
<td>42 (.79)</td>
<td>9 (.10)</td>
</tr>
<tr>
<td>2. Transformation of Raw to Scaled Scores</td>
<td>6 (.12)</td>
<td>5 (.09)</td>
<td>-1 (.03)</td>
</tr>
<tr>
<td>3. Addition of Scaled Scores</td>
<td>2 (.04)</td>
<td>5 (.09)</td>
<td>3 (.05)</td>
</tr>
<tr>
<td>4. Transformation of Scaled to IQ Scores</td>
<td>7 (.15)</td>
<td>1 (.02)</td>
<td>-6 (.13)</td>
</tr>
<tr>
<td>Total Errors</td>
<td>48</td>
<td>53</td>
<td>5</td>
</tr>
</tbody>
</table>

Note. The parentheses show proportion of error per 126 protocols.
Category 1: Addition of Raw Scores
Category 2: Transformation of Raw to Scaled Scores
Category 3: Addition of Scaled Scores
Category 4: Transformation of Scaled to IQ Scores

Figure 3. Clerical Errors by Category
by far, for both tests was the addition of raw scores. On the WISC-R, 69% of clerical errors came as a result of incorrect addition of raw scores (error-category #1), while on the WISC-III, the proportion was slightly higher: 79% of the clerical errors were as a result of incorrect addition of raw scores.

Although addition of raw score errors were most common, other error types were more likely to have a greater impact on IQ scores. Whereas incorrect addition of raw scores only affects one subtest, and then often only minimally, the other error types are errors involving scaled scores and standard scores and are more certain to change the IQ scores.

The single error category showing the most change was error category #4, transformation of scaled to IQ scores, which accounted for 15% of errors on the WISC-R, but only 2% of errors on the WISC-III. No chronological-age computation errors (error-category #5) were found on any protocols for either test.

Comparison between WISC-R and WISC-III

A series of t-tests for paired or dependent observations were run to analyze the difference between means on the two tests for total errors, Full Scale IQ-changing errors, and for each error category for the two tests (see Table 7). No significant difference was found between the means on the two tests for any of these categories.
### Table 7

#### Comparison of Error-types for WISC-R and WISC-III

<table>
<thead>
<tr>
<th>Error Type</th>
<th>WISC-R</th>
<th>WISC-III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>S.D.</td>
</tr>
<tr>
<td></td>
<td>(n=7)</td>
<td></td>
</tr>
<tr>
<td>1. Overall Errors</td>
<td>6.86</td>
<td>4.41</td>
</tr>
<tr>
<td>2. FSIQ-changing</td>
<td>3.00</td>
<td>2.08</td>
</tr>
<tr>
<td>errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Category-#1&lt;sup&gt;b&lt;/sup&gt;</td>
<td>4.71</td>
<td>2.81</td>
</tr>
<tr>
<td>4. Category-#2</td>
<td>0.86</td>
<td>1.21</td>
</tr>
<tr>
<td>5. Category-#3</td>
<td>0.29</td>
<td>0.76</td>
</tr>
<tr>
<td>6. Category-#4</td>
<td>1.00</td>
<td>1.41</td>
</tr>
</tbody>
</table>

**Note.** The t-tests used were for paired or correlated observations.

<sup>a</sup>Mean number of errors per 18 protocols.

<sup>b</sup>Type-#1 errors: Addition of Raw Scores

Type-#2 errors: Transformation of Raw to Scaled Scores

Type-#3 errors: Addition of Scaled Scores

Type-#4 errors: Transformation of Scaled to IQ scores
Examiner Experience and the WISC-III

One of the questions asked in this study pertains to the effects that examiner experience on a test has on the rate of clerical error. Table 8 (and Figure 4) displays each examiner's overall errors over three consecutive half-year time periods from the beginning of WISC-III administration. This table shows the mean error rate decreasing over each of the three time periods from 3.0 errors per six protocols at the beginning of administration, to 2.0 errors per six protocols in the third six-month period.

