COGNITIVE STRATEGIES AND HEURISTICS UNDERLYING PSYCHOLOGISTS' JUDGMENTS ON THE WISC-R VERBAL SCALES: A PROTOCOL ANALYSIS

by

JOSETTE ANNE-MARIE PEROT

B.A., YORK UNIVERSITY, 1988

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

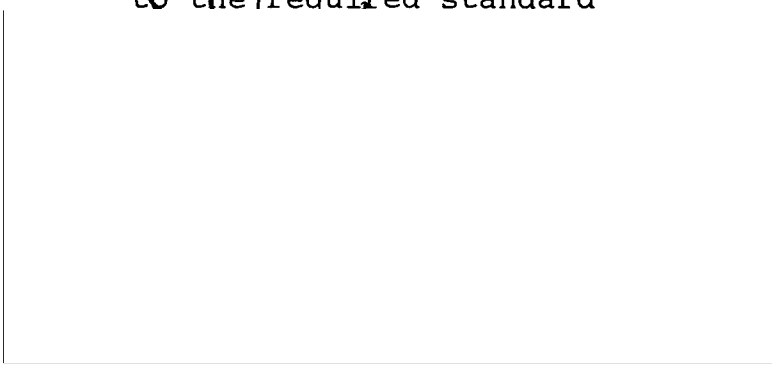
MASTER OF ARTS

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

March 1992

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) _____

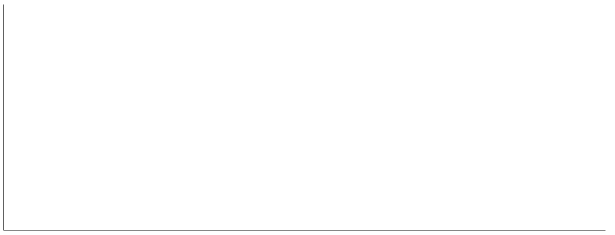
Department of Educational Psychology and Special Education
The University of British Columbia
Vancouver, Canada

Date March 31/92

ABSTRACT

The primary purpose of this study was to investigate psychologists' natural, interactive decision-making behaviour while scoring difficult verbal responses on the Wechsler Intelligence Scale for Children-Revised. A total of 23 psychologists participated in the study. First of all, in order to obtain scoring information descriptive of the sample, psychologists scored a WISC-R protocol. This protocol comprised four verbal scale subtests: the Vocabulary, Similarities, Information, and Comprehension subtests. In order of difficulty, the Vocabulary, Comprehension, Similarities, and Information subtests were found to be most prone to scoring differences. The Verbal IQ was found to vary by 11 points. Differences in point assignment within subtests accounted for variance in scoring. Following the completion of the first measure, a sub-sample of 8 psychologists provided think-aloud protocols in a separate session while scoring a second fabricated Comprehension subtest. The complexity of the task involved the consideration of administration errors and response judgment while scoring. Rather than focus solely on quantitative analysis of error differences as has been done in prior research, this study conceptualized these sources by providing additional analysis of specific strategies psychologists used while making scoring decisions.

The results of the verbal protocol analysis identified cognitive strategies inherent in the scoring of difficult type responses. The type and frequency of cognitive strategies identified in the study appear to be related to individual scoring accuracy. At the end of the session, psychologists were asked to identify strategies that were useful to them in difficult scoring situations. All psychologists identified the manual as the primary heuristic; however, percentage frequencies of verbalized strategies across subjects indicated that only four of the subjects used the manual as their primary aid on this task. These findings are further discussed, as well as their implications and inferences.



William T. McKee, Ph.D.

Research Supervisor

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to Dr. William McKee, my research supervisor, for his open door policy, and his unwaivering support throughout the course of this research, from the problem formulation, insightful commentary on the drafts, to the finished product.

I am also especially grateful to Dr. Marion Porath for her constant support, her valuable comments, insights, and critical analysis throughout the proposal and thesis stage that has contributed immeasurably to the quality of the work.

I gratefully acknowledge the contribution of Dr. Nand Kishor, who provided valuable feedback on my proposal and who planted the "conceptual seed" that helped formulate the framework for this study.

I wish to express my appreciation to Dr. John Slate for his interest in this study as well as for his generosity for permitting me to use his fabricated protocols for the purpose of this project.

I gratefully acknowledge the help of Drs. Suzanne Jacobsen and John Carter and the school board officials who have been instrumental in the subject solicitation process. A word of appreciation goes to the psychologists who volunteered for this study without whose efforts this study would not have come to be. Thanks folks!

Finally, I wish to acknowledge the love and encouragement I have received from my family: my parents, Elton and Victoria Perot, my sister, Giovanna, and my brother, Jean-Francois who have been patient in my academic pursuits from a distance. I also wish to thank Indar, who was here for the writing of the thesis, for his moral support and encouragement during this period.

CHAPTER 1

INTRODUCTION

School psychologists often make decisions under uncertain conditions. For example, when a child is under consideration for placement in special services, school psychologists must make judgments based on different sources of information, then weigh the probabilities and outcomes as to whether the child needs these services (Fagley, 1988, p.311). Given the prominence of the Wechsler Scales and standardized tests, one of the subareas within psychologists' profession which contributes to this uncertainty is their judgments regarding the scoring of verbal responses on the Wechsler Intelligence Scale for Children-Revised (WISC-R). Research has demonstrated that there is a high degree of subjectivity involved in the scoring of responses on the Verbal subtests (Slate & Hunnicut, 1988). The verbal subtests are prone to elicit problematic responses. These responses are usually ambiguous responses that demand considerable judgment on the part of the examiner (Brannigan, 1975) and therefore are difficult to score. Additionally, the difficulty of scoring "novel" responses has been widely acknowledged. Sattler (1988, p. 147) amplifies the challenged posed in the scoring of verbal responses as illustrated in Figure 1.

HOW WOULD YOU SCORE THIS?

Figure 1.
Scoring Dilemma
Used with the permission of Sattler

In this respect, it is inevitable that psychologists often have differences of opinion in their evaluations of the same response. Despite the knowledge that psychologists differ in their judgments of verbal responses, there lacks is a lack of descriptive evidence in the literature linking these differences to the actual judgmental strategies and heuristics that psychologists habitually employ in their task of scoring difficult-to-score verbal responses on the WISC-R. This is unfortunate since psychologists make extensive use of the WISC-R in their practice, and knowledge of the heuristics that they employ as well as their related thought processes may shed light on how they cope with areas that are not clearly delineated in the test manual. In the broader scope of psychologists' professional judgments in making complex decisions, Barnett (1988) calls for

conceptual links in order to analyze psychologist's behaviours in "problem framing, planning and implementing strategies" (p.667). In this regard, one may extend such conceptual links to the analysis of psychologists' task performance on the WISC-R from an information processing framework.

Context of the Study

As highlighted, the heuristics or "guiding strategies" that psychologists engage in have not been well articulated in the context of scoring WISC-R verbal responses. A probable reason for the absence of such data is that studies involving scoring differences on the WISC-R have not viewed the testing process from the perspective of the psychologist, that is, as a cognitive task requiring significant judgment. Consequently, with the focus on standardization and objectivity in the testing practice (Hanna, Bradley, & Holen, 1981; Slate & Jones, 1989), psychologists themselves have often been overlooked as an active part of the measuring process - a test process that calls for specific cognitive skills making often complex judgments. For example, the difficulty of the scoring task increases when the psychologist must make subjective judgments surrounding the "appropriateness" of a child's response to a certain test item. The role of subjectivity is especially increased by responses that are not clearly scorable by the test manual (Slate & Hunnicut, 1988).

According to Wechsler (1974), such exercises do indeed rely on the professional judgment capabilities of the examiner. Yet, psychologists are not specifically trained in optimal judgment strategies; therefore there may be a discrepancy between the guidelines in the manual and what psychologists actually do.

The Purpose of the Study

The purpose of this study was to investigate psychologists' judgments from the standpoint of cognitive psychology. Through analysis of verbal protocols this study investigated the strategies utilized by a group of psychologists while engaged in making evaluative judgments of difficult-to-score responses. The aim of the study was to describe the particular cognitive strategies arising from the verbal data. Since this study was mainly descriptive in nature, it is hoped that specific cognitive strategies identified in this research can be more fully detailed in the future.

Assumptions of the Study

An underlying assumption of this research is that the type and frequency of psychologists' underlying judgment processes affect the frequency of their scoring of verbal responses. DeNisi, Cafferty, and Meglino (1984) have suggested that some strategies lead to more accurate ratings in the appraisal process than others. Therefore, some psychologists may more ably appraise problematic responses

than others as a result of the efficiency and effectiveness of the judgmental processes they employ.

A second assumption is that the degree of clinical experience is not a factor in attaining accuracy. Both graduate students (Slate & Chick, 1989; Slate & Jones, 1989; Warren & Brown, 1972) and experienced psychologists (Brannigan, 1975; Miller & Chansky, 1972; Oakland, Lee, & Axelrad, 1975; Plumb & Charles, 1955) alike have been found to be prone to errors in scoring the verbal subtests. In other words, scoring difficulties do not seem to diminish with experience.

Justification of the Study

The WISC-R is one of the most commonly administered tests in clinical practice (Slate & Chick, 1989). Additionally, the extensive use of the WISC-R is reflected in the graduate classroom where it is the most frequently taught individual intelligence test (Slate & Chick, 1989). The use of the Wechsler Scales may be traced as far back as 1939 (Wechsler Bellevue Scales) which has afforded these tests a lengthy history in assessment (Plumb & Charles, 1955). Moreover, regarding the assessment of children, the Wechsler Scales have continued in the form of the WISC and WISC-R (and more recently, the WISC III). The functional use of the WISC-R has emerged not only as an IQ test but as the most widely used diagnostic tool in making important decisions regarding educational placement and special

services (Bradley, Hanna, & Lucas, 1980). Furthermore, especially with the emergence of the WISC III, it appears that the Wechsler Scales will continue to be used extensively in clinical practice as an important diagnostic aid.

Although psychologists go through extensive training on the WISC-R, they do differ in their scoring. Psychologists' scoring differences impact negatively on the integrity of the test scores which in turn affects the validity of subsequent decisions based on these scores. Thus, if the processes behind psychologists' differences in scoring can be studied, then greater understanding as to why there is such high variability in marking verbal responses may be helpful for training. Moreover, according to Pitz and Sachs (1984), "Errors in judgment suggest ways in which performance might be improved, especially if one understands why the errors occurred" (p.141).

Summary

It is apparent that scoring differences persist despite traditional training and detailed scoring guidance from the test manuals. The same differences have been encountered across several similar measures where verbal responses are scored (Warren & Brown, 1972). In light of this evidence, this study sought to investigate the cognitive strategies employed by psychologists as they engaged in the scoring of verbal scale responses.

The next chapter introduces a conceptual framework relevant to the interaction of cognitive psychology and the psychometric tradition. Included in the review of literature is a discussion on the usefulness of verbal reports as data since this form of data was used in the study.

Chapter 2

REVIEW OF THE LITERATURE

Theoretical Framework: Cognitive Psychology and the Psychometric Link

The process of making evaluative judgments and the strategies underlying this process have traditionally been studied within the field of work or organizational psychology. In such instances where judgment is required, employers are often called upon to evaluate employee performance (De Nisi, Cafferty, & Meglino, 1984; Kishor, 1987; Mount & Thompson, 1987; Murphy & Balzer, 1986). In employee evaluation two separate traditions that bear on the issue of the judgmental process have been studied (Feldman, 1981). These are the instrument-psychometric tradition and the social psychology tradition. The former deals with the study of random error and systematic biases in performance ratings while the latter focuses on the cognitive systems underlying attribution processes, person perception, and stereotyping. In the past, both schools of thought have been kept separate; however, more interest is being developed in the interactions between the individual's cognition in making a rating and the psychometric instrument that the individual uses. According to Krzytofiak, Cardy, and Newman, 1988:

> A significant body of appraisal research has approached the problem of error and bias by focusing on improved instrumentation. Because this attention to format has been of limited success, appraisal research has been focusing on the role of the rater as an information processor. (p.515)

The psychometric tradition has until recently overlooked the fact that the individual him/herself comprises part of the rating process. This is because the act of successfully making judgments requires individuals to sample from more than one source of information. Anderson (1977) sees this success as being contingent upon the "ability to interpret, integrate, and deferentially weight information to arrive at an appropriate decision" (p.68). Additionally, a number of investigators have combined both traditions to study the relationship between cognition and the processes of making rating judgments (Borman, 1977; DeNisi, Cafferty, & Meglino, 1984; Mount & Thompson, 1987; Murphy & Blazer, 1986). Questions such as, "[What] cognitive processes [or strategies] are engendered by the various types of rating scales ...?" are being asked (Feldman, 1980, p.128). And, how does a rater's cognition intercede with an evaluative judgment to produce a specific rating judgment? For example, in judging an employee's performance, schematic processing is a fundamental cognitive mechanism used in human judgment that may affect the final rating (Kishor, 1987). It has been suggested that an employer-rater has certain schematic categorizations that guide him/her to

notice specific employee attributes; these influence how he/she makes a rating judgment (Mount & Thompson, 1987). As in the perception of people, the perception of nonsocial stimuli which are ambiguous "is often determined by what the perceiver expects to see" (McArthur, 1981, p.204). For instance, decision frames may guide the way an individual conceptualizes "acts, outcomes, contingencies associated with a particular choice" (Tversky & Kahneman, 1984). Therefore, rating judgments on the same employee by different supervisors do not necessarily have to agree.

Along the same lines, Slate and Jones (1988) have suggested that psychologists may conceptualize verbal item responses differently and, as a result, need to learn to clarify response categories. Because psychologists may conceptualize information differently, it is possible that they may rely heavily on individual strategies and heuristics and exhibit variations in these processes to categorize responses. Such heuristics may be reflective of the systematic scoring patterns that fit their method of weighing and integrating information. Much like employee rating behaviour, these mental processes are a part of the WISC-R rating behaviour. However, the mode of judging will differ between tasks because of the basic task structure and requirements of each. For instance, Payne (1982) found that judgments change as the presentation of the task itself changes. Hence, one may infer that the mode of processing

information in WISC-R ratings is reflected by the contextual judgmental strategies employed, and knowledge of the task structure. However, since the study of judgment processes is a relatively new endeavour within the field of cognitive psychology (Rappoport & Summers, 1973), models describing encoding processes, judgmental heuristics and decision strategies are not as yet common. Most of the work in this area has been aimed primarily at providing descriptions of task heuristics in the hope that, at a later date, a "systematic theoretical presentation" can be developed (Pitz & Sachs, 1984, p.146). However, if appraisal practice is to advance, then further investigation into information-processing involved in making appraisals needs to be addressed (Cardy & Dobbins, 1986).

A Perspective: The WISC-R as a Cognitive Task

> The information-processing revolution that has taken place in cognitive psychology over the past 2 decades has been characterized by an increased emphasis on the processes rather than the products of task performance. The major goal of the task-analytic approach that has dominated the study of information processing has been to discover the elementary processes people use in performing tasks and to understand the strategies into which these processes and strategies act. As a result of this often successful pursuit of this goal, we now have a good understanding, at least at some level, of how people approach a large variety of tasks. (Sternberg & Ketron, 1982, p.399)

Information processing psychologists study the mind "in terms of mental representations and the processes that underlie observable behavior" (Sternberg, 1985, p.1). In other words, information processing theory attempts to

describe the processes and strategies that underlie human judgment and problem solving (Schulman & Elstein, 1975). These may be qualitative processes such as the strategies that individuals employ in acquiring information and the ways they use this information in certain problem-solving situations (Pitz & Sachs, 1984). Since "cognitive tasks vary dramatically in well-definedness and specificity" (Ericsson & Simon, 1974, p.119), it is possible to conceptualize the scoring of WISC-R verbal responses as a cognitive task; the scoring of responses of the Verbal Scale requires psychologists to actively process, weigh, and integrate information when judging a response.

To reiterate, an advantage to this approach is that cognitive activities can be related to observed performance or behaviours. Therefore, one can then make inferences as to what strategies the individual used to perform the task. Classic studies of problem-solving in chess (Chase & Simon, 1973; de Groot, 1965,1966; Simon & Chase, 1973) and physics (Chi, Feltovich, & Glaser, 1981) from the perspective of expert and novice knowledge bases (cognitive content methodology approach) have provided insight into cognitive processes. Such an approach often studies the comparative performance between experts and novices in different content domains. More recently, an information-processing framework has been found useful in research on human performance in clinical diagnostic settings such as medicine and in

interactive instructional contexts, such as teaching (Fogarty, Wang, & Creek, 1983). These studies have been useful in describing the interaction between knowledge, cognitive processes and heuristics in problem-solving and decision making. Sternberg (1985) summarizes the basic questions of interest to researchers of this persuasion:

1. What are the mental processes that constitute intelligent performance on various tasks?
2. How rapidly and accurately are these processes performed?
3. Into what strategies for task performance do these mental processes combine?
4. Upon what forms of mental representation do these processes and strategies act?
5. What is the knowledge base that is organized into these forms of representation, and how does it affect, and how is it affected by, the processes, strategies, and representations that individuals use? (pp.1-2)

Psychologists' Task Performance on the WISC-R

Questions such as those posed by Sternberg may be helpful in the investigation of the factors involved in psychologists' evaluations and scoring of WISC-R responses. That is, the study of the underlying cognitive variables that affect psychologists' WISC-R task performance may be fruitful in conceptualizing errors in scoring. The traditional manner of conceptualizing psychologists' performance on their use of the Wechsler Scales has been from a quantitative or psychometric perspective. This approach has tallied the number and kinds of errors

committed by psychologist-examiners. The usefulness of this approach has been that it has identified and quantified a significant problem within the psychological profession; that is, the Wechsler Scales have been found to lend themselves to highly variable scoring because of the difficult-to-score responses that they elicit. However, despite the recurrent scoring problems evidenced in the Wechsler Scales, studies focusing on these tests have simply continued to acknowledge this problem through the documentation of quantitative data underlying scoring accuracy in test protocols. Such studies have not focused reseach on understanding the nature and causes which give rise to these problems.