A repeated-measures 2-factor ANOVA was run in order to compare the rate of overall examiner clerical error on the WISC-III in each of the three time periods. No significant effect was found ($F_{[2,12]} = 1.10, MSE = 1.59$) for the time/experience factor (the variance for the person factor cannot be calculated because there is only one observation per cell). Thus, there was no significant difference between overall error rates between the three six-month time periods.

Table 9 (and Figure 4) displays examiners' Full Scale IQ-changing errors over the three time periods from the beginning of WISC-III administration. A decrease of IQ changing errors was noted: from 2.0 per six protocols at the beginning of WISC-III administration, to .57 in the third six-month period of administration.
**Table 8**

**WISC-III Overall Errors Over Time**

<table>
<thead>
<tr>
<th>Examiner</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

M 3.0 2.6 2.0

**Note.** Errors listed consist of clerical errors per six protocols for each of the three six-month time periods for each examiner.
Figure 4. WISC-III Mean Errors (per 6 protocols)
Table 9

**WISC-III FSIQ-Changing Errors Over Time**

<table>
<thead>
<tr>
<th>Examiner</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

|M| 2.0 | 1.6 | .6 |

**Note.** Errors listed consist of Full-Scale IQ-changing clerical errors per six protocols for each of the three six-month time periods.
Another repeated-measures ANOVA was run to investigate the differences between means of serious errors - that is, errors which cause changes in Full Scale IQ - for the three time periods. It was found that there was a significant difference \( (F_{2,12} = 4.06, p < .05, MSe = 0.93) \) between the three means.

Through further investigation using the Tukey honest significant difference multiple comparisons technique \( (HSD_{.05} = 1.37) \), it was found that the mean for time period 1 was significantly higher than the mean for time period 3. In other words, significantly more Full-Scale IQ-changing clerical errors were made at the very beginning of WISC-III administration than after a year to a year-and-a-half of experience.

Summary of Results

Clerical errors were found on 38% and 42% of WISC-R and WISC-III protocols, respectively, with 17% and 23% of protocols having Full Scale IQ changes as a result. The mean FSIQ change on the WISC-R was 2.3 points (range 1 - 8); on the WISC-III the mean change was 1.8 points with a range of 1 - 7 points. Only one error led to a change in descriptive category, from Mildly Mentally Handicapped to Borderline.

The study's first hypothesis stated that the rate of practitioners' clerical error on completed and scored WISC-III protocols will be equal to their rate on the WISC-R. No significant differences were found between the mean errors on the two tests
for overall errors, Full Scale IQ-changing errors, or any of the four error-categories examined. A positive, but statistically non-significant correlation was found between examiners' years of experience and number of clerical errors made on the WISC-III.

The second hypothesis stated *practitioners' clerical error rates will not change significantly with increasing experience on the WISC-III*. The examiners' mean overall error rate was reduced from 3.0 errors per six protocols at the beginning of administration, to 2.0 errors per six protocols in the third six-month period. The difference was not statistically significant ($F_{[2,12]} = 1.10$). The examiners' Full Scale IQ-changing errors were also compared over the three-month time period. At the beginning of test administration, 2.0 errors were made per six protocols; during the third six-month time period, 0.57 errors were made which changed the Full Scale IQs. The difference between these means was significant ($F_{[2,12]} = 4.06; \ p < .05$).
Chapter V

Discussion

The purpose of this study was to investigate the difference in error rates for practitioners between the WISC-R and its revision, the WISC-III. The first hypothesis presented in this study stated that the rate of practitioners’ clerical error on completed and scored WISC-III protocols will be equal to their rate of error on the WISC-R. The results of this study show that the WISC-III error rate was not significantly different from the error rate on the WISC-R.

Also of interest was the pattern of error rate on the WISC-III over three time periods, starting from the beginning of its administration. The second hypothesis stated that practitioners’ clerical error rates will not change significantly with increasing familiarity (experience) on the WISC-III. The results from this study suggested that whereas overall clerical error rates showed no significant change, Full Scale IQ-changing errors showed a significant decrease over the three time periods; that is, school psychologists made significantly more errors at the beginning of test administration than after 18 months of administration.

Error Rates

Of the 252 total protocols, 40% contained clerical errors, with almost 20% of the total number of protocols having incorrect Full Scale IQs as a result of these errors. Slightly more clerical errors were found on the WISC-III (53, or 42% of the 126 protocols) than on
the WISC-R (48, or 38% of the 126 protocols) but the difference was not significant. There was also no significant difference between the two tests for errors that caused changes in Full Scale IQs.

Other studies have reported similar rates of clerical error. For example, Sherrets et al. (1979) reported that 46.5% of the 200 WISC-R protocols examined in their study contained at least one clerical error. Wagoner (1988) found that 34% of the 50 WISC-R protocols contained errors which changed Full Scale IQs, although the errors examined in her study included scoring errors as well as clerical errors, thus inflating the rate of errors in comparison to this study. Error rates in studies using student examiners are generally much higher than the rates reported in the present study. For example, Slate and Chick (1989) found that 67.4% of their student-examiners made errors which changed Full Scale IQ scores.

Of the WISC-R protocols with FSIQ changes, the mean change was 2.3 points, with a range of 1-8 points. The standard deviation of change in FSIQ scores was 1.14. On the WISC-III, the mean change on protocols with errors was 1.8 points, with a range of 1-7 points. The standard deviation of change in FSIQ scores was 1.05. The $SE_M$ of the WISC-III is 3.20, but because of the method used in calculating this $SE_M$, it is evident that clerical errors must be seen as a source of additional error.
Standard Error of Measurement

As stated earlier (in Chapter 1), the $SE_M$ on the Wechsler scales is calculated from a composite of split-half reliability coefficients taken from individual subtests. The resulting $SE_M$ therefore, cannot be thought to account for error from sources outside of these subtests. The $SE_M$ published for a Wechsler test might be seen to account for one type of clerical error - addition of raw scores - in the protocols used in the reliability studies in the production of the test. However, it is almost certain that the rate of clerical errors is higher in actual practice than is the rate given from the test’s original reliability study. Those errors not found within the actual subtests - transformation of raw to scaled score, addition of scaled scores, transformation of scaled scores to IQ scores - are not accounted for at all by the published $SE_M$ of the test.

It has been suggested that the $SE_M$ of a test should include not only errors brought about through content sampling, as in the case of the Wechsler scales, but also time sampling errors, errors that result from mistakes in scoring, and from improper test administration (Bradley, Hanna, & Lucas, 1981). The Bradley et al. study suggested that a composite $SEM$ be composed in the following manner: $SE_M\,comp = \sqrt{SE_{Mcs}^2 + SE_{Mts}^2 + SE_{Ms}^2 + SE_{Ma}^2}$ (cs=content sampling; ts=time sampling; s=scoring; a=administration). Their estimation of a composite $SEM$ on the WISC-R came to 6.5 IQ points (as opposed to the published 3.19 points).
The results of the present study suggest that because clerical errors exist and have an effect on test reliability, a composite $SE_M$ should also include clerical errors ($SE_{Mc}$). The standard deviation of IQ changes on the WISC-R was 1.14 points (1.05 on the WISC-III). When included in the above formula, the composite $SE_M$ rises to 6.7 points for the WISC-R, as opposed to the published $SE_M$ of 3.19 points. This pattern is likely similar for the WISC-III.

**Examiners**

Six of the seven examiners (86%) made clerical errors on both tests, with one examiner completing error-free protocols on all 18 protocols of the WISC-R and all 18 of the WISC-R. In comparison, Sherrets et al. (1979) found 89% of the examiners in his study had committed clerical errors; Levenson (1988) found 57% of examiners made clerical errors on the 162 WISC-R protocols used in the study. The results from this study in this area are similar to results in previous studies and suggest that the sample of examiners is similar to the sample in previous research.