Problems of the Verbal Scales

The problems of scoring the Wechsler Verbal Scales have been acknowledged since the inception of the Wecshler-Bellevue Scales in 1939 (Plumb & Charles, 1955). Scoring problems are also apparent in the WAIS, WAIS-R and the WISC as well as other standardized tests. An extensive body of research has shown that psychologists frequently commit serious errors when administering and scoring test protocols (Franklin, Stollman, Burpeau, & Sabers, 1982; Hunnicutt, Slate, Gamble, & Wheeler, 1990; Miller & Chansky, 1972; Miller, Chansky, & Gredler, 1970; Oakland, Lee, & Axelrad, 1975; Plumb & Charles, 1955; Slate & Jones, 1989; Walker, Hunt, & Schwartz, 1965; Warren & Brown, 1972).

More particularly, the most problematic tests have been shown to be those which comprise the Verbal Scale (Oakland, Axelrad & Lee, 1975; Slate & Chick, 1989; Slate & Jones, 1988). The studies that have examined the nature or types of errors on the Wechsler Scales demonstrate with an overwhelming consensus that the Verbal subtests are the most difficult to score and that the source of much variability in scoring stems from these subtests rather than from the Performance subtests. Slate and Jones (1988) ranked the WISC-R subtests in terms of scoring difficulty. In the order of most difficult to least difficult to score, these were the Vocabulary, Comprehension, and Similarities subtests. Information was ranked seventh out of the ten scales. Digit Span and Mazes were omitted.

The problem of scoring protocols has been illustrated in studies where psychologists have been given identical protocols to score and have awarded different scores to the same items (Slate & Jones, 1988). The nature of these errors usually involved giving more credit than required. For example, psychologists are more prone to giving 2 points for a one point response, and 1 point for a 0 point response (Slate & Jones, in press). Errors also occur in failing to record subject responses verbatim (Warren & Brown, 1972), as well as through differences in questioning of ambiguous subject responses (Brannigan, 1975). Additional questioning usually occurs for item responses that are not clearly

scorable by the test manual. Although a wide range of empirical evidence addresses the impact of scoring difference on the Full Scale and Verbal IQ, a brief review of some studies relating to this problem is warranted.

In an early study, Miller, Chansky, and Grendler (1970) investigated the degree of agreement among 32 psychologists-in-training in the scoring of WISC protocols containing fabricated responses. Although the authors hypothesized that ratings would be highly comparable, they found a wide range of scoring. The full scale IQ ranged from 76-93. They also found that verbal subtests lend themselves to highly variable scoring. The Comprehension and Vocabulary subtests were found to be most vulnerable to scoring errors.

A later study was conducted by the same authors (Miller & Chansky, 1972) in which they investigated the agreement among professionals in the scoring of WISC protocols. Surprisingly, professional psychologists seemed to fare no better. Sixty-four professional psychometricians scored identical WISC protocols. Again the greatest interscorer variability was produced by the Verbal subtests. This same protocol elicited an IQ range from 78-95 points which indicates that psychologists typically vary in their scoring. The authors commented that psychologists seem to use additional criteria other than the manual, however they did not expand on these criteria. Similarly, Kasper,

Throne, and Schulman (1968) have suggested that individuals may rely more readily on memory and experience than on the manual as they gain experience.

Again, in a study involving 94 psychologists, Oakland, Axelrad, and Lee (1975) found the Verbal Scale to have a lower interrater agreement than the Performance Scale on WISC protocols.

Additionally, Babad, Mann, and Mar-Hayim (1974) investigated the effects of experimenter bias. Eighteen graduate students scored a prepared protocol. They were told that the responses were those either of a under-achieving disadvantaged child or a high-achieving upper middle class child. For both subtests, results indicated that means for the Comprehension subtest differed significantly, as well as the Verbal IQ score.

As seen, variability in performance on the Verbal subtests is a serious and recurrent problem, such that significant IQ discrepancies have often been brought to light after corrections. This is extremely worrisome since intelligence tests are routinely administered to children who are functioning at a marginal level (Boem, Duker, Haesloop, & White, 1974; Warren & Brown, 1973).

Nature of the WISC-R Verbal Scales

The WISC-R Verbal Scale consists of six subtests: Information, Comprehension, Arithmetic, Similarities, Vocabulary, and Digit Span. However, only Information,

Comprehension, Similarities, and Vocabulary will be described since these were the subtests used in the study. These subtests are untimed.

Information measures memory of a wide range of general information and knowledge gained from experience and education (Sattler, 1988;Searles, 1975; Truch, 1989). Such information gives the psychologist an idea of the child's "general range of information, alertness to the environment, social or cultural background, and attitudes towards school and school-like tasks" (Sattler, 1988, p. 147). The nature of the questions asked pertains to "questions concerning names of objects, dates, historical and geographical facts, and other such information" (Sattler, 1988, p.147). An example of a Similarities item is, "What are the four seasons of the year?" [item 11]. A more difficult question is, "Who was Charles Darwin?" [item 29]. The starting point of the test is determined by the age of the child. Each item is either given 0 or 1 point depending on the quality of the response. The psychologist is allowed to question the child by saying "Explain what you mean or Tell me more" if the response is not clear. This subtest consists of 30 items, and is discontinued after 5 consecutive failures.

Similarities subtest measures essential relationships between facts and ideas, namely associated relationships between word-pairs (Searles, 1975; Truch, 1989). The Similarities subtest consists of 17 items. For each item,

the psychologist asks the child to differentiate between two words. For example, for the first item of the Similarities subtest, the psychologist asks, "In what way are a wheel and a ball alike? How are they the same?" They are both round would be a correct answer (Wechsler, 1974, p.74). All children are administered the first item. The first four items are either given 1 or 2 points. Additional items are given a score of either 2, 1, or 0 depending on the sophistication or conceptual level of the child's response. Two points are given for a general classification which is primary to both words, 1 point for less pertinent but specific properties common to both words, and 0 points are given for clearly wrong responses. For specific items, additional questioning is permitted to clarify ambiguous responses. This subtest is discontinued after 3 consecutive failures.

Vocabulary consists of words that need to be defined. This subtest measures learning ability, word knowledge acquired from experience, education, richness of ideas, kind and quality of language, and level of abstract thinking. This subtest is considered to be the best single measure of intelligence of all the subtests (Searles, 1975; Truch, 1989). The Vocabulary subtest consists of 32 items arranged in increasing order of difficulty. The psychologist asks "What does _____ mean? Or what is a _____? (Wechsler, 1974, p.89). A score of 2, 1, or 0 is credited

to each item depending on the level of sophistication of the response. Examples of a 2 point response are, "a good synonym","a major use", "definitive...or primary features of objects" (Wechsler, 1974, p.161). One point is given for partially correct responses, synonyms that are less pertinent, or a definition of a minor use of an object (Sattler, 1988, p.151). Psychologists are allowed to question vague responses. Some responses must be queried, if a (Q) appears in the scoring rules, and some responses indicated in the manual, must be scored without further questioning or clarification. Similar to the Information subtest, the starting point of this subtest is also determined by the age of the child.

Comprehension questions reflect the child's level of moral development and understanding of surrounding societal conventions. Success depends on social judgment, practical information, as well as knowledge pertaining to past experiences in reaching solutions (Sattler, 1988, p.153). Knowledge of one's body and interpersonal relations are also reflected in the questions. This subtest consists of 17 items. Examples of items are, "What is the thing to do when you cut your finger?" [item 1], or, " What are you supposed to do if you find someone's wallet or pocket-book in a store?" [item 2]. Responses to items are either scored 2, 1, or 0. The child must express at least two of the general ideas listed in the manual in order to be awarded 2 points.

The child receives only 1 point for one idea, and 0 points for an incorrect response. The psychologist is permitted to question vague responses in accordance with the querying procedures in the manual. Additionally, if the child replies with only one idea, the psychologist may ask for a second response. This subtest is discontinued after 4 consecutive failures.

Task Summary

Each of the above subtests requires the examiner to present test items orally to the child. The test questions are presented as written in the manual so that the examiner does not depart from standardized procedures. The child is expected to answer and the examiner immediately records the child's response as accurately as possible in the test booklet. The examiner is not expected to indicate the appropriateness of the child's answer by providing feedback. However, if the examiner is uncertain as to what the child has said, the examiner may ask the child to repeat the response in order to clarify ambiguous responses. If the child's response is absolutely wrong as stated in the manual then the child's response should not be questioned further. The examiner is expected to insure that the child is comfortable in the testing situation so that the child may do his/her best. The examiner should be aware of instances of distractibility that may affect performance. The examiner can increase his/her awareness of the overall

testing situation by knowing the "task well enough, so that the test flows almost automatically, leaving [the examiner] maximally free to observe all aspects of the child's behavior" (Sattler, 1988, p.103). Each subtest is discontinued when the child reaches the ceiling.

Although this study focuses on the scoring judgment aspect of the test administration process, the procedures leading up to and following this phase are outlined for their contextual value.

The Administration Process

In the administration process, the psychologist usually progresses through various decision points (see Figure 2). During an input phase, the psychologist must administer the

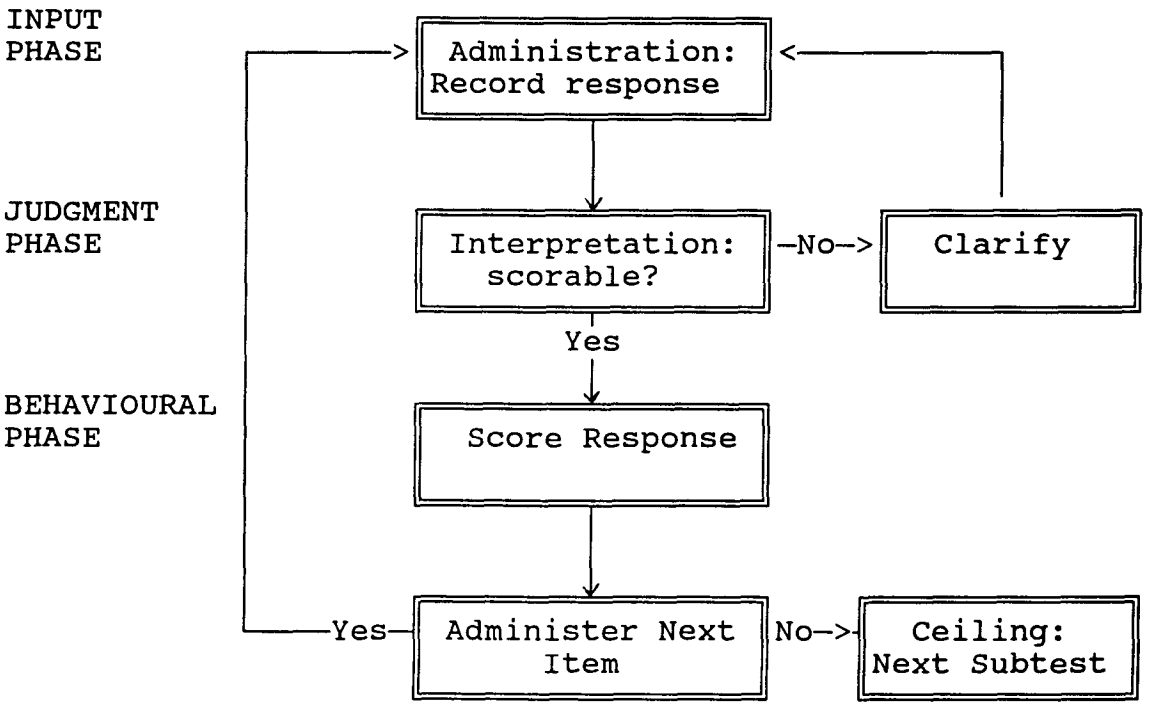


Figure 2. Model of Psychologists' Judgmental Process

test item as well as record the child's response as accurately as possible.

The judgment phase of the process involves reading the child's response, and determining whether the response is scorable according to standardized procedures. If the response is clearly scorable, the psychologist may automatically proceed to the behavioural phase and score the response. As established in the literature, in many cases the verbal response may not be clearly scorable by the manual and hence be difficult to judge. Therefore, in the actual testing situation the psychologist must seek clarification with additional questioning and record the new response. The psychologist must then reinterpret the answer to the item in light of the new response given by the child. The psychologist may then score this new response. What the psychologist actually does at this point - the interpretive phase of the model - is the focus of this study.

Finally, the psychologist proceeds to the next item and the administration cycle begins again. If a ceiling is obtained, the next subtest is administered instead.

An underlying assumption is that a smooth administration requires a high level of competence. The psychologist must be able to access specific knowledge relevant to task performance. "Intelligent performance of complex tasks means doing the tasks correctly, with little or no waste [sic] motion - with few or no mistakes or detours along the

way" (Simon, 1976, p.65). Psychologists must then have a solid foundation of declarative knowledge, [the background or factual knowledge of a particular domain (Gavelek & Raphael, 1985)] as well as procedural knowledge. Procedural or conditional knowledge refers to knowing the specific steps involved in carrying out the task (Lesgold, 1990). Expert performance requires an enormous amount of such knowledge (Simon, 1976).

The task domain related to the scoring of verbal responses is knowledge of the basic rules needed to perform the task properly. These basic rules may include knowledge of the scoring procedures, knowing the starting point of each subtest according to the age of the child, when a ceiling is obtained, and, when probing for clarification is appropriate. Such knowledge would determine how well the model works for each psychologist. However, it is necessary to mention that the basic declarative knowledge necessary to do well on the task may be interactive with other sources of complex declarative knowledge that are related to the field, but not necessarily task-specific. An example of this type of knowledge would be what the psychologist already knows about the child from school reports, or teacher conferences. Such knowledge may influence psychologists' interpretations of responses. For example, Sattler (1988) states that "the child's performance on the WISC-R should be interpreted in relation to all other sources of data" (p.182). However,

when this information is known beforehand it may influence scoring responses, especially in instances where responses are marginal.

Cognitive Strategies and Heuristics

One of the ways in which cognitive psychologists study the manner in which individuals treat information is by investigating the strategies that individuals employ when heeding information during problem solving. However, as Webb (1975) suggests, "The language used to describe problem-solving processes is cumbersome. The meaning of strategy or heuristic varies from study to study" pp.103-104). Tversky and Kahneman (1983) define judgmental heuristic as "a strategy - whether deliberate or not - that relies on a natural assessment to produce an estimation or prediction" (p.294). Additionally, Burns (1990) defines heuristics as "cognitive shortcuts" (p. 343) and Fischoff (cited in Kahneman, Slovic, & Tversky, 1982) extends this definition to include individual strategies or non-optimal rules of thumb which are effective in some cases in guiding judgments. For the purpose of this study Fischoff's definition provides the definitional framework. This is because one may think of a heuristic as a cognitive strategy that sometimes leads to systematic bias in making judgments. In other words, not all cognitive strategies are effective in bringing about appropriate judgments because of an incorrect problem-solving procedure - these types of

strategies are called heuristics.

An example of a biased cognitive strategy, or heuristic is illustrated in the errors that children sometimes make in mathematical problem-solving. Buggy algorithms are apparent where children fail to borrow in subtraction problems (Van Haneghan, Baker, 1989, cited in McCormick, Miller, & Pressley). For example, Brown and Burton (1978, cited in Gagne, 1985) found that some children consistently used the incorrect procedure of subtracting the two numbers in each column. These children ignored the position of the smaller number. For example, if the smaller number was in the top position of the column the children still proceeded to subtract the larger number from the smaller in order to find the difference without first borrowing. It may be suggested, then, that psychologists' judgments may involve the application of similar types of biases in the difficult task of scoring verbal responses.

The next section discusses the limitations and advantages of the cognitive psychology laboratory approach to study heuristics.

Cognitive Psychology and the Laboratory Method

One criticism of cognitive psychology is that the study of strategies that individuals employ in the laboratory lacks external validity; that is, these strategies may not necessarily generalize to everyday problem solving tasks. Laboratory experiments give rise only to "anecdotal

evidence" (Galotti, 1989) in relation to people's actual heuristics away from the laboratory (Burns, 1990). For instance, Galotti (1989) describes the limitations of the laboratory approach to problem solving in the following way:

> The premises are usually already identified, the amount of irrelevant information...[is] restricted, the number of inferences to be performed...[are limited] to one or a few, there often exist normatively correct answers. (p.343)

In making decisions under uncertain conditions, trained professionals performing tasks in their fields, do rely on judgmental heuristics to make their exercises easier (Fagley, 1988). However, since all aspects of a problem situation cannot be studied at once, an advantage of the laboratory approach is that external variables can be controlled so that an identifiable aspect a problem can be more ably studied. This is especially useful in an exploratory study. For example, in this study all subjects were given the same protocols to score under similar conditions. This methodology allowed for the comparative study of the cognitive strategies and heuristic processes among psychologists engaged in the same task.