The examiners' years of experience were correlated with the number of errors made on the two tests. On the WISC-R, no relationship was seen between years of experience and number of clerical errors made. On the WISC-III, however, a positive correlation was seen between overall errors and experience ($r = .54$). Even stronger was the correlation between IQ-changing errors and experience ($r = .60$). These positive correlations (although not statistically significant) suggest that, in this sample, those
examiners with the greatest years of experience had the most difficulty in adapting to the changes in a newly revised standardized test, and were more likely to make careless errors. These more experienced examiners made more clerical errors, and they reported more inaccurate results than did examiners with less experience.

The error rates in the present study are consistent enough with those in other studies to support the assumption that the examiners are representative of the population of examiners. Conclusions drawn in this study should be applicable to other practitioners using these tests. Also, the demographic characteristics of the community from which the data were taken suggest that the examinees were likely representative of the population of examinees in the province.

Error Categories

The most common error type by far, on both tests, was addition of raw scores. On the WISC-R, 69% of all errors were of that type, and on the WISC-III, 79% of all errors were in that category. The difference between the two tests was not significant. This type of error, though most frequent, is least likely to make a major difference in the IQ score at the end of the test.

Category #2 errors - transformation of raw to scaled scores - was the second most frequent error on the WISC-R, with 15% of errors, but was the least-frequent error type on the WISC-III, with only 2% of errors. The difference between the two tests for this
error type was not statistically significant, but suggests a
difference between the two tests.

Little (1992) suggested that "Of greatest concern with
regard to administration and scoring is the process of converting
raw scores to standard scores ... This different format of
presentation is sure to be confusing and cause recording errors" (p.
151). That the WISC-III seems less prone to error in the area of
transformation of raw to scaled scores is somewhat surprising. The
changes made on the WISC-III protocol appear to make the process of
transformation of raw to scaled, or standard scores more confusing.
Contrary to Little's predictions, however, the new format does not
seem to be more error-prone in this area, with the present study
showing relatively few errors being committed in that location.

The second most frequent error-types on the WISC-III were
transformation of raw to scaled scores (9%), and addition of scaled
scores (9%). These errors were found in similar number on the
WISC-R, with no significant difference between the two tests in
this area. That these error rates are the same should be of some
concern to those using the WISC-III: these two error categories are
among the most likely to have an impact on Full Scale IQ.

Most studies report at least some clerical errors in the area of
computation of chronological age (e.g., Levenson et al., 1988;
Sherrets et al., 1979). No errors of this type were found in this
study, probably as a result of the referral and reporting process used
in the school district from which the protocols were taken. In the school district selected for this study, the school psychologists are given referrals with age and birth-date supplied; after the assessment, the reports with birthdate and age are distributed to parents and school personnel involved with the child. It is likely that any chronological-age computation error would be noticed and the test protocol would be corrected.

The WISC-III Over Time

The WISC-III protocols were also scrutinized for the rate of examiner errors over three time periods, starting with the beginning of administration and continuing for three half-year time periods. All of the examiners came to the WISC-III with training and experience on the WISC-R, but no formal re-training on the WISC-III. The revised test is similar enough to the WISC-R that examiners are expected to be able to make the transition with a minimum of difficulty.

Two types of errors were looked at: overall clerical errors, and serious errors (i.e., those errors which cause the Full Scale IQs to change). The overall error rate declined over the three time periods from 3.0 per six protocols, to 2.0 per six protocols, but the decrease was non-significant. However, for the errors which cause Full Scale IQ change, a significant difference was found between the first and third time periods. The mean number of IQ-changing errors (2.0 per 6 protocols, or 33%) at the beginning of administration was
significantly higher than during the third six-month period, which had an IQ-changing error rate of 0.6 per 6 protocols (10%).

Past studies have shown some decrease in examiner errors in graduate students' administrations of standardized tests, but few, if any studies to date have looked at practitioners. Slate and Jones (1990c) found no decrease in graduate students' scoring error rates after 5 practice administrations of the WISC-R, and a slight decrease after 10 administrations. In their study, error rates were very high: 79.7% of the protocols completed contained errors which changed the Full Scale IQ. They concluded that the training for administration of standardized tests was inadequate, and should consist of more than simple repetition of test administrations.