Therefore, one may speculate that, in scoring the verbal responses, psychologists do have systematic ways of indexing information. Where little ambiguity exists, assignment of a stimulus to a category should be an automatic process for most psychologists (Feldman, 1981). Judgmental strategies in these instances are instantaneous and perhaps similar

across psychologists. A response that is easily decipherable and not cognitively demanding will automatically be rewarded a consensual point value. On the other hand, heuristic processing - a more rudimentary application of judgment - is more prevalent when a judgmental situation is cognitively demanding. Strategies in these instances are more deliberate (Pitz & Sachs, 1984) such that, in the face of difficult-to-score responses, one may speculate that psychologists do sometimes find it necessary to use certain "rules of thumb" or heuristics, for example, referring to the test manual to match with similar examples. Because heuristics often result in systematic errors or biases (Svenson, 1985), individual psychologists may become insensitive to variations in data which may account for scoring errors. Interestingly, Kasper, Throne, and Schulman (1968, cited in Conner & Woodall, 1983, p.378) have suggested that, as a psychologist generally becomes more experienced in the scoring of WISC-R protocols, "s/he may rely more heavily on his/her memory than on the manual...resulting in individual scoring patterns" .

Cognitive psychologists generally refer to tasks as problem situations to which a solution is sought. In the context of this study, such a solution is a judgment choice regarding a specific point value to award a response. Pitz and Sachs (1984) reiterate that "[whenever] information processing occurs as part of the [judgment and

decision-making process], the only observable behavior is a response - usually a... choice" (p.152). In order to better understand why individuals make certain choices, psychologists try to trace the solution paths by analyzing the underlying thinking and mental strategies that are associated with solving a particular task. One way of studying the thinking processes involved in the judgmental process has been by means of verbal protocols (Ericsson & Simon, 1984; Klein, 1983; Pitz & Sachs, 1984).

Verbal Protocol Analysis

The seminal work of Ericsson and Simon (1984) has drawn attention to the benefits and the applicability of verbal reports to investigate the underlying cognitive processes in decision-making tasks. The technique of concurrent self-reports such as thinking-aloud and talking-aloud techniques have traditionally been used to provide this verbal data. The self-report technique requires that subjects verbally express all thoughts which come into their minds as they perform a task (Ericsson & Simon, 1984). In thinking-aloud, the more complex of the two self-report techniques, subjects are asked to verbalize both simple and complex thoughts while engaged in the particular task. Complex thoughts may include detailed information pertaining to sub-goals, goals, motives, reasons, and comments on the domain-specific knowledge necessary to complete the task. Additionally, think-aloud reports are detailed enough that

decision rules to the solution process can be inferred. Alternatively, subjects may also be asked to report these decision rules (Crow, Olshavsky, & Summers, 1980; Klein, 1983) or to report any hypotheses they used in problem-solving (Ericsson & Simon, 1980). As opposed to think-aloud techniques, talk-aloud techniques are most useful when the experimenter is interested in general types of information related to cognitive processes. This technique requires subjects simply to say out loud whatever they are saying silently to themselves in a problem-solving episode. Although there appears to be an overlap between think-aloud and talking-aloud techniques, they do seem to differ in the conceptual level of information they generate. This is because the type of instruction and probing questions asked by the experimenter guides the subject as to whether events should be reported in general or more particular terms. This affects the depth of information present in the verbal protocol. However, the usefulness of both techniques are such that they:

> Can reveal in remarkable detail what information [subjects] are attending to while performing the tasks...[they] can provide an orderly picture of the exact way in which the tasks are being performed: the strategies employed, the inferences drawn from information, the accessing of memory by recognition". (Ericsson & Simon, 1984, p.220)

On the other hand, the underlying assumption that verbal reports are a reflection of what is in awareness or working

memory has been the subject of criticism by some cognitive psychologists. Nisbett and Wilson (1977) hypothesized that the very act of verbalizing while engaged in the task changes the task environment and, therefore, the nature of the data. However, Ericsson and Simon (1984) have argued that verbal reports are not altered by thinking aloud, and therefore will not interfere with ongoing thinking processes. For example, Newell and Simon (1972) compared the think-aloud protocols of 7 subjects discovering proofs in propositional logic exercises to 64 subjects under the same conditions. When the structure of search trees and solution paths were compared, no differences were found between the two groups. Other work is in agreement that verbal protocols do not change judgmental behaviour (Payne, 1980; Karph, 1973). Earlier research found that verbalizing actually aids in improving performance in terms of uncovering general problem-solving principles and new reasons for specific choices (Benjafield, 1969; Dansereau & Gregg, 1966; Davis, 1968; Gagne & Smith, 1962).

Methodology of Verbal Protocol Analysis

Ericsson and Simon (1984) make the important point that "thinking aloud does not by itself enforce an analytical approach" (p.88) to understanding cognitive processes. In order to bring some level of conceptual understanding to verbalized thoughts, this raw data must be treated in some manner in order for conceptual interpretations to be made.

The question, then, is, "How do we characterize cognitive structures, or thoughts?" In verbal protocol analysis, thoughts are usually characterized by separating the protocol into smaller units, or segments.

Segmentation

Segmentation refers to the breaking apart of an entire protocol into smaller units. Protocols may be segmented in different ways depending on the nature of the study and the research question. However, when a protocol is segmented, each segment usually represents one instance of a general cognitive process (Ericsson & Simon, 1984). According to Ericsson and Simon (1984), the appropriate and most frequently used cues in segmentation are: "pauses, intonation, contours...as well as syntactical markers for complete phrases and sentences - the cues for segmentation in ordinary discourse" (p.205). In some cases, relying solely on this type of segmentation may be inappropriate. Where the actual content of the protocol is the important factor, it may be more appropriate to use idea or semantic units as the criteria underlying segmentation (Smith, 1971). Therefore, a segmented element in a protocol may be defined as a semantic unit judged to represent a complete thought. After each protocol is segmented, the next step is to encode these segments.

Encoding

The process of encoding may be described by the actual

matching of a segment to a category. In model-based encoding categories are usually already defined. The choice of categories may be based on existing theory, or categories already existing in the literature (Glaser, 1978). Alternatively, categories may be constructed through knowledge and procedures of the experimenter, such as pilot studies (Ericsson & Simon, 1984; Kilpatrick, 1968). For example, in a study investigating the relationship between the thinking aloud technique and problem solving ability of mathematics problems, Flaherty (1974) was able to devise and revise categories that were appropriate to the task through a pilot study. A schematic representation of this process is seen in Figure 3.

Input----->Encoding----->Output

(Segment) (Category)

Figure 3.

Encoding Process

Adapted from Ericsson & Simon, 1984, p.276

In contrast, where a study is exploratory and appropriate categories do not exist or existing categories are not suitable, it is not uncommon to segment and code simultaneously (Kilpatrick, 1968; Glaser, 1978; Glaser & Strauss, 1967).

The next section describes method for deriving validity

of verbal data.

Deriving Validity from Think-Aloud Data

The type of validation procedure used in verbal report methodologies generally depends on the nature of the study, and the question asked. However, data obtained from verbal reports generally begins with the credibility of constructing conceptual codes from transcribed data (Glaser & Strauss, 1967). Thus, validation usually involves the comparison between the coded verbal report and some other type of measure that refers or is related to the same events in the same fashion (White, 1980). This procedure allows for a valid comparison between two related events, each concerned with the same question.

In studies involving numerous subjects and where quantitative data is available, external validity may be measured by predicting success on a criterion variable. For example, in a study involving problem-solving patterns of 8th grade students, Kilpatrick (1968) used verbal protocol data coded according to two schemas, heuristic strategies and processes. He used the methodology of correlating the verbal data with performance scores, such as solution times and solution scores. Through this method he was also able to refine his data by eliminating artifactual correlations resulting from properties inherent in the coding system.

Validation may also be a measure of a subject's

self-awareness of the particular activity studied (Smith & Miller, 1978). In a study involving students' ability to report their decision criteria surrounding the choice of what college to attend, Berl, Lewis, and Morrison (1976; cited in Smith & Miller, p.360) found that students were able to report on the cognitive processes that underpinned their decision concerning what school to attend. They found that the students' reports correlated well with reports relating to the actual criteria that governed their decisions.

Another way to validate verbal protocol data is to construct a subject's problem space from the initial step to the solution in such detail that each minute step in problem solving can be evaluated. Newell and Simon (1972) used this methodology when they investigated subjects' capabilities to problem-solve crypto-arithmetic problems. Subjects were required to decode arithmetic problems where digits were replaced by letters. In this way the extent of the subject's awareness of problem-solving could be determined. It was found that subjects were able to clearly report on their steps and also give reasons as to why they were undergoing certain steps. In exercises as detailed as these, Smith and Miller (1978) report that "there is no reason to believe that anything is going on besides what subjects can report" (p.360). However, White (1980) warns that since an individual's cognitive activity and strategies

are the variables of interest, comparison between group means in control and experimental situations does not best represent the nature of the question.

Thinking Aloud Methodology

The motivation of employing a thinking aloud methodology in this study was to acquire a veridical description of the thought processes involved in making scoring judgments and the explanations underlying these judgments. In a study involving intransitive preferences, Montgomery (1977) found "that think-aloud protocols from single individuals can give valuable information about decision processes in asmuch as it is possible to describe the [subjects'] choices by means of choice rules that were derived from the think aloud reports" (p.360). Additionally, the thinking-aloud technique has been used successfully to study cognitive processes reflecting financial decisions of bank trust officers, chess moves of chess players, and diagnostic reasoning of clinical psychologists and physicians (Peterson, Marx, & Clark, 1978).

Ericsson and Simon (1984) have found that verbal reports from even a small number of subjects are useful in terms of generality of cognitive processes and strategies within-subjects and generalizability between-subjects over tasks.

Lastly, since the scoring of the WISC-R verbal subtests is based on psychologists' professional experiences and

activities, familiarity of the task environment is a methodological advantage with regard to the facility and the accuracy with which subjects report their thoughts. White (1980) suggests that:

> The making of the judgment and of the report should be a matter of some ease for the subject, not a task in itself requiring concentration and effort. The subject should not have to be preoccupied with the mechanics of an unfamiliar task or the problems of comprehending difficult instructions. (p.107)

Summary

This chapter has focused on the importance of WISC-R scores in psychological testing. The literature reviewed established that psychologists make errors in scoring, particularly on the Verbal subtests. The types and frequencies of errors have been identified in previous research, but the underlying reasons have not been analyzed.

In order to help explain these differences in scoring, a cognitive psychology framework was adopted, and the nature of research in cognitive psychology on other tasks was represented. One assumption is that the task of scoring presents cognitive demands which might account for differences in scoring.

Thirdly, the methodology of verbal protocol analysis was reviewed.

The research questions are presented below:

1) What are the common mental strategies that underlie psychologists' judgments in their approach to the scoring of

the same task?

2) To what degree are their strategies similar?

3) To what degree are psychologists aware of their own cognitive strategies?

In the following chapter the methodology for the investigation is described.

CHAPTER 3

METHODOLOGY

This chapter presents a description of the sample of psychologists who volunteered for this study. Also presented is a description of the procedures underlying the development of the categories for the verbal data, the analysis of this data, and the training of the independent coder.

Sample

A sample of 23 psychologists was solicited from 4 school districts in the Greater Vancouver area. The majority of the psychologists worked primarily within the school system. Two also held positions in hospitals. Subjects were solicited through representatives in their respective school districts. The names of potential volunteers were then given to the investigator. The investigator contacted each person by phone in order to gain their participation in the study.

The median number of years that subjects worked as psychologists was 5.00 years. The psychologists' levels of education varied. One psychologist had a bachelor's degree; fifteen were master's level psychologists; and the remainder were doctoral level psychologists. Seventy-eight percent of the psychologists were trained on the WISC-R as well as the

Wechsler Preschool and Primary Scale of Intelligence (WPPSI), 70% also had training on the Stanford-Binet IV. Seven psychologists described themselves primarily as school psychologists, two as educational psychologists, two as developmental psychologists, and one as a special educator. The remaining 11 in the sample described their jobs as eclectic in nature. That is, they worked in the capacity of at least two of the following categories: school psychologist, educational psychologist, counselling psychologist, or special educator.

Procedures

Session 1: The 23 psychologists were mailed a completed WISC-R protocol that contained fabricated responses from four verbal subtests (Information, Similarities, Vocabulary, and Comprehension). Also included in the package were two consent forms (one to be retained by the participant) which described the nature of the study and a background information form. The subjects scored the subtests at their own convenience and mailed them back to the investigator with the background form and a copy of the consent form. The psychologists were asked also to indicate on the consent form whether they were willing to participate in the think-aloud session. Each psychologist was assigned an identification number and all data was coded with this number in order to preserve subject confidentiality.

The Stimulus Protocol

The stimulus protocol is a fairly new instrument. It is one of a series of protocols developed by Dr. J. Slate of Arkansas State University. The protocols are part of an unpublished text, Guide to administering and scoring the WISC-R (Slate, 1991). The protocols are currently being used for research purposes at other universities and for training in Dr. Slate's Intelligence Testing course.

The protocols were employed in a study by Dr. Slate in the summer of 1991. Although the results have not been written up as yet, the mean error rate per protocol was found to be about 3 per protocol. The protocols were constructed to be as difficult as possible to score correctly (J. Slate, personal communication, Nov. 22, 1991).

Session 2: A subsample of 9 psychologists participated in session two. These persons were contacted to set up a time and a place of convenience to participate in the think-aloud exercise. Usually, the investigator met with the psychologists in the district for the exercise. One protocol spoiled yielding a total of 8 protocols for session two. The median number of years that subjects worked as psychologists was 5.00 years. Five were master's level psychologists, the other three were doctoral level psychologists. The sample described themselves mainly as district school psychologists. They all had training on the WISC-R as well as on the Stanford-Binet IV; all except one

psychologist had undergone training on the WPPSI.

A Comprehension subtest, WISC-R II measure (see Appendix B), was selected for this think-aloud session since it is one of the verbal subtests that have been shown to produce a large amount of scoring variability (Slate, 1988). Although the Vocabulary subtest has been shown to produce the most variability in scoring, this test was not appropriate due to its length. Time constraints were a concern of the subjects.

For this exercise, subjects were first given a brief warm-up think-aloud task. This task involved analogy-type questions that required the psychologist to reason out loud. In order to acquaint them with the think-aloud method, the experimenter first demonstrated this exercise for the subject. When subjects were comfortable with the think-aloud method they then proceeded to the actual think-aloud task on the Comprehension WISC-R II measure, that is, to verbalize what they would normally be thinking as they were judging a response. At the end of the session, the subjects were asked this final probing question, "Were there any strategies that you were conscious of that aided you in deciding what point value to award a response?" Such additional information acted to validate the think-aloud data through the comparison of actual performance on the task in response to the third research question. The duration of session two was, on average, 25 minutes (see

Appendix B for complete script for warm-up exercise).

The next section describes the treatment of the verbal data derived from this sample.

Development of Verbal Categories

The methodology behind category development first involved transcribing all tape-recorded responses of each protocol into written text. The written text was then segmented into semantic units for analysis. A semantic unit was previously defined as a phrase or sentence representing a complete thought (Smith, 1971). All units were coded from categories that were derived from the data itself. It is customary to derive coding categories from the data present in the protocols themselves (Ericsson & Simon, 1980). According to Glaser (1978) it is desirable to enter research without predetermined ideas; this methodology allows the investigator to remain open to the data generation process.

> Merely selecting data for a category that has been established by another theory tends to hinder the generation of new categories, because the major effect is not generation but data selection. (Glaser & Strauss, p.37)

The analysis of verbal protocols proceeded in three stages. In the initial stage, before segmentation, a pilot study was conducted involving a logical task-analysis of think-aloud protocols of a student-psychologist and a practising psychologist. The basic coding scheme was derived from this method. Seven basic coding categories

that reflected trends across both protocols were derived: self-questioning behaviour, self-regulatory behaviour, general metastrategic statements, memory, manual, and recommendations/evaluations.

The second stage of category development involved category finetuning. This stage involved the analysis of protocols that were subsequently collected, and then segmented into semantic units. The basic method of finetuning involved the matching of examples (segmented units) from these protocols with an appropriate category derived from the pilot study. The investigator noted instances where an example could not be matched in a category, or seemed to fall within two categories. Consequently, two categories were altered, and two additional categories were added to the set. Self-questioning behaviour and self-regulatory behaviour were collapsed into monitoring statements, recommendations and evaluations were also collapsed into one category, and planning behaviour and self-explanations categories were added. Again, a total of seven categories was obtained. Such revision is not uncommon in the treatment of verbal data. According to Glaser and Strauss (1967), the methodology of jointly collecting, coding, and analysis of data should be an interactive process.

Analysis of Semantic Units

Each protocol was first segmented according to idea or

semantic units. For illustrative purposes, an example of the segmented units from two protocols as well as the complete protocols are presented in Appendix C. Each segment was then assigned to a category by two coders. In order to prevent coders from using contextual information and to preserve objectivity in the coding scheme, all segments were printed on separate cards and then randomly coded.

Training of the Coder

Second party verification was necessary to obtain a reliability index regarding coding. Therefore, a description of coder training is given.

The investigator first defined the categories for the coder. The coder was able to ask questions at this point so that nuances in category definitions could be clarified. In order that further misconceptions could be clarified the coder was first trained on practice units especially developed by the investigator for this purpose. The coder first sorted about 15 cards out loud into categories, and gave reasons for specific category choices. If the coder made an error during this process, the investigator stopped the exercise and clarified the categories. The coder sorted 10 more cards without interruptions. The investigator went though any corrections with the coder. Next the coder sorted 10 more cards, and at this point the coder was ready for the actual coding exercise.

The coder then proceeded to sort the 281 segments derived from all subjects into the seven categories. A coding reliability index was obtained by computing the percentage agreement between the two coders. The percentage agreement was found to be 93% - 261 units. The 20 units on which no agreement was found were dropped from further analysis. This was not a major problem in affecting any subsequent analysis since the units were few in number and the reliability was already quite high.

Summary of Instrumentation

1. DEMOGRAPHIC (BACKGROUND) QUESTIONNAIRE - Psychologists were asked to provide information relating to professional experience, level of education, and formal training in testing. This information was required in order to have descriptive data for the sample (see Appendix A for copy of questionnaire).

2. WISC-R MEASURE I consisted of four verbal subtests: Vocabulary, Comprehension, Similarities, and Information. An age was attached to the protocol. Each subtest contained some administrative errors which consisted of responses that were inappropriately cued and items administered beyond a ceiling level. The scoring criteria then required the psychologist to assign the correct point value to each response item. If an item was inappropriately queried (Q), the psychologist was expected to assign a point value according to the procedure outlined in the manual. For

example, if a one point response was given initially, and it was incorrectly probed on the protocol and a two point answer was subsequently recorded, it should still be assigned a one point. On the other hand, if a zero point answer was given initially, and an incorrectly queried response elevated the point value, the point value should still be recorded as zero. Other errors on the subtest included inaccurate starting points (see Appendix B for copy of WISC-R I measure). Each subject's protocol was checked against Slate's scoring key for point assignment differences per item. Each subject provided a total raw score which were checked for addition errors. There were no errors in addition. For each subtest the total raw score was converted into a scaled score. A prorated verbal IQ was also calculated for each subtest.

3. WISC-R MEASURE II consisted of a single Comprehension subtest from a different fabricated protocol. No age was attached to this subtest. As with the WISC-R I measure the complexity of the task involved the simultaneous consideration of whether an item was correctly administered or not, that is, a decision as to whether a response was correctly or incorrectly queried and response scoring judgment. The task also involved the judgment of ambiguous type responses, for example, "treat it (Q) treat it with things at home", and a multiple response, "catch bad people, arrest crooks, enforce laws" [item 4]. Lastly, the measure

also consisted of items administered beyond the ceiling (see Appendix A for copies of instrumentation).

As with the WISC-R I measure, the same scoring verification procedure was followed here except that a verbal IQ was not computed since this measure consisted of only one subtest. Additional data obtained from this measure involved the development of verbal categories from audiotape transcriptions.

The following chapter summarizes the demographic data for the sample and presents the results of the analysis of the WISC I and WISC II data.

CHAPTER 4

RESULTS

This chapter presents the research findings of the two sessions. There were two sources of data for this study. Twenty-two WISC-R verbal protocols were obtained from session one. One subject was not able to participate in the first session but was able to do so in the second session. These protocols comprised the Information, Similarities, Vocabulary, and Comprehension subtests, which provided scoring information descriptive of the sample. Secondly, eight verbal protocols were analyzed from a think-aloud exercise using a Comprehension subtest from another measure. This exercise provided general descriptive information of psychologists' cognitive strategies. The complete results of the demographic questionnaire is also presented in this chapter.

Demographic Characteristics of Sample

The demographic data pertaining to the 23 subjects are presented in Table 1. Separate demographic data are also presented for the 8 psychologists whose data was analyzed from the think-aloud session. The primary variables of interest included experience, educational level, and professional training.

The median number of years of experience indicated that

the sub-sample had a slightly higher degree of experience than the total sample however, all members of the second group had formal training on the WISC-R as opposed to 78% of the total sample. Another point is that about half (52%) of the total sample described their profession as varied, as opposed to 62% of the second sample who preferred the label of school psychologist.

Table 1

Demographic Characteristics of Sample

Characteristic	Total Sample (23)	Subsample (8)
Experience (years)		
Median	5.00	6.00
Range	1-27	1-12
Educational Level		
B.A.	1(4%)	0
M.A.	16(70%)	5(62%)
PhD/EdD	6(26%)	3(38%)
Formal Training		
WISC-R	18(78%)	8(100%)
WPPSI	18(78%)	7(88%)
Stanford-Binet IV	16(70%)	8(100%)
Profession		
School Psychologist	7(30%)	5(62%)
Educational Psychologist	2(13%)	2(25%)
Counselling Psychologist	0	0
Psychometrician	0	0
Special Educator	1(4%)	0
Eclectic	11(52%)	1(13%)
Developmental Psychologist	2(9%)	0

As some members of the total sample did not have formal classroom training on the WISC-R, a two-tailed t-test was conducted to compare the total number of errors made between

those subjects who had formal training and those who did not. The t-test revealed no significant differences between the two groups, $t(20)=.17$, $p>.1$. The nature of these errors are presented in the next section.

Session One WISC-R I Results

The number and types of errors were computed across subtests for the overall sample. The analyses revealed that the Vocabulary subtest was most prone to scoring errors. The Comprehension subtest was found to be the next highest in errors, then Similarities, and Information. Additionally, when psychologists erred, they were prone to give more credit than less credit. This result was obtained by comparing instances where psychologists awarded more credit when they should have given less (64 instances) to giving less credit when more credit was necessary (50 instances). Frequencies of types of errors are summarized in Table 2.

Next, in order to determine how great inter-psychologist differences were, scaled scores as well as a prorated Verbal IQ for each protocol were calculated. The Verbal IQ (PRO) was found to vary by as much as 11 points. The average Verbal IQ for the sample as well as the average scaled scores for each subtest is presented in Table 3. Slate's scoring key for the WISC-R I measure is also presented in Appendix B for comparative purposes.

Table 2

Types of Errors Across Subtests

Error type	Info	Sim	Voc	Comp*
0 point for a 2 point answer	0	1	6	4
0 point for a 1 point answer	2	0	7	7
1 point for a 2 point answer	0	0	15	7
1 point for a 0 point answer	10	6	2	3
2 points for a 1 point answer	0	9	12	4
2 points for a 0 point answer	0	8	2	1
Inappropriate questioning**	6	24	36	44
Failure to question***	10	6	15	14
Failure to obtain a correct ceiling	3	0	0	1
Failure to credit items below basal	13	NA	0	NA
Total (n=22)	44	54	95	85

* Comprehension scores based on 21 subjects
** Instances where the subject agreed with inappropriate questioning on the protocol, as well as introduced inappropriate additional questioning
***Failure to indicate on protocol additional (correct) questioning

As a second check of scoring variability, standard deviations [SD] were computed for each subtest and were compared with the Wechsler norms, that is, the standard error of measurement for each of the scaled scores. This comparison was also done for the Verbal IQ (PRO). A similar data analysis was previously used by Oakland, Lee, and Axelrad (1975). They suggested that the "SDs represent a range of scores reflecting the degree of interrater

Table 3

Comparison of Scaled Scores to Slate's Key

Subtest	Scaled Scores		Range	Scoring key
	Mean	SD		
Information	7.23	.92	6-9	7
Similarities	9.00	1.54	7-13	8
Vocabulary	7.05	.58	6-8	7
Comprehension	5.43	.87	4-7	6
Verbal IQ	82.52	3.88	80-91	81

variability. The higher the SD, the greater the variability; thus higher SDs reflect lower interrater consistency or reliability" (p.229). Therefore, if the standard deviation of each subtest is less than the standard error of measurement as reported by the manual then the scoring may be said to be relatively homogeneous and within the bounds of measurement error. Results indicated that except for the Similarities subtest and Verbal IQ, all standard deviations were substantial, although less than the SEMs as reported in the manual. The results of these analyses are summarized in Table 4.

Table 4

Means and Standard Deviations and Standard Errors of Measurement for Scale Scores and Verbal IQ

Subtest	Mean	SD	SEM
Information	7.23	.92	1.12
Similarities	9.00	1.54	1.28
Vocabulary	7.05	.58	.87
Comprehension	5.43	.87	1.51
Verbal IQ	82.52	3.88	3.57

Ancillary Analyses

When differences in assigning credit to specific items as compared to Slate's key was the sole consideration, differences averaged 6.45 items per protocol for the total sample. A breakdown of the sample revealed that those who participated in session one alone (n=15) had a mean of 6.27 point differences, while those who also participated in the second session (n=8) had slightly higher mean of 6.85 point differences (Table 5). A two-tailed t-test between these two means revealed no significant differences between the two samples, $t(20)=.35$, $p>.1$.

Table 5

Comparison of Point Differences Between Groups

Error of Item Credit Differences	Mean Errors Session One Alone	Mean Errors Both Sessions
6.45 (SD=3.65) (n=22)	6.27 (SD=3.92) (n=15)	6.85 (SD=3.24) (n=7)

Secondly, a Bartlett-Box F test was conducted to test for homogeneity of variance between point differences for all groups. The test revealed no significant difference between the group variances, $F(2)=.14$, $p>.1$.

A second analysis compared the mean number of errors with subjects who participated in session one alone to those who participated in both sessions. Those who also participated in the second session did not differ considerably from those who did not. Both groups made a similar number of errors when results were averaged across protocols. Table 6 summarizes these results. Furthermore, a two-tailed t-test found no significant differences between the two group means, $t(20)=.08$, $p>.1$.

Table 6

Comparison of Total Errors Between Groups

Mean Errors of Total group	Mean Errors Session 1 Alone	Mean Errors Both Sessions
12.64 (SD=7.74) (n=22)	12.73 (SD=8.76) (n=15)	12.43 (SD=5.53) (n=7)

Furthermore, a Bartlett-Box F test found no significant difference between the three group variances for total errors, $F(2)=.74$, $p>.1$. These results of the variance tests indicated that the groups were relatively homogeneous in respect to the overall errors made as well as differences made in assigning credit to items. However, the overall results also indicated that psychologists do exhibit differences in scoring as illustrated by the Verbal IQ range (Table 3). In this respect, in order to explain this variability within the sample, it was important to investigate psychologists' descriptions of their thought processes through verbal protocols.

Results of Session Two

This section presents the types of categories derived from the verbal data. These categories represent the common mental processes and strategies that underlie psychologists'

scoring judgments. Following the presentation of the development of category types, results of the analyses of category use within and across subjects are presented.

Types of Verbal Categories

1) Monitoring Statements (MS): This category either reflects differing states of comprehension or assessments of one's progress, that is, reflecting upon what one knows or does not know. The subject may verbalize statements referring to a lack of information in the response, an unclear response, or a response needing clarification that hinders this progress.

These statements may refer to instances of:

a) "self-questioning" behaviour.

Examples: - I wonder whether that should have been queried.
- I'm just not too sure what to do.
- I'm not sure.

b) "self-regulatory" behaviour: includes checking, revision, monitoring, or regulating one's progress.

Examples: - I'll just do a quick run through.
- I'm going back here.
- I'll just confirm that.
- I don't need that.

2) Planning (PL): This category refers to the subject's awareness of the task demands and goals that help make the problem easier to solve. Mechanisms to resolve the problem

include organizing ideas into goals for attacking the problem. That is, subjects may break up the problem into constituent parts. These are usually stated as intentions. Examples: - Okay, now I have to do two things.

- I remind myself that I'm looking for a couple of things here.
- Okay for two points you need both of those general ideas.

3) Self-explanation (SE): In this category the subject must go beyond the information given. This is because "the quantity and quality of a response is not sufficient to warrant a confident judgment" (Flavell, 1979). The subject overcomes the incompleteness of an example by deriving implications and/or making inferences by expanding on the information present. This subject may also explain the inadequacy of an answer by giving reasons. This means that more conceptual information is necessary if the subject is to properly evaluate a response.

Examples: - We can imply that "to go over myself" would be to get help.

- That really doesn't expand on the information.

4) General Metastrategic Statements (GMS): This category refers to any individualized or personal compensatory strategies or personal feelings that subjects fall back on in the evaluation process. These types of strategies may be

personalized strategies. Such strategies may include additional probing and/or proceeding beyond the ceiling.

Examples: - I always give more points.
- When in doubt I always give more.
- I never score as I administer the test, but even when I think I have a ceiling I still go on.
- I would like to know more about the kid.
- What a great answer!

5) Memory (MEM): The subject may first recall information from memory before checking the manual. That is, does the child's response match the subject's concept in memory? This is indicated by a fairly quick response indicating a concept already held in memory. For instance, the subject may automatically classify the response by giving it a point value before checking the manual.

Examples: - It's a fairly clear response, it's somewhere in the manual.
- It's in the guidebook; I'll check it though.
- It sounds/looks/appears like a one point answer.
- My initial reaction is that...

6) Manual (MAN): In this category the subject actually reads from the manual. The subject tries to find a corresponding match between the child's response and a concept in the

manual. This is done when the subject matches the response to the sample response in the manual or to the general criteria.

Examples: - I always consult the manual.
- That falls under the "Insulation" category.
- "Treat it" is in the manual.

7) Recommendations/Evaluations(R/E): This category refers to the domain-specific information that reflects the subject's interpretations about rules, procedures and scoring practices that bring about accurate scoring.

Examples: - This kid should have been questioned at the protocol.
- That didn't need to be cued.
- It was administered correctly.

Category Frequency of Verbal Categories Across Subjects

An analysis of category frequencies indicated that psychologists depended quite often on the WISC-R manual, although not overwhelmingly. This was represented by 27.6% of the reported verbalizations. In instances where psychologists actually consulted the manual but did not verbalize this behaviour, the investigator took note of this fact. So as not to misrepresent the verbal data, this were accommodated for in the actual coding.

Next to the use of the manual, psychologists reported the next greatest amount of time engaging in cognitive

activity that made reference to evaluations and recommendations. Comments were usually made about whether items that were appropriately queried. This category accounted for about 22.2% of the verbalizations.

Memory accounted for 14.2% of the verbalizations, and monitoring statements were reflected in 13.4% of verbalizations. For self-explanations, psychologists generally did not make their own interpretations of seemingly problematic responses before checking the manual. This was only done 9.9% of the time. Planning statements were minimal, they accounted for only 6.5% of the reported verbalizations. Similarly, metastrategic strategies (personal statements) were also underrepresented by the lowest statistic of 6.5 of the verbal categories. A percentage summary of the cognitive strategies is represented in Table 7.

Strategy Use Across Subjects

An index of the degree of similarity of strategy use was reflected in category frequency data across subjects. An inter-subject comparison across categories indicated that psychologists varied widely in the extent of particular strategies pertaining to the scoring task. For example, the highest percentage frequency of using the manual was 54% for one psychologist (Subject #8) as opposed to 18% for another psychologist (Subject #13). One psychologist monitored his/her work 23% of the time (Subject #2), another

Table 7

Frequency and Percentage Categories of Verbal Behaviours

Categories	Frequency (Percentage)
Manual	72 (27.9%)
Recommendations/Evaluations	58 (22.2%)
Memory	37 (14.2%)
Monitoring Statements	35 (13.4%)
Self-explanations	26 (9.9%)
Planning	17 (6.5%)
General Metastrategic Statements	17 (6.5%)

only 4% of the time (Subject #8), and one psychologist made no references at all to general metastrategic statements (Subject #8) while all the others did make some metastrategic references. Percentage-wise, Subject #15 accounted for most of the metastrategic statements such as "But let's just try to find out more about the kid" or "Sometimes I'm able to put thing in context if I know how old he is". Interestingly, one psychologist never verbalized planning statements (Subject #13) and another only 3% of the time (Subject #11). These results are summarized in Table 8.