Gilbert (1992) investigated an alternative method of teaching test administration and found students' WAIS-R administration errors significantly reduced from the beginning to the end of the semester. Gilbert's study, however, looked only at administration errors, and not at scoring or clerical errors, which may or may not have followed a similar pattern.

The results of the present study indicate that practicing school psychologists make a significantly higher number of IQ-changing errors at the beginning of administration of a newly-revised instrument than they do 18 months later. Interestingly, though, the rate of overall errors does not change significantly over time. The reason for this difference in pattern likely lies in the
make-up of each of these two broad categories of errors (overall and IQ-changing).

The overall errors consisted mostly of category #1 errors, that is, addition of raw scores. These errors often result in minimal change in scaled scores and, consequently, IQ scores. The WISC-R and WISC-III do not differ vastly in the way raw scores are added up, and therefore no change over time is noted when the examiners begin using the revised test. The IQ changing errors, though, are more likely to be category #2, #3 or #4 errors which have more potential to change IQ scores. These types of errors can be directly attributed to changes in the protocol format of the newly-revised test, and to the examiners’ inexperience with the test. When all three time periods of WISC-III administration were examined, the WISC-III did not show a significantly higher rate of category #2, #3, and #4 errors than the WISC-R. It may be construed that the cause of the high rate of IQ-changing errors at the beginning of WISC-III administration can be attributed to examiners’ inexperience and lack of retraining rather than to a deficiency in protocol design.

**Strengths and Weaknesses of the Study**

The relatively small number of examiners who provided protocols is perhaps a limitation of this study, though the total number of protocols used compares favorably with most other similar studies in the literature. Also, the variety of training backgrounds (school and clinical psychology, many different graduate programs) may lessen the impact of the limitation stated
above. Another limitation of this study is that only Master’s level psychologists participated in this study, thus likely underestimating the rate of clerical error (Ryan, 1983). However, error rates in this study were similar to those in other studies (Sherrets et al., 1979; Wagoner, 1988).

Another possible limitation is that this study focussed only on clerical errors, and ignored scoring and administration errors. The error rate in this study therefore likely underestimates the actual rate of total examiner error.

Several key areas of strength were noted in this study. For example, unlike most research studies completed in this area, this study used actual protocols from practitioners. Many other studies have used either graduate students’ protocols, or fabricated protocols scored by volunteer professionals. The size of the sample of protocols also compares favourably with other studies. This study also examined examiner error from two new perspectives: very few, if any, studies to date have made comparisons between a test and its revision, or looked at the pattern of error rates over time.

Conclusions

Two central questions were addressed in this study. The first was regarding the change in the rate of examiner clerical error between the WISC-R and the WISC-III. The second involved the rate of error over time on the WISC-III, from the beginning of the test’s administration to a period 18 months later.
A claim made in the WISC-III examiner's manual (1991) states: "Changes from the WISC-R ... have made it easier for the examiner to use" (p. iv). I interpreted the claim to mean that the test, including the newly-revised protocol, has been designed so as to be less prone to examiners' clerical errors. Even if "easier to use" wasn't intended to be interpreted as reduced clerical errors, the problem of clerical errors is well-enough documented so that it should have been addressed in the design of the revised test.

The results from this study suggest examiners continue to make a substantial number of clerical errors on the WISC-III: 42% of protocols had clerical errors, with Full Scale IQ-changing errors on 23% of all protocols. The results suggest that there is no significant difference in clerical errors between the two tests, and therefore, the WISC-III does not seem to be easier to use than the WISC-R, if ease of use is interpreted as propensity to make clerical errors.