Table 8

Frequencies and Percentages* of Cognitive Strategies Across Subjects in each Category

	Subjects							
	2	8	10	11	12	13	15	18
MAN	5(23)	15(54)	11(31)	11(31)	6(21)	3(18)	12(19)	9(29)
R/E	5(23)	3(11)	5(14)	9(25)	8(29)	8(47)	14(22)	6(19)
MEM	2(9)	1(4)	6(17)	8(22)	4(14)	1(6)	10(16)	5(16)
MS	5(23)	1(4)	7(20)	3(8)	4(14)	1(6)	9(14)	5(16)
SE	1(4)	5(18)	2(6)	4(11)	3(11)	3(18)	5(8)	3(10)
PL	3(14)	3(11)	2(6)	1(3)	2(7)	0	4(6)	2(6)
GMS	1(4)	0	2(6)	1(3)	1(4)	1(6)	10(16)	1(3)
Total Number	22	28	35	36	28	17	64	31

*percentages are presented within parentheses

Scoring by Item on the WISC-R II Measure

Psychologists' scoring on the Comprehension WISC-R II measure is presented in Table 9. The total raw score for each subject is presented as well as the total scoring difference from Slate's key. This difference is also represented as a distribution reflecting the degree of lenience in scoring. For example, Subject #2's total raw score was greater by 3 points comparative to Slate's key (Total scoring diff). The source of scoring differences is represented by a scoring distribution (Score dist) where this subject gave more points to 5 items and less points to

Table 9
Patterns of Scoring on WISC-R II Measure

Items	Key	Subjects 2	8	10	11	12	13	15	18
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
3	1	2	1	1	1	1	1	1	1
4	1	1	2	2	1	1	1	1	2
5	2	2	2	2	2	2	2	2	2
6	2	0	2	0	0	0	2	2	2
7	2	2	2	0	2	2	2	2	2
8	1	1	1	1	1	0	1	1	1
9	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0
11	0	1	1	0	0	0	0	1	1
12	0	0	0	0	0	0	0	0	0
13		1	0					0	0
14		1	0					1	0
15		1	1					1	1
Raw Score Total	2	15	14	9	10	9	12	15	15
Total Scoring Diff		+3	+1	-3	-2	-3	0	+3	+3
Score Dist		+5 -2 (7)	+3 5 (3)	+1 -4 (5)	-2 (2)	-3 (3)	0	+3 (3)	+3 (3)

2 items, with a total of 7 items scored differently than the key. Overall, only one subject scored in agreement with the key. Four of the others awarded more points than the key, the other three psychologists awarded fewer points. Additionally, four of the psychologists awarded points past the ceiling (Subjects 2, 8, 15, and 18); four others obtained the correct ceiling (Subjects 10, 11, 12,and 13).

Non-Problematic Items

Items where psychologists scored consensually were 1, 2, 5, 9, 10, and 12. All psychologists awarded 1 point for the response to item 1, 2 points for the responses to items 2 and 5, and 0 points for the responses from items 9, 10, and 12. As for items 3, 7, and 8, the scores differed by one point assignment.

Difficult Items

Items 4, 6, and 11 posed more difficult to score. Three psychologists awarded 2 points for the response to item 4, while the other five gave 1 point to the same response. As for item 6, half of the psychologists gave no credit to the response, while the other half gave 1 point. The scoring on item 11 was also divided. Half of the psychologists gave 1 point while the other half awarded no points. For illustrative purposes the number and percentage of verbalized responses across subjects for three of the non-problematic items (1, 2, and 9) and three of the difficult items (4, 6, and 11) are presented in Appendix D

in Tables D1 and D2. Strategies varied across items, however, memory seemed to be more frequent for the less difficult items, while recommendations/evaluations was more common for the more difficult items. For examples of psychologists' actual verbalizations to these items see Appendix C. Appendix C presented the actual verbalized statements of the psychologist who had no errors (Subject #13) and the psychologist who erred the most (Subject #2).

General Strategies

At the end of the think-aloud session, psychologists were asked to relate any general strategies that were helpful to them in scoring difficulties. Not surprisingly, all psychologists said that their general strategy was to refer to the manual. However, the percentage frequencies of only four of the subjects (Subjects 8, 10, 11, and 18) indicated the use of the manual was a primary decision aid.

Summary

The results of both sessions showed that, on average, psychologists participating in this study were within an acceptable range of error (Table 3). However, there were some individual scoring differences as represented by the 11 point spread for the Verbal IQ. To explore this variance, inter-individual scoring patterns were represented in the second session (Table 9). Scoring strategies were previously identified in the second session (Table 7) as well as the frequency of strategy use across subjects (Table

8). These results are discussed in the next chapter.

Chapter 5

SUMMARY AND CONCLUSIONS

The primary purpose of this study was to explore psychologists' use of cognitive strategies when scoring difficult verbal responses on the WISC-R. Because the psychometric approach does not address mental processes, it was helpful to look at how these processes might affect psychologists' judgments, and ultimately their scoring. This study investigated the processes behind psychologists' decision-making behaviour as it pertained to judging verbal responses.

Summary of Results and Discussion

Session one and session two psychologists scored similarly on the WISC-R I measure. Therefore, with respect to their performance, one may infer that the results on the WISC-R II measure have some degree of generalizability to those psychologists who only participated in the first session. T-tests found no significant differences for point differences and errors between session one only and session two psychologists. Although the Verbal IQ was found to vary by 11 points (80-91), overall group differences in scoring were not sufficient to affect subtest scaled scores or the Verbal IQ (PRO) score. Collectively, the psychologists in this study were found to assign scores comparable to those

on Slate's fabricated protocol. Additionally, the SD of the Verbal IQ approached the SEM value of each subtest suggests that scoring variability is a an error component that should be incorporated into the SEM (Table 4). This suggests in turn, that although a high degree of objectivity is assumed in the administration and scoring of WISC-R verbal responses, test users need to be cognizant of the limitations inherent in the tests themselves. Limitations in the test are only acknowledged in the form of content sampling (estimated from studies of internal consistency), and time sampling (estimated from score stability studies) as sources of measurement error (Hanna, Bradley, & Holen, 1981; Slate & Chick, 1989).

Although the net scoring differences did not appear to affect the overall IQ there was apparent inter-subject variability. Scores often cancel themselves out within a protocol and items that reflect differences of opinion are not identified. In other words, the range of scores on these verbal subtests reflected psychologists' differences in attaching credit to specific items.

Lastly, in order of scoring difficulty, the Vocabulary, Comprehension, Similarities, and Information subtests were found to be prone to scoring errors. These results are consistent with those obtained by others (Miller, Chansky, & Grendler, 1970; Oakland, Axelrad, & Lee, 1975; Slate & Chick, 1989; Slate & Jones, 1988). Next, the results of

session two will be discussed in more detail since this session was the primary focus of the study. The Comprehension subtest used in the think-aloud exercise was a test that the psychologists had not used in the first session. This test was employed so as not to repeat previous material.

Discussion of Research Question 1: Are there common mental processes and strategies that underlie psychologists' scoring judgments?

The results of the second session indicated that psychologists generally relied on the manual; some psychologists referred to the manual more than others. Psychologists referred to the manual's principles and criteria as an aid in judging responses, although at times this strategy was not productive.

Recommendations/evaluations, the next most frequently used category, reflected psychologists' knowledge of task-specific rules, for example, "I think it should have been queried" (Subject #18). Furthermore, some responses seemed to remind psychologists of an example already held in memory. After reading a response, one psychologist simply responded, "That looks like a hit on a one point response (Subject #15). Thus, psychologists seemed to have a repertoire of responses where they were able to learn from examples and make abstractions across categories. Thirdly, because psychologists did not engage frequently in

self-questioning and self-regulatory behaviour, one may speculate that performance in this regard may be highly automatic. Additionally, the lack of planning statements indicated that again steps in the task were highly procedularized. Psychologists generally did not make their own interpretations of seemingly problematic responses before checking the manual. When interpretations were made it was to generate more conceptual information in order to more clearly evaluate a response. "We can imply to go over myself would be to get help" (Subject #12). Psychologists generally found it helpful to verbalize planning statements when the task became difficult. This was a key heuristic when psychologists had to search for at least two key ideas in a response. Moreover, general metastrategic statements were found not to account for a large portion of judgments pertaining to scoring behaviour. However, the low percentage of metastrategic statements does not adequately reflect this strategy that can lead to errors by affecting ceilings. This point will be discussed further.

Discussion of Research Question 2: To what degree are psychologist cognitive strategies similar?

Individual cognitive strategies were found to vary depending on the complexity of the response. Because intra-individual processes were diverse across the group, two of the psychologists' cognitive patterns may be highlighted in particular: the cognitive patterns of the

psychologist who had no scoring errors and the cognitive activity of the psychologist who erred the most.

Psychologist With No Errors (Subject #13)

Ironically, the psychologist that referred to the manual the least amount of time (18%) made no errors. Rather, this psychologist spent most of the time verbalizing statements referring to recommendations and evaluations (47%). For example, verbalizations included such statements as, "that kind of response is not meant to be queried" or, "okay on the first one I would give it a one, it was appropriately queried". The manual was consulted only for clarification. The underlying reason was that this psychologist was also more prone to notice errors within the protocol itself and marked accordingly. This was also indicated by the relatively high rate of evaluation/recommendation responses which was an index of task knowledge. Additionally, in ambiguous circumstances, such as in item 1, "treat it (Q) treat it with things at home", scoring accuracy was increased because the psychologist also tried to make judgments based on self-reasoning when the manual was of no help. For example, "the additional answer doesn't add anything that I feel tells me that the child knows any more than what he knew in the first answer." This is evidenced in this psychologist acquiring the highest frequency of self-explanation episodes. In sum, 65% of the psychologist's verbalizations referred to self-explanations

and recommendations/evaluations. Additionally, since this psychologist took the least amount of time to score the subtest, and did so with no errors, this indicated a high level of proceduralized knowledge.

Psychologist with the Most Errors (Subject #2)

Although Subject #2 used the manual more frequently (23%) than Subject #13, this psychologist erred the most overall. The differentiating factor was that although this psychologist used the manual more often, s/he made half as many recommendations and evaluations (23%). This appears to reflect a lack of declarative knowledge necessary for accurate scoring, that is, the questioning rules for scoring. For example, for the response, "so people can get meat (Q) it might be bad", this person failed to disregard the incorrect cue and awarded the latter part of the response one point. This action impeded obtaining the correct ceiling. Additionally, this psychologist engaged in the least amount of self-explanatory behaviour (4%) when presented with ambiguous responses. In contrast, for Subject #13 self-explanatory behaviour seemed helpful in the differentiation between category responses or vague responses.

Psychologists' Scoring by Item on the WISC-R II

As expected, responses clearly scorable by the manual were awarded a consensual point value by all psychologists. An example of such an instance is in question two where

there were no errors, "What are you supposed to do if you find someone's wallet or pocket-book in a store?". The response given in the fabricated protocol was "Turn it into lost and found". In these types of situations psychologists would usually rely on memory-based examples, that is, instances from previous experience. In other words, it would appear that the structural aspects of the response reminded psychologists of examples already held in longterm memory. Examples of memory-based comments to this question were as follows:

> It's a fairly clear response, I'll give it a two -it's somewhere in the manual it's as close as such (Subject #10).
>
> The second one is straight forward. It's the same as in the book (Subject #13).
>
> Turn it into lost and found is two points. So that's a two (Subject #2).

On the other hand, after a memory-based impression was made, some psychologists preferred to use an additional monitoring strategy by also checking the criteria in the manual:

> Number two. Turn it into lost and found. My first reaction is I think it's a two pointer (memory)...and when I check it actually is (manual; Subject #18).

Thus, in some cases the manual was never consulted; in other cases, the manual functioned primarily as a "backup" aid if psychologists were confident of a response that resembled a memory-based example.

Problematic Responses

For responses that were more problematic to score, psychologists' judgments may be explained by their personal frames of references, the way they conceptualized responses as well as by their rule-based interpretations. The difficulty of scoring "novel" subject responses has been widely acknowledged. Sattler (1988) amplifies the challenge posed in the scoring of verbal responses as illustrated in Figure 3. Some of the items in the Comprehension verbal exercises seemed problematic in this regard since there was variance in scoring among subjects. An example of these items to be discussed are 4, 6, and 11.

Personal Frames of Reference

Psychologists' personal frames of references may be accounted for by statements referring to their general metastrategic statements (personalized strategies). Although such statements were few in number (5.4% of the overall verbalizations), they were still found to exert an influence on psychologists' scoring. This was represented by the fact that half of the subjects obtained ceilings by the 12th response, while the others did not obtain ceilings. Obtaining the correct ceiling is an important aspect of accurate scoring since subsequent scoring is determined by the correct finishing point. Again, this was evidenced in item 11. Those who did not obtain a ceiling assigned a one point credit to the 11th response to the question "Why is it

important for the government to hire people to inspect the meat in meat packing plants?" The fabricated response was "So people can get meat(Q)It might be bad". In these instances psychologists ignored the query and preferred to award the child a point, as in the following examples: "He [the child] demonstrated that he knew it" (Subject #15), "So, it's to the child's advantage if you make this kind of mistake because you can then raise the score" (Subject #2), and "I think it's verging on the concept of protecting the consumer" (Subject #18). Generally, the psychologists were interested in whether the child really knew the correct response despite the knowledge that the response should not have been queried. A more general account of one psychologist's personal philosophy is summed up in the following by subject #1:

> The administration sets up a "structured interview" in order to get to know the child. I look at the WISC-R as a tool to get to know the child. I test a lot of ESL children. You know that they are intelligent but they may not know the words, so you try to get the global response. I also test a lot of language pathology children. You have to get the global response of what they mean because they may not be able to express it (Subject #1).

Subject #2 expressed similar thoughts. "I write all over my protocol. I tend to make observations because the numbers themselves don't have any meaning on their own". For instance, this psychologist said:

> The questions in the Comprehension subtest basically are morally loaded questions. What a child learns at home

> is basically his/her view of the world. The questions in the WISC-R may not be reflective of that. For the question, "What would you do if someone younger than you hits you?" Well, I knew a child whose younger brother was always hit upon. When the family system is set up this way it's hard to ignore. So if you pose such a question to the child, the natural response would be to hit him.

Thus, it appears that psychologists' personal frames of references affect their individual judgment strategies as well. For example, psychologists appear to be lenient if the child's response shows promise or if it seems that the child may know the correct answer depending on the context or situation. One psychologist suggested that, "If this is an older person I think I would give that two points; if it was a younger person I would consider it to be queried...to see if they know what they are talking about" (Subject #18). In such circumstances, one psychologist (Subject #15) said it is better to always give more points. On the other hand, another psychologist related a different approach:

> I don't know exactly if I come to a situation where I really don't know...I mean I know some people tend to score up and some people tend to score down and I guess if I were stuck in that situation I would tend to score down...and then I would just keep that in mind how many times I had to do that when I'm coming up with the scores to see if they would have an effect on any of the scales or the general scores (Subject #8).

A Multitude of Responses: Differences in Conceptualizations of Responses

Another item where psychologists did not achieve a consensus was on item four. Where psychologists

conceptualized responses differently was reflected primarily in self-questioning behaviour and monitoring behaviour. Self-questioning served an important function by offering clarifications and giving reasons in instances of problematic responses. Monitoring behaviour was reflective of psychologists' checking their actions in the face of such unclear responses. For the question, "What are some reasons we need police men?" The fabricated response was, "Catch bad people, arrest crooks, enforce laws". Five of the subjects categorized the three responses as similar and awarded the response one point as a unitary response. Although psychologists checked the manual for matching criteria it did not seem to be overly helpful to all. For example, Subject #13 said, "I would have queried that as well to look for a second answer, they all fall within the same category." Subjects 8, 10, and 18 awarded two points. They gave reasons such as "I think that categorically catching bad people and arresting crooks are probably one, and enforcing laws is definitely another" (Subject #18). "The first two are the same level...I'm checking to see if the two ones are the same as 'enforce laws'... I think they're two separate categories so I'm giving it two points" (Subject #10).

To (Q) Or Not To (Q):Rule-based Interpretations

When psychologists made reference to rule-based interpretations (22.2% of the time), it was generally in

reference to examiner errors already present on the protocol. This aspect of scoring was reflected in the recommendations/evaluations category, which revolved mainly around whether a response should or should not be cued. An important finding emerged involving biased heuristics. This heuristic involved always scoring the response after a (Q). Half of the subjects used this biased heuristic for question #6, "What is the thing to do if a boy (girl) much smaller than yourself starts to fight with you?". For the response, "Let him be(Q)and get mad", four subjects gave 0 points and the other four 2 points. Psychologists gave reasons such as, "...I think that he would get a zero when he spoils it like this" (Subject #2), as opposed to "I'd give that a two because his first answer is correct and it should not have been queried, and while his second answer is inadequate it should not have been asked in the first place so his first answer is worth a two" (Subject #13) - the correct algorithm. An application of an incorrect heuristic would involve always crediting a second response to an incorrectly cued first response. To reiterate, the correct algorithm would be to award credit only to the first response in accordance with standardization procedures irrespective of an answer that elevates the credit value. Also, differences in rule-based interpretations and thus judgments in scoring were apparent in the scoring of item eleven, "So people can get meat (Q) It might be bad". Half of the psychologists

awarded no credit, the other half awarded 1 point credit. Again, the biased heuristic involved crediting the latter response if it elevated the value of the first response despite inappropriate cueing. As discussed previously, for this example a personal frame of reference that favours leniency may explain this biased heuristic of awarding credit to the more sophisticated response because some psychologists did note that the response should not have been cued. According to Brehmer, Hagafors, & Johansson (1980):

> Subjects do not perform optimally in a judgment task even when they know exactly what rule to use for making optimal judgments ... knowing the rule for a task is not enough for producing a correct response...being told how to perform a judgment task does not guarantee perfect judgments. (p. 373)

Therefore, being told how to score responses does not guarantee correct scoring judgments. Alternatively, general knowledge deficiencies pertaining to scoring rules is an explanatory factor underlying differences in judgments (Subject #2).

Discussion of Research Question 3: To what degree are psychologists aware of their own cognitive processes and strategies?

Results of percentage frequencies indicated that only half of the subjects accurately reported the manual as their primary heuristic. It could be that although psychologists are aware that the manual is the most important heuristic

aid for checking responses, they are unaware that there are other underlying factors, or alternative primary strategies influence their scoring judgments just as well. Especially in instances where the manual is not overly helpful psychologists may unknowingly have developed secondary strategies to aid in their scoring judgments. This is evidenced by the fact that psychologists did not report extensively on individual metastrategic statements (5.4%) as pressure to conform to standardization procedures is mandatory in the scoring of WISC-R responses.

One subject did have a better idea than the others that the manual was the most important decisive factor in scoring judgments and commented upon this:

> Well, I guess the first thing I would do is to try and stick to the manual. So, if I can see that there is something clearly in the manual like in the example where responses weren't queried even though they should have been then I don't because it's been drilled into me that you have to stick to the standardization because otherwise it's not as valid and reliable.

However, this psychologist differed from Slate's key by 4 points, the second highest error rate. One may then conclude that use of the manual does not itself guarantee flawless scoring. As seen, there seems to be important cognitive variables such as the strategies identified in this study that psychologists are largely unaware of in their scoring experience.

Conclusions

An advantage of a study of this nature is that it identified general cognitive strategies and a biased heuristic that conceptualized sources of error. This was done by relating cognitive strategies to scoring differences. Secondly, through the validity question (research question 3), it brought to light that psychologists are largely unaware of how their underlying cognitive variables affect judgments. Thirdly, a biased heuristic of some psychologists was identified that involved identifying an inclination towards leniency even if it meant that standardized procedures were not adhered to. One may speculate, then, that other members of the general population may use similar processes in their scoring, especially for difficult or ambiguous verbal responses. This leniency aspect of scoring also became apparent in the quantitative performance data in session one where subjects were more prone to award more credit to responses of lesser value (Table 2).

Another point is that in the present study psychologists generally found the think-aloud exercise helpful to them in looking at their underlying reasons in difficult scoring situations. In reference to an inappropriately questioned item on a protocol, one psychologist commented that, "What do I do there...this is good review for me...this is actually an interesting exercise... to do, because it is

interesting to look at why do I think that..." (Subject #2).

Limitations of the study

Although the majority of educational research is conducted with volunteer subjects (Borg & Gall, 1989), the fact that a volunteer sample was used in the study evokes several considerations. The first is that the psychologists who volunteered may have been different than those who did not. Those who volunteered were perhaps more confident in their scoring, and perhaps were more at ease to opening themselves up to a stranger for evaluation. Others who did not volunteer perhaps declined to do so due to the anxiety of having their work evaluated. Additionally, those who volunteered for the thinking aloud exercise may be a different group in themselves. Although think-aloud exercises generally involve a small number of Subjects (Ericsson & Simon, 1984), participation of all 23 subjects was not obtained for the verbal analysis. Caution should therefore be maintained in generalizing the results of this study to the general population of Greater Vancouver psychologists. Ideally, if a larger sample of scorers were obtained, then perhaps more general kinds of heuristics and scoring biases would have become apparent, as reflected in different category types and frequencies.

A second limitation of the study is that the protocols were fabricated. In a real testing situation psychologists often find alternative cues helpful in their scoring. For

example, Subject #12 commented that "in looking at these answers, it's like dealing with the kid, and the strategies that the kid goes through to get these answers are helpful, so that's what you miss when you're looking just at black and white". An alternative methodology to overcome this disadvantage would be to videotape an assessment process involving the administration of the WISC III to a child. A videotape would capture verbatim responses as well as make visible important testing facets that would be lost through fabricated protocols. A further research question may ask, "To what extent would scoring judgments obtained from a more realistic setting differ from those obtained on a fabricated protocol constructed with the same verbal responses?" For instance, for a multiple response answer in this study, one psychologist said that s/he may have marked differently if the child had taken a breath in between a response - indicating two concepts rather than one. Such cues would be apparent on videotape.

Another limitation related to the analogue nature of the protocols is that psychologists were already presented with cued responses. These facets may have actually functioned as distracters in this exercise.

Implications of the Study

Differences among individuals who are in a judgment role have usually been treated as error (Slovic & Lichtenstein, cited in Rappoport & Summers, 1973) irrespective of the

underlying mental processes. This is because researchers have paid minimal attention to the inferences and the particular cognitive processes that are responsible for examiners' performance judgments (Kishor, 1987). Consequently, it is often overlooked that examiners are an active part of a test process that calls for specific cognitive skills surrounding interpretive judgment.

Unfortunately, it does not appear as though studies have systematically explored the systems underlying psychologist-examiners' scoring behaviours. According to Barnett (1988, cited in Burns, 1990), applications of judgment research in cognitive psychology to school psychology are lacking. Rather the end product of the scoring process, that is, the behavioural or scoring phase has received the attention rather than the systems underlying this process. The study of these systems may help in the understanding of psychologists' differential judgments of difficult-to-score and ambiguous verbal responses. Slate and Hunnicut (1988) analyzed literature pertaining to the Wechsler protocol errors in order to identify probable reasons. They suggested that ambiguity in test manuals and poor instruction were contributing factors; however, again, there empirical studies addressing these specific issues are lacking. One may speculate, as suggested in this study, that, in practice, psychologists are often left to their own devices in terms of scoring

problematic verbal responses. It follows that studies need to focus on the cognitive activity surrounding psychologists' actions in doubtful situations. Do they often rely on their own compensatory strategies when evaluating difficult responses? Psychologist-examiners must often make subjective judgments because many responses given by children and adults alike are not clearly scorable by the test manual (Slate & Hunnicut, 1988).

Furthermore, this study brings to light that variations in psychologists' cognitive activity appear to affect scoring outcomes. This was demonstrated by comparing the performance of Subject #8 and Subject #13. Furthermore, implications for future training is that in order to finetune scoring, the need for revision of questioning rules by psychologists, as well as the clarification of such rules in the manual may be helpful for some psychologists. Additionally, these problematic areas should be stressed in the teaching of the Wechsler Scales in the graduate classroom.

Finally, the similar task structure of the new versions of the WISC-R, WISC III, implies that the findings of this study will have some degree of generalizability to the new measure. This is because the nature of underlying cognitive processes were studied, rather than an item analysis. Therefore, the decision-making strategies and techniques that psychologists employ in scoring may also transfer to

problem areas on the WISC III as well. However, since this was primarily an exploratory study, the possibility remains for more indepth research using the similar WISC III task.

References

Anderson, B. (1977). Differences in teachers' judgment policies for varying numbers of verbal and numerical cues. Organizational Behavior and Human Performance, 19, 68-88.

Babad, E., Mann. M., & Mar-Hayim, M. (1974). Bias in scoring the WISC subtests. Journal of Consulting and Clinical Psychology, 43, 268.

Barnett, D.W. (1988). Professional judgment: A critical appraisal. School Psychology Review, 17, 658-672.

Benjafield, J. (1969). Evidence that "thinking aloud" constitutes an externalization of inner speech. Psychonomic Science, 15, 83-84.

Berl, J., Lewis, G., & Morrison, R.S. (1976). Applying models of choice to the problem college selection. In E.R. Smith & Miller, F.D. (1978). Limits on Perception of Cognitive Processes: A Reply to Nisbett and Wilson. Psychological Review, 85, 355-362.

Boehm, A., Duker, J., Haesloop, M., & White, M. (1974). Behavioral objectives in training for competence in the administration of individual intelligence tests. Journal of School Psychology, 12, 150-157.

Borg, W., & Gall, M. (1989). Educational Research: An Introduction, Fifth Edition. New York: Pitman.

Borman, W.C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 20, 238-252.

Bradley, F., Hanna, G., & Lucas, M. (1980). The reliability of scoring the WISC-R. Journal of Consulting and Clinical Psychology, 48, 530-531.

Brannigan, G. (1975). Scoring difficulties on the Wechsler intelligence scales. Psychology in the Schools, 12, 313-314.

Brehmer, B., Hagafors, R., & Johansson, R. (1980). Cognitive skills in judgment: Subjects' ability to use information about weights, function forms, and organizing principles. Organizational Behavior and Human Performance, 26, 373-385.

Brown, J.S., & Burton, R. (1978). Diagnostic models for procedural bugs in basis mathematical skills. Cognitive Science, 2, 155-192.

Burgess, R. (1985). Field Methods in the Study of Education. Philadelphia: Falmer.

Burns, C.W. (1990). Judgment theory and school psychology. Journal of School Psychology, 28, 343-349.

Cardy, R., & Dobbins, G. (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. Journal of Applied Psychology, 72, 672-678.

Chase, W.G., & Simon, H.A. (1973). Perception in chess. Cognitive Psychology, 4, 55-81.

Chi, M.T.H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Conner, R., & Woodall, F.E. (1983). The effects of experience and structured feedback on the WISC-R error rates made by student-examiners. Psychology in the Schools, 20, 376-370.

Crow, L., Olshavsky, R., & Summers, J. (1980). Industrial buyers' choice strategies: A protocol analysis. Journal of Marketing Research, 17, 34-44.

Dansereau, D., & Gregg, L. (1966). An information processing analysis of mental multiplication. Psychonomic Science, 6, 71-72.

Davis, H. (1968). Verbalization, experimenter presence and problem solving. Journal of Personality and Social Psychology, 8, 299-302.

de Groot, A.D. (1965). Thought and Choice in Chess. The Hague: Mouton.

DeNisi, A.S., Cafferty, T.P., & Meglino, B.M. (1984). Organizational Behavior and Human Performance, 33, 360-396.

Ericsson, K.A., & Simon, H.A. (1984). Protocol Analysis: Verbal Reports as Data. Cambridge: The MIT Press.

Ericsson, K.A., & Simon, H.A. (1980). Verbal Reports as Data. Psychological Review, 87, 215-251.

Fagley, N.S. (1988). Judgmental heuristics: Implications for the decision making of school psychologists. School Psychology Review, 17, 311-321.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 66, 127-148.

Flavell, J.H. (1979). Metacognition and cognitive monitoring: A new area of cognitive developmental enquiry. American Psychologist, 34, 906-911.

Franklin, M., Stollman, P., Burpeau, M., & Sabers, D. (1982). Examiner error in intelligence testing: Are you a source? Psychology in the Schools, 19, 563-569.

Fogarty, J.L., Wang, M.C., & Creek, R. (1983). A descriptive study of experienced and novice teachers' interactive thoughts instructional thoughts and actions. Journal of Educational Research, 77, 22-32.

Forrest-Pressley, D.L., MacKinnon, & Waller, T.G. (1985). Metacognition, Cognition, and Human Performance: Instructional Practices (Vol.2). New York: Academic Press.

Gagne, E.D. (1985). The Cognitive Psychology of School Learning. Boston: Little, Brown and Company.

Gagne, M., & Smith, E., Jr. (1962). A study of the effect of verbalization on problem solving. Journal of Experimental Psychology, 63, 12-18.

Galotti, K. (1989). Approaches to studying formal and everyday reasoning. Psychological Bulletin, 105, 331-351.

Gavelek, J.R., & Raphael, T.E. (1985). Metacognition, instruction, and the role of questioning activities. In D.L. Forrest-Pressley, G.E. MacKinnon, & T.G. Waller (Eds.), Metacognition, Cognition, and Human Performance: Instructional Practices (Vol.2). New York: Academic Press.

Glaser, B.G., & Strauss, A.L. (1967). The Discovery of Grounded Theory: Strategies for Qualitative Research. Chicago, IL: Aldine.

Glaser, B.G. (1978). Theoretical Sensitivity. Mill Valley, CA: The Sociology Press.

Hanna, G., Bradley, F., & Holen, M. (1981). Estimating major sources of measurement error in individual intelligence scales: Taking our heads out of the sand. Journal of School Psychology, 19(4), 370-376.

Higgins, E., Herman, C., & Zanna, M. (1981). Social Cognition: The Ontario Symposium (Vol. 1). Hillsdale,NJ: Erlbaum.

Hunnicutt, L., Slate, J., Gamble, C., & Wheeler, M. (1990). Examiner errors in the Kaufman Assessment Battery for Children: A preliminary investigation. Journal of School Psychology, 28, 271-278.

Kasper, J.C., Throne, F.M., & Shulman. J.L. (1968). A study of the interjudge reliability in scoring the responses of a group of mentally retarded boys to three WISC subscales. Educational and Psychological Measurement, 28, 469-477.

Karph, D.A. (1973). Thinking-aloud in Human Discrimination Learning. Unpublished doctoral dissertation, State University of New York at Stony Brook.

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

Kilpatrick, J. (1968). Analyzing the solution of word problems in mathematics: An exploratory study (Doctoral dissertation, Stanford University, 1967). Dissertation Abstracts International, 42, 4380-A.

Kishor, N. (1987). Cognitive Strategies in Judgment. Unpublished doctoral dissertation, The University of British Columbia, Vancouver.

Klein, N. (1983). Utility and decision strategies: A second look at the rational decision maker. Organizational Behavior and Human Performance, 31, 1-25.

Krzytofiak, F., Cardy, R., & Newman, J. (1988). Implicit personality and performance appraisal: The influence of trait inferences on evaluations of behavior. Journal of Applied Psychology, 73, 515-521.

Lesgold, A. (1990). Problem solving. In R.J. Sternberg &, E.E. Smith (Eds.). The Psychology of Human Thought (pp. 188-213). Cambridge:Cambridge University Press.

McArthur, L. (1981). What grabs you? The role of attention in impression formation and causal attribution. In E. Higgins, C. Herman, and M. Zanna (Eds.), Social Cognition: The Ontario Symposium (Vol.1). Hillsdale, NJ: Erlbaum.

McCormick, C., Miller, G., & Pressley, M. (1989). Cognitive Strategy Research. Springer-Verlag: New YorK.

Miller, C., & Chansky, N. (1972). Psychologists' scoring of WISC protocols. Psychology in the Schools, 9, 144-152.

Miller, C., Chansky, N., & Gredler, G. (1970). Rater agreement on WISC protocols. Psychology in the Schools, 7, 190-193.

Montgomery, H. (1977). A study of intransitive preferences using a think aloud procedure. In H. Zungermann &, G. De Zeeuw (Eds.), Decision Making and Change in Human Affairs. Boston: D. Reidel.

Mount, M.K., & Thompson, D.T. (1987). Cognitive categorization and quality of performance ratings. Journal of Applied Psychology, 72, 240-246.

Murphy, K.R., & Balzer, W.K. (1986). Systemic distortions in memory in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. Journal of Applied Psychology, 71, 39-44.

Newell, A., & Simon, H.A. (1972). Human Problem Solving. New Jersey: Prentice-Hall.

Nisbett, R., & Wilson, D. (1977). Verbal reports on mental processes. Psychological Review, 84, 231-259.

Oakland, T., Lee, S., & Axelrad, K. (1975). Examiner differences on actual WISC protocols. Journal of School Psychology, 13, 227-233.

Payne, J. (1980). Information processing theory: Some concepts applied to decision research. In T.S. Wallsten (Ed.), Cognitive Processes in Choice and Decision Behavior (pp. 95-115). Hillsdale, NJ: Erlbaum.

Peterson, P.L., Marx, R.W., & Clark, C.M. (1978). Teacher planning, teacher behavior, and student achievement. American Educational Research Journal, 15, 417-432.

Pitz, G., & Sachs, N. (1984). Judgment and decision: Theory and application. Annual Review of Psychology, 35, 139-163.

Plumb, G., & Charles, D. (1955). Scoring difficulty of Wechsler comprehension responses. Journal of Educational Psychology, 46, 179-183.

Rappoport, L., & Summers, D. (1973). Human Judgement and Social Interaction. New York: Holt, Rhinehart, & Winston.

Sattler, J.M. (1988). Assessment of Children (3rd ed.). San Diego: Jerome M. Sattler.

Searles, E. (1985). How to Use WISC-R Scores in Reading Diagnosis. Newark, DE: International Reading Association.

Sherrets, G., Gard, G., & Langer, H. (1979). Frequency of clerical errors on WISC protocols. Psychology in the Schools, 16, 495-496.

Schulman, L.S., & Elstein, A.S. (1975). Studies of problem solving, judgment, and decision making: Implications for educational research. In F. Kerlinger (Ed.), Review of Research in Education (Vol.3, pp. 3-42). Itasca, Illinois: F.E. Peacock.

Simon, H.A., & Chase, W.G. (1973). Skill in chess. American Scientist, 61, 394-403.

Simon, D.P., & Simon, H.A. (1978). Individual differences in solving physics problems. In R. Smith, E.T. (1985, September 30). Are you creative? Business Week, 80-84.

Slate, J.R. (1991). Guide to administering and scoring the WISC-R. Unpublished manuscript, Arkansas State University, Arkansas.

Slate, J.R. & Chick, D. (1989). WISC-R examiner errors: Cause for concern. Psychology in the Schools, 26, 78-83.

Slate, J.R., & Hunnicutt, L. (1988). Examiner errors on the Wechsler scales. Journal of Psychoeducational Assessment, 6, 280-288.

Slate, J.R., & Jones, C.H. (In press). Errors on the Wechsler scales: Commonly mis-scored examinee responses. Social and Behavioral Sciences Documents.

Slate, J.R., & Jones, C.H. (1990). Identifying students' errors in administering the WAIS-R. Psychology in the Schools, 27, 83-87.

Slate, J.R., & Jones, C.H. (1989). Examiner errors in the WAIS-R: A source for concern. The Journal of Psychology, 124, 343-345.