A more extensive retraining on the WISC-III might reduce the frequency of IQ-changing errors in the initial months of test administration, when these errors are most likely to occur. This retraining might not have a significant effect on reducing overall clerical errors, which in this study occurred at a high rate through WISC-R administration, and remained high throughout all three time periods of WISC-III administration.
Because clerical errors remain a problem with the WISC-III, and because the provided $SE_M$ does not account for these clerical errors, it is suggested that the amount of error in a test could be better indicated with a composite $SE_M$ which would include not only content sampling errors, but all other error types, including clerical errors. This suggestion, a variation of a suggestion made by Hanna et al. (1981), would provide test users with important additional information that is not presently published. The issue of reliability and $SE_M$ is not just a theoretical or psychometric issue, but one which is of practical significance. Scores from standardized tests are frequently given within a surrounding confidence interval, which is directly related to the reliability coefficient of the test.

This study also examined the rate of clerical error on the WISC-III over three six-month time periods, beginning with the first six months of administration of the test. It was found that although the overall clerical error rate did not change significantly between the three time periods, the Full Scale IQ-changing errors were significantly higher in the first time period than in the last time period. In other words, examiners made a relatively high number of serious (FSIQ-changing) errors at the beginning of their administration of the WISC-III.

It was earlier hypothesized that there would be no significant change in the rate of clerical error with examiners' increasing familiarity with the WISC-III. Although little, if any previous research was conducted in this area with practitioners, and studies...
with graduate students showed a decrease in error rates with experience, it was thought that practitioners who had no formal retraining would show special vigilance in avoiding errors with a new and somewhat unfamiliar test revision. This present study suggests that practitioners with no formal retraining make a relatively high number of IQ-changing errors at the beginning of the test’s use. It was found that the examiners who had the most difficulty with the transition from the WISC-R to the WISC-III were those with the most experience on the WISC-R. There is a clear need for the re-training of examiners who use the revised version of a standardized test.

Implications of the Study

Clerical errors continue to be a problem with the Wechsler scales. The results from this study suggest that examiner errors continue unabated in the WISC-III, with no significant reduction in either overall or IQ-changing errors. The revised protocol, with one possible exception (transformation of raw to scaled scores) does not seem to be less prone to clerical errors. And although one might attribute all clerical errors to examiner carelessness, and consider these errors unavoidable, valuable suggestions have been made which may lessen their rate and impact. Levenson et al. (1988) made the suggestion that a “revised record form might include directions to the psychologist to verify and recheck all response scoring, mathematical, administrative, IQ, and birth date computations” (p. 663). Suggestions like these have been wholly ignored on the WISC-III protocol, with the result that the error rate remains high and
unchanged. The WISC-III would be a more reliable instrument if more emphasis was placed by the test designer on avoiding clerical errors. A few key instructions on the actual protocol - perhaps a step-by-step guide to check and recheck scoring and mathematical computations - would likely result in more reliable scores.

Recently-introduced computer scoring programs for standardized tests may help reduce the incidence of some types of clerical errors. For example, the Scoring Assistant for the Wechsler Scales, introduced in 1992, completes all calculations for transformation of raw to scaled scores, addition of scaled scores, and transformation of scaled to IQ scores. The most frequent of clerical errors - addition of raw scores - would not be reduced through the use of a computer scoring program, nor would scoring or administration errors. Also introduced through the use of computer scoring programs would be another source of error: mistakes in transferring data from the protocol to the computer.

Another issue that has been largely ignored is the issue of the accuracy of the reported reliability coefficients and $SE_M$s. The results from this study underscore the need to include in a composite $SE_M$ not only the content sampling, time sampling and test administration errors suggested by Hanna et al. (1981), but to also include the component of clerical error.

Professional examiners are not often formally re-trained on revisions of standardized tests. For the most part, it has been
assumed that they possess the knowledge and experience to undertake the necessary changes in practice brought on by the test's revision. This study suggests that professionals make a relatively higher number of serious clerical errors when they are in the process of learning to use a revised test. It has been suggested (e.g., Franklin et al., 1982) that practicing psychologists are lacking in opportunities for continuing education. From the results of this study, it can be recommended that practitioners should undertake a more extensive retraining when a standardized test is revised, even when the changes appear to be largely cosmetic, as is the case with the WISC-III.
REFERENCES


and type of clerical errors made on WISC-R protocols. 
*Psychological Reports*, 62, 659-664.