Slate, J.R., & Jones, C.H. (1989). Can teaching of the WISC-R be improved? Quasi-Experimental Exploration. Professional Psychology: Research and Practice, 20, 408-410.

Slate, J.R., & Jones, C.H. (1988). Strategies to reduce examiner error on the Wechsler scales. Social and Behavioral Sciences Documents, 18 (Ms. No. 2840)

Slovic, P., & Lichtenstein, S. (1973). Comparison of Bayesian and regression approaches to the study of information processing in judgment. In L. Rappoport & D. Summers, Human Judgment and Social Interaction (pp. 15-109). New York: Holt, Rhinehart, & Winston.

Smith, C.O. (1971). The Structure of Intellect Processes Analyses System: A Technique for the Investigation and Quantification of Problem Solving Processes. Unpublished doctoral dissertation, University of Houston.

Smith, E.R., & Miller, F.D. (1978). Limits on Perception of Cognitive Processes: A Reply to Nisbett and Wilson. Psychological Review, 85, 355-362.

Smith, R., E.T. (1985, September 30). Are you creative? Business Week, 80-84.

Sternberg, R.J. (1985). Introduction: What is an information-processing approach to human abilities? In R. Sternberg (Ed.), Human Abilities: An Information-Processing Approach (pp. 1-4). New York:W.H. Freeman.

Sternberg, R.J. (1981). Testing and cognitive psychology. American Psychologist, 36, 1181-1189.

Sternberg, R.J., & Ketron, J.L. (1982). Selection and implementation of strategies in reasoning by analogy. Journal of Educational Psychology, 74, 399-413.

Svenson, O. (1985). Cognitive strategies in a complex judgment task: Analyses of concurrent verbal reports and judgments of cumulated risk over different exposure times. Organizational Behavior and Human Decision Processes, 36, 1-15.

Truch, S. (1989). The WISC-R Companion. Calgary: Foothills Educational Materials.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, 90, 293-315.

Tversky, A., & Kahneman, D. (1984). The framing of decisions and the psychology of choice. In G. Wright (Ed.)., Behavioral Decision Making. New York: Plenum.

Van Haneghan, J.P., & Baker, L. (1989). Cognitive monitoring in mathematics. In C. McCormick, G. Miller, & M. Pressley, Cognitive Strategy Research (pp.215-238). Springer-Verlag:New York.

Wallsten, T.S. (1980). Cognitive Processes in Choice and Decision Behavior. Hillsdale, NJ: Erlbaum.

Walker, R., Hunt, W., & Schwartz, M. (1965). The difficulty of WAIS comprehension scoring. Journal of Clinical Psychology, 21, 427-429.

Warren, S., & Brown, W. (1972). Examiner scoring errors on individual intelligence tests. Psychology in the Schools, 10, 118-122.

Webb, N.L. (1975). An exploration of mathematical problem-solving processes. Dissertation Abstracts International, 36, 3689A. (University Microfilms No. 75-25625).

Wechsler, D. (1974). Manual for the Wechsler Intelligence Scale for Children-Revised. New York: Psychological Corporation.

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. Psychological Review, 87, 105-112.

Wiggins, N. (1973). Individual differences in human judgements: A multivariate approach. In L. Rappoport, & D. Summers (Eds.), Human Judgment and Social Interaction. New York: Holt, Rinehart and Winston

Woods, P. (1985). Ethnography and theory construction in educational research. In R. Burgess (Ed.), Field Methods in the Study of Education (pp.51-78). Philadelphia: Falmer.

Wright, G. (1984). Behavioral Decision Making. New York: Plenum.

Zungermann, H., & De Zeeuw (1977). Decision Making and Change in Human Affairs. Boston: D. Reidel.

APPENDIX A

WISC-R I

1. INFORMATION Discontinue after 5 consecutive failures.		Score 1 or 0
1. Finger		
2. Ears		
3. Legs		
4. Ball		
5. Nickel		
6. Cow		
7. Week	SEVEN	
8. March	APRIL	
9. Bacon	CALF	
10. Dozen	13	
11. Seasons	SP,SU,W,F	
12. America	C. COLUMBUS	
13. Stomach	GROWLS	
14. Sun	WEST	
15. Leap Year	FEBRUARY	
16. Bulb	EDISON	
17. 1776	BRITISH	
18. Oil	LIGHTER	
19. Border	DK	
20. Ton	2000	
21. Chile	N. AMERICAN	
22. Glass	LIME	
23. Greece	ROME	
24. Tall	5'11 & 1/2"	
25. Barometer	DK	
26. Rust	OXYGEN	
27. Los Angeles	2799	
28. Hieroglyphics	DK	
29. Darwin	NR	
30. Turpentine	SAP	
	Total	Max. = 30

3. SIMILARITIES Discontinue after 3 consecutive failures.		Score 1 or 0
1. Wheel—ball	ROUND AND ROLL	
2. Candle—lamp	LIGHT UP	
3. Shirt—hat	WEAR BOTH	
4. Piano—guitar	MAKE WONDERFUL MUSIC	
5. Apple—banana	FRUITS	Score 2, 1 or 0
6. Beer—wine	ADULT DRINKS (Q) GET YOU DRUNK	
7. Cat—mouse	MAMMALS	
8. Elbow—knee	JOINTS	
9. Telephone—radio	MAKE NOISE (Q) COMMUNICATION DEVICES	
10. Pound—yard	BOTH AMOUNTS	
11. Anger—joy	FEELINGS (Q) HOW SOME PEOPLE REACT	
12. Scissors—copper pan	BOTH COPPER	
13. Mountain—lake	ANIMALS IN BOTH (Q) BOTH NATURE	
14. Liberty—justice	CIVIL RIGHTS	
15. First—last	POSITIONS (Q) IN A SERIES	
*16. The numbers 49 and 121	ODD NUMBERS	
17. Salt—water	CHEMICAL ELEMENTS THAT ARE USED	
	Total	Max. = 30

*If the child gives a 1-point response to Item 16, say, "How else are the numbers 49 and 121 alike?"

7. VOCABULARY Discontinue after 5 consecutive failures.		Score 2, 1, or 0
1. Knife		
2. Umbrella		
3. Clock		
4. Hat		
5. Bicycle		
6. Nail	SHARP OBJECT	
7. Alphabet	ALL THE LETTERS IN THE ALPHABET FROM A TO Z	
8. Donkey	ANIMAL	
9. Thief	SOMEONE WHO STEALS	
10. Join	PUT TOGETHER	
11. Brave	LIKE WHEN YOU SAVE SOMEBODY FROM DYING	
12. Diamond	GEM (Q) SHINY THING	
13. Gamble	PLAY WITH DICE (Q) CHEATING ON CARDS AND STUFF	
14. Nonsense	FOOLISHNESS	
15. Prevent	KEEP SOMETHING (Q) FROM HAPPENING	
16. Contagious	LOTS OF PEOPLE GET THE FLU (Q) INFECTIOUS	
17. Nuisance	A BOTHER OR PEST (Q) MY KID BROTHER WON'T LEAVE ME ALONE	
18. Fable	BOOK (Q) BOOK	
19. Hazardous	POISONOUS (Q) IT COULD KILL YOU	
20. Migrate	TO FLY SOUTH	
21. Stanza	SOMETHING TO DO WITH PAPER AND WRITING	
22. Seclude	IN, LIKE INCLUDE SOMEBODY	
23. Mantis	ANIMAL	
24. Espionage	MISSION (Q) A SPY WITH A MISSION FOR THE GOVERNMENT	
25. Belfry	A CHURCH TOWER LIKE THING	
26. Rivalry	DK	
27. Amendment	NR	
28. Compel	DK	
29. Affliction	YOU HURT SOMEBODY, YOU INFLICT SOMETHING ON THEM	
30. Obliterate		
31. Imminent		
32. Dilatory		
	Total	Max. = 64

9. COMPREHENSION Discontinue after 4 consecutive failures.		Score 2, 1, or 0
1. Cut finger	LET IT BLEED (Q) PUT A BAND-AID ON IT	
2. Find wallet	CALL A RADIO STATION	
*3. Smoke	CALL THE FIRE DEPARTMENT (Q) CALL THE POLICE	
*4. Policemen	TO STOP CRIME (Q) HELP PEOPLE WITH PROBLEMS THAT AREN'T ILLEGAL	
5. Lose ball	BUY ANOTHER ONE TO REPLACE THE ONE I LOST	
6. Fight	BEAT THEM UP, NOBODY IS GOING TO HIT ME, LET ME TELL YOU	
*7. Build house	COOLER IN SUMMER AND WARMER IN WINTER	
*8. License plates	IDENTIFY THE CAR AND TO REPORT ACCIDENTS WHEN YOU HAVE TO DO SO	
*9. Criminals	AS AN EXAMPLE, FOR PUNISHMENT, SO THEY WON'T DO IT AGAIN, BAD PEOPLE	
10. Stamps	TO PAY FOR POSTAGE	
11. Inspect meat	TO PROTECT THE MEAT FROM BEING BAD FOR PEOPLE	
*12. Charity	TAX DEDUCTIBLE, SAFER FOR YOU TO DO SO	
13. Secret ballot	IT IS THE RIGHT WAY	
*14. Paperbacks	CHEAPER, NOT SO BAD IF YOU LOSE A PAPERBACK BOOK	
15. Promise	IT IS A MATTER OF TRUST AMONG PEOPLE (Q) A MATTER OF CONSCIENCE	
*16. Cotton	SOFT AND WARM (Q) CONVENIENT (Q) COMFORTABLE AND COOL	
*17. Senators		
*If the child replies with only one idea, ask him for a second response. Rephrase the test item appropriately, saying, "Tell me another thing to do (reason why, advantage of)...."	Total	Max.=34

WISC-R I KEY

1. INFORMATION Discontinue after 5 consecutive failures.		Score 1 or 0
1. Finger		1
2. Ears		1
3. Legs		1
4. Soil		1
5. Nickel		1
6. Cow		1
7. Week	SEVEN	1
8. March	APRIL	1
9. Bacon	CALF Should NOT have given	1
10. Dozen	13	1
11. Seasons	SP,SU,W,F	1
12. America	C. COLUMBUS	1
13. Stomach	GROWLS	0
14. Sun	WEST	1
15. Leap Year	FEBRUARY	1
16. Bulb	EDISON	1
17. 1776	Should have BRITISH Q	0
18. Oil	LIGHTER	1
19. Border	DK	0
20. Ton	2000	1
21. Chile	N. AMERICAN	0
22. Glass	Should have LIME Q	0
23. Greece	ROME	0
24. Tall	5'11 & 1/2"	0
25. Barometer	DK	0
26. Rust	OXYGEN	
27. Los Angeles	2799	
28. Hieroglyphics	DK	0
29. Darwin	NR	0
30. Turpentine	SAP	
	Total	Max.=30 17

3. SIMILARITIES Discontinue after 3 consecutive failures.		Score 1 or 0
1. Wheel—ball	ROUND AND ROLL	1
2. Candle—lamp	LIGHT UP	1
3. Shirt—hat	WEAR BOTH	1
4. Piano—guitar	MAKE WONDERFUL MUSIC	1
5. Apple—banana	FRUITS	Score 2, 1, or 0 2
6. Beer—wine	ADULT DRINKS (Q) GET YOU DRUNK	2
7. Cat—mouse	MAMMALS	2
8. Elbow—knee	JOINTS	2
9. Telephone—radio	MAKE NOISE (Q) COMMUNICATION DEVICES Should not have Q	0
10. Pound—yard	BOTH AMOUNTS	1
11. Anger—joy	FEELINGS (Q) HOW SOME PEOPLE REACT Should not have Q	2
12. Scissors—copper pan	BOTH COPPER	0
13. Mountain—lake	ANIMALS IN BOTH (Q) BOTH NATURE NO	0
14. Liberty—justice	CIVIL RIGHTS	1
15. First—last	POSITIONS (Q) IN A SERIES Should not have Q	1
*16. The numbers 49 and 121	ODD NUMBERS should have Q	1
17. Salt—water	CHEMICAL ELEMENTS THAT ARE USED	0
*If the child gives a 1-point response to item 16, say, "How else are the numbers 49 and 121 alike?"	Total	Max.=30 18

7. VOCABULARY Discontinue after 5 consecutive failures.		Score 2, 1, or 0
1. Knife		2
2. Umbrella		2
3. Clock		2
4. Hat		2
5. Bicycle		2
6. Nail	SHARP OBJECT should not have given	2
7. Alphabet	ALL THE LETTERS IN THE ALPHABET FROM A TO Z	2
8. Donkey	ANIMAL	2
9. Thief	SOMEONE WHO STEALS	2
10. Join	PUT TOGETHER	2
11. Brave	LIKE WHEN YOU SAVE SOMEBODY FROM DYING	1
12. Diamond	GEM (Q) SHINY THING should not have Q	2
13. Gamble	PLAY WITH DICE (Q) CHEATING ON CARDS AND STUFF NO	1
14. Nonsense	FOOLISHNESS	2
15. Prevent	KEEP SOMETHING (Q) FROM HAPPENING	2
16. Contagious	LOTS OF PEOPLE GET THE FLU (Q) INFECTIOUS NO	1
17. Nuisance	A BOTHER OR PEST (Q) MY KID BROTHER WON'T LEAVE ME ALONE NO	2
18. Fable	BOOK (Q) BOOK	1
19. Hazardous	POISONOUS (Q) IT COULD KILL YOU	2
20. Migrate	TO FLY SOUTH should have Q	1
21. Stanza	SOMETHING TO DO WITH PAPER AND WRITING should have Q	0
22. Seclude	IN, LIKE INCLUDE SOMEBODY	0
23. Mantis	ANIMAL	0
24. Espionage	MISSION (Q) A SPY WITH A MISSION FOR THE GOVERNMENT NO	0
25. Belfry	A CHURCH TOWER LIKE THING	1
26. Rivalry	DK	0
27. Amendment	NR	0
28. Compel	DK	0
29. Affliction	YOU HURT SOMEBODY, YOU INFLICT SOMETHING ON THEM	0
30. Obliterate	no ceiling	
31. Imminent		
32. Dilatory		
	Total	Max.=64 36

COMPREHENSION Discontinue after 4 consecutive failures.		Score 2, 1, or 0
Cut finger	LET IT BLEED (Q) PUT A BAND-AID ON IT	2
Find wallet	CALL A RADIO STATION	1
Smoke	CALL THE FIRE DEPARTMENT (Q) CALL THE POLICE	1
Policemen	TO STOP CRIME (Q) HELP PEOPLE WITH PROBLEMS THAT AREN'T ILLEGAL	2
Lose ball	BUY ANOTHER ONE TO REPLACE THE ONE I LOST	2
Fight	BEAT THEM UP, NOBODY IS GOING TO HIT ME, LET ME TELL YOU	0
Build house	COOLER IN SUMMER AND WARMER IN WINTER Q needed	1
License plates	IDENTIFY THE CAR AND TO REPORT ACCIDENTS WHEN YOU HAVE TO DO SO	1
Criminals	AS AN EXAMPLE, FOR PUNISHMENT, SO THEY WON'T DO IT AGAIN, BAD PEOPLE	2
Stamps	TO PAY FOR POSTAGE	2
Inspect meat	TO PROTECT THE MEAT FROM BEING BAD FOR PEOPLE	2
Charity	TAX DEDUCTIBLE, SAFER FOR YOU TO DO SO	1
Secret ballot	IT IS THE RIGHT WAY SHOULD HAVE Q	0
Paperbacks	CHEAPER, NOT SO BAD IF YOU LOSE A PAPERBACK BOOK	1
Promise	IT IS A MATTER OF TRUST AMONG PEOPLE (Q) A MATTER OF CONSCIENCE NO	2
Cotton	SOFT AND WARM (Q) CONVENIENT (Q) COMFORTABLE AND COOL	0
Senators	NO CEILING	—
If child replies with only one idea, ask him for a second response. Rephrase the test item appropriately, saying, "Tell me one thing to do (reason why, advantage of)...."	Total	Max.=34 20

WISC-R II

9. COMPREHENSION Discontinue after 4 consecutive failures.		Score 2, 1, or 0
1. Cut finger	TREAT IT (Q) TREAT IT WITH THINGS AT HOME	
2. Find wallet	TURN IT INTO LOST AND FOUND	
*3. Smoke	ASK AN ADULT TO HELP AND GO OVER MYSELF	
*4. Policemen	CATCH BAD PEOPLE, ARREST CROOKS, ENFORCE LAWS	
5. Lose ball	LOOK ALL OVER FOR THE BALL (Q) PAY FOR IT IF LOST	
6. Fight	LET HIM BE (Q) AND GET MAD	
*7. Build house	COOLER (Q) FIREPROOF	
*8. License plates	SO THE GOVERNMENT CAN KEEP TRACK OF CARS (Q) WON'T GO TO JAIL	
*9. Criminals	BAD PEOPLE AND AREN'T NICE	
10. Stamps	IT'S THE LAW	
11. Inspect meat	SO PEOPLE CAN GET MEAT (Q) IT MIGHT BE BAD	
*12. Charity	CHARITY NEEDS IT MORE I THINK	
13. Secret ballot	SO POLICE DO NOT CATCH YOU (Q) DEMOCRATIC WAY	
*14. Paperbacks	IT IS OKAY TO BEND THEM HOWEVER YOU WANT (Q) OK	
15. Promise	PEOPLE ARE DEPENDING ON YOU TO KEEP YOUR WORD	
*16. Cotton		
*17. Senators		
*If the child replies with only one idea, ask him for a second response. Sequence the next item appropriately, saying, "Tell me another thing to do (reason why, advantage of)...."	Total	Max = 34

WISC-R II KEY

9. COMPREHENSION Discontinue after 4 consecutive failures.		Score 2, 1, or 0
1. Cut finger	TREAT IT (Q) TREAT IT WITH THINGS AT HOME	1
2. Find wallet	TURN IT INTO LOST AND FOUND	2
*3. Smoke	ASK AN ADULT TO HELP AND GO OVER MYSELF	1
*4. Policemen	CATCH BAD PEOPLE, ARREST CROOKS, ENFORCE LAWS	1
5. Lose ball	LOOK ALL OVER FOR THE BALL (Q) PAY FOR IT IF LOST	2
6. Fight	LET HIM BE (Q) NO AND GET MAD	2
*7. Build house	COOLER (Q) FIREPROOF	2
*8. License plates	SO THE GOVERNMENT CAN KEEP TRACK OF CARS (Q) WON'T GO TO JAIL	1
*9. Criminals	BAD PEOPLE AND AREN'T NICE	0
10. Stamps	IT'S THE LAW should have Q	0
11. Inspect meat	SO PEOPLE CAN GET MEAT (Q) NO IT MIGHT BE BAD	0
12. Charity	CHARITY NEEDS IT MORE I THINK	0
13. Secret ballot	SO POLICE DO NOT CATCH YOU (Q) DEMOCRATIC WAY	
14. Paperbacks	IT IS OKAY TO BEND THEM HOWEVER YOU WANT (Q) OK	
15. Promise	PEOPLE ARE DEPENDING ON YOU TO KEEP YOUR WORD	
16. Cotton		
17. Senators		
the child replies with only one idea, ask him for a second response. Rephrase the test item appropriately, saying, "Tell me other thing to do (reason why, advantage of)...."	Total	Max.=34 12

WISC-R RECORD FORM

Wechsler Intelligence Scale for Children—Revised

NAME WILLIAM JAMES AGE 16-8 SEX M

ADDRESS

PARENT'S NAME

SCHOOL GRADE

PLACE OF TESTING TESTED BY

REFERRED BY

WISC-R STUDY INSTRUCTIONS

Dear Psychologists,

There are four subtests to be scored in this protocol. Please score the responses of each subtest, giving a total after you have scored each one. This protocol does contain some administrative errors. For example, some responses may not have been probed (Q) properly. You may wish to indicate whether you thought the probing was appropriate or not by writing or briefly commenting on the sheets. I am interested mainly in your decisions/judgments in arriving at a score for those responses, not in the score itself.

Some abbreviated responses present in the protocol are: SP = Spring, SU = Summer, W = Winter, F = Fall; DK = don't know, and NR = no response.

Once again thank you for your time volunteered.

THE UNIVERSITY OF BRITISH COLUMBIA

Department of Educational Psychology
and Special Education
Faculty of Education
2125 Main Mall
Vancouver, B.C. Canada V6T 1Z4
Tel: (604) 822-8229
Fax: (604) 822-3302

WISC-R STUDY

Psychologists,

Please find enclosed the following materials:

Two consent forms
A background information form
Four fabricated WISC-R Verbal subtests

Please complete the package, and return all materials to the envelope. However, please retain the second consent form for your own records.

If you have any questions regarding the study procedures or general comments, please feel free to contact me at 737-0866, or Dr. Bill McKee at 822-6572. We will be pleased to hear from you.

All information will be treated as strictly confidential. Thank you for your time volunteered.

Yours sincerely,

Josette Perot

CONSENT FORM (COPY TO BE RETAINED BY PARTICIPANT)

Project Title: Psychologists' Scoring on the WISC-R Verbal Scales

Subject no._____

Thank you for cooperating in this project regarding the differences in psychologists' decision-making in the scoring of Verbal responses on the WISC-R. The data obtained from this study will provide helpful scoring information for other psychologists. Secondly, participation in this study will have a significant bearing on the training of student psychologists. Therefore, your experience in the WISC-R testing practice will be an important contribution to this study.

Participation in this study will require you to provide some personal descriptive information, and score the Verbal Scales of a WISC-R protocol. Arrangements can be made to score the WISC-R protocol at your convenience. The estimated time for this exercise is no more than 30-40 minutes. In a second session, a subsample of psychologists will be asked to participate in a short talk-aloud exercise while scoring some items. The estimated time of the second session is approximately 30 minutes. Again, arrangements will be made in accordance with a time and place that is convenient for you.

Participants will be asked to provide a code name of their own choosing and a phone number only as a means for contact. Additionally, all participants will be provided with a code number to preserve their identities since confidentiality is of the utmost concern in this study.

It is the right of any subject to refuse to participate or withdraw from the study at any time. Such a decision will neither jeopardize nor influence you in any way. Please indicate your willingness to participate in this project by providing your consent below. Please also sign and retain this copy for your records.

If you have any questions or enquiries about this study please feel free to contact me at this number: 737-0866. Or, you may contact Dr. Bill McKee at 822-6572. We will be please to hear from you. Thank you for your cooperation.

Josette Perot
(Master's student, U.B.C.)

****

I consent to participate in the study of psychologists' decision-making, and agree to allow the use of data acquired in this study, and possibly recorded data for research purposes. I acknowledge that I have received a copy of this consent form.

_____ _____

Signature Date

If you are selected for the second session please indicate whether you are willing to participate: yes or no (please circle appropriate response). Please retain this copy for your records.

BACKGROUND INFORMATION

Directions: Please provide the following information about yourself. Your responses will be coded and used to provide information descriptive of participants. This information, as well as all other data you provide during this project, will be treated as confidential.

Subject number: _____

1. Please provide a code name of your choice, and a telephone number where you may be reached. (You may use your first name if you wish; this information will only be used to contact if you wish to participate in the study.)

_____ _____
Name Telephone number

2. Years of college: _____

3. Highest degree earned: _____

4. Number of years as a psychologist: _____

5. Which of the following best describes your professional training (please check the appropriate spot):

a) School psychologist _____

b) Educational psychologist _____

c) Counseling psychologist _____

d) Psychometrician _____

e) Special educator _____

f) Other (please specify) _____

6. Please indicate the appropriate date if you have received formal training (e.g., coursework, supervised administration) on the administration and scoring of each of the following instruments:

WISC-R _____

WPPSI-R _____

Stanford-Binet IV _____

7. Name of school district: _____

The information you provided will remain confidential. Thank you for your cooperation.

APPENDIX B

SCRIPT FOR THINKING-ALOUD PROTOCOL

INTRODUCTION:

1) "You have scored the WISC-R many times in the past. What we are going to do during this time is not much different from your previous scoring experience. You are going to score some items on a Verbal Scale. The only difference is that I am going to ask you to think aloud as you do this scoring. Sometimes when we are working on a problem alone we say out loud whatever passes into our head. Here let me show you. Let's take this example:

TROUT:FISH :: WHALE:_____

Example of experimenter's think-aloud thoughts to above problem:

- They all live in the water - a trout is a kind of fish - a whale is a kind of fish too, but at the same time it is different from a fish, in the sense that it does not really belong to the fish class - if I remember correctly from Biology class in high school a whale is a warm-blooded animal, I think it has mammary glands too - a whale therefore belongs to the general class of mammals - I think that's the correct choice - yes mammal.

2) WARM-UP TASK FOR SUBJECT:

Before you begin the Verbal Scale think-aloud exercise, why don't you practice on this warm-up exercise first. Say whatever comes into you head as you try to think about the response. This is just to get you accustomed to thinking aloud.

CLOCK:TIME :: YARDSTICK:_____

HAIR:SCALP :: TOOTH_____

FRAGILE:BREAK :: FLEXIBLE:_____

3) When the subject is ready for the actual exercise:

"Please say your thoughts as you decide on what specific point value to award the response to the item. Your thoughts can be simple or complex. You may give reasons for your choices. Some responses have a (Q) beside them. The (Q) means that these particular responses have been probed. You may or may not agree with whether the probing was appropriate. You may make comments on this too. You have the freedom to verbalize, or make reference to <u>whatever</u> you feel will help you make the best choice. No matter how irrelevant it may seem, I am interested in all that you have to say. Please begin."

4) Final end of session probing question:

"The information that you have volunteered will be very valuable, however to summarize this experience, can you tell me what strategies generally help you most in scoring responses?"

APPENDIX C

Segmented Units for Subject #13

Item 1: Okay on the first one I would give it a one, it was appropriately queried (recommendations/evaluations).

The additional information doesn't add anything that I feel tells me that the child knows any more than what he new in the first place (self-explanations).

Item 2: The second answer is straight forward, it's the same answer as in the book (manual).

Item 3: I think those are basically the same type of answer (self-explanation).

I would have queried to see if there is an additional type of answer because in my opinion those are the same (recommendations/evaluations).

...but I would have queried in the actual testing situation (recommendations/evaluations).

Item 5: I'd give that a two because he explained it (self-explanation).

Item 6: And number six, I'd give that a two because his first answer is correct and it should not have been queried (recommendations/evaluations).

...and while the second answer is inadequate it should not have been asked in the first place (recommendations/evaluations).

Item 7: And number seven is a two, he's given two different answers (memory).

Item 8: So the government can keep track of cars, um...I'm not really sure what he means by that, um...it's very similar to the way the government keeps a record of vehicles (manual).

I don't know if that would have been given to clarify what this child was saying (monitoring statements).

...that kind of response is not meant to be queried (recommendations/evaluations).

Item 9: Number nine would be a zero because bad people is a zero answer (manual).

Item 10: And number ten that is a zero answer but should have been queried because there is no query you'd have to treat is as a zero (recommendations/evaluations).

Item 11: ...that's a zero response and shouldn't have been queried, although that was an appropriate answer the first one was not (recommendations/evaluations).

Item 12: I would stop as soon as I hit the ceiling, I would not mark the rest of them (general metastrategic statements).

Protocol for Subject #13

*(E): Okay the first thing that I want to say is that you've scored the Comprehension subtest many times, and today is not much different except that I'm just going to ask you to think aloud or talk aloud as you do this scoring so that I can follow what you are doing and your thoughts as you approach the task of scoring.

But before we begin that and just to get you used to thinking aloud I've prepared a few cards, and what I'm going to do I'm going to practice one just to show you how it's done, and then we can do a couple more. This is just to let you see how I solve the problem.

**(S): Okay.

(E): Okay when I look at this card trout is to fish as whale is to something, I know that they're all sea animals. But if I remember correctly from Biology class a long, long time ago, a whale is a little bit different from the other two because it's warm blooded and it has mammary glands. So I think that even though a whale is to a fish as a trout is to a fish a whale is from the mammal class.

So that's just my example of what goes through my head as I'm thinking about the problem. And, I'll just ask you to try one or two before we begin.

(S): Hair is to scalp as tooth is to...okay hair sits on the scalp so I'd say tooth is to mouth, it's in the mouth.

(E): Good, okay, just one more.

(S): Clock is to time as yardstick is to measurement because a yardstick measures something.

(E): Good , okay before we begin I'm just going to, read you a few specific instructions. Okay what I what you to do is just to please say you thoughts as you decide on what specific point value to award the response to the item. Your thoughts can be simple or complex. You may give reasons for your choices. Some responses have a (Q) beside them. The (Q) means that these particular responses have been probed or queried. You may or may not agree with whether the probing was a appropriate. You may make comments on this too. You have the freedom to verbalize or make reference to whatever you feel will help you make the best possible choice. No matter how irrelevant it may see, I am interested in all that you have to say. It's important to me. So, if you don't have any questions you may begin and start scoring, and just tell me whatever comes into you head as you are doing this task.

(S): So you are only worried about what I'm thinking not what's written here so I don't really need to read those as I'm going through.

*Experimenter
**Subject

(E): That's right.

(S): Okay.

(E): If it helps to read you can.

(S) [Number one]: Okay on the first one I would give it a one it was appropriately queried, but the additional answer doesn't add anything that I feel tells me that the child knows anymore than what he knew in the first answer. So I'd only give that a one.

[Number two]: The second one is straight forward. It's the same answer as in the book (gives a two).

[Number three]: And number three, um, I think that those are really basically the same type of answer, um, I would have queried to see if there is an additional type of answer because in my opinion those are the same. So I would just have to give a one because that second answer is not there, but I would have queried it in the actual testing situation.

[Number four]: And number four I would have given it a one and I would have queried that as well to look for a second answer because they all fall within the same category.

[Number five]: Number five I'd give that a two because he explained it, if he couldn't find it after the query then he'd pay for it...(gives a two).

[Number six]: And number six, I'd give that a two because his first answer is correct and it should not have been queried, and while the second answer is inadequate it should not have been asked in the first place so his first answer is worth a two.

[Number seven]: And number seven is a two, he's given two different answers.

[Number eight]: So the government can keep track of cars, um... I'm not really sure what he means by that, um...it's very similar to the way the government keeps a record of vehicles, um, the query not having given the test myself, the query, I don't know if that would have been given to clarify what this child is saying but even if it was that kind of response is not meant to be queried. So, I think I would give a one for the first part, the won't go to jail doesn't mean anything.

[Number nine]: Number nine would be a zero because bad people is a zero answer.

[Number ten]: And number ten that is a zero answer but should have been queried because there is no query you'd have to treat it as a zero.

[Number eleven]: Okay again in number eleven, that's a zero response and shouldn't have been queried, although that was an

appropriate answer the first one was not.

[Number twelve]: Number twelve is a zero response again...now that would be a ceiling, do you still want the other ones marked?

(E): If you would usually stop there.

(S): I would stop as soon as I hit a ceiling, I would not mark the rest of them.

(E): Alright, so we can stop here, but just to kind of summarize your experience, if you are presented with the kinds of difficult type response that you're not sure about are there any general kinds of strategies that you try to fall back on?

(S): What I do is that I read through all the responses that are given in the manual and try and identify the quality...first of all does that fit into any one of the responses that are given in the manual, and if it does not then I try and identify the quality of the answer with what I think the intent of the quality was in the manual. And, usually I'm only marking these if I'm the one who's given it, so I try and recall what the child was talking about that I might not have written everything down, but I do try and write it all down, but mostly I just weigh it against the quality in the book versus the quality of what the child said.

(E): Okay, thank you very much.

(S): That's all.

(E): That's it.

Segmented Units for Subject #2

Item 1: So...that doesn't really expand...so that should be a one still (self-explanation).

Item 2: I'd give that a two, go over myself, go see what's the matter (manual).

Item 4: So we need two reasons (planning).

Arrest crooks is the same one...and enforce laws...is the same category so we need to ask for another response on that one (planning).

Item 5: Okay, look all over for the ball is part of a one, yes and there's a (Q), pay for it if lost makes it a two (manual).

Item 6: Let him be is two points, so that doesn't need to be cued (recommendations/evaluations).

And if it's cued and he spoils it with get mad...you see he spoiled it and I'm just not sure what to do with that, so what do I know about spoiling responses (monitoring statements).

Actually I'm curious, I'm curious [about spoiling rules] (general metastrategic statements).

Item 8: But that interesting you see because keeping track of cars you might even cue 'cause it really...what it says here is, um, the way the government keeps a record of the vehicles is different (manual).

Item 9: Bad people are criminals, or bad people aren't nice is zero...and that a zero with no question (memory).

Item 10: It should be questioned (recommendations/evaluations).

Item 11: That should not have been questioned (recommendations/evaluations).

I'm going to check again though (monitoring statements).

...I'll just check out my thinking...see where my thinking is coming from (monitoring statements).

Item 12: Charity needs it more I think, well that's another one that's scorable. That's just a zero (memory).

Item 13: So people do not catch you, what does that mean (monitoring statements)?

I would question that one just like they did (recommendations/evaluations).

It's the democratic way doesn't give you any of the criteria (manual).

I would have questioned that one. That was not clear (monitoring statements).

Item 14: It's okay to bend them however you want...that's one to question because...that's under the cheaper category...it's okay to bend it or fold it (manual).

...that one should not have been questioned (recommendations/evaluations).

Item 15: People are depending on you to keep your word..it's a one, and now we have to ask for another [response] (planning).

Protocol for Subject #2

*(E): What I'd like to say is that you've scored the WISC many times in the past but today we're not going to do anything that much different except I'm going to ask you to think aloud as you do this scoring, but just in case you're not quite aware of what a think aloud is, I'm going to try a little exercise.

**(S): Certainly, sure.

(E): I brought along just a couple of cards of some analogy type questions, and I'm going to do a little think aloud myself which will take just a few seconds. What it involves is saying whatever comes into your mind, for instance: trout, fish, whale, it's an analogy question...and I'm going to tell you whatever flows into my mind. First of all, they're all from the fish family, um, but at the same time a whale is just a little bit different if I remember from Biology class in high school because a whale is warmblooded and it has mammary glands, therefore, it's not quite from the fish class but I would say it's from the mammal class, so I think the answer is mammal. Trout is to fish, as whale is to mammal.

Now I just want you just to try one more on your own, just to get accustomed to saying these things out loud.

(S): Okay, so I should try it like you did?

(E): Sure, yes whatever comes into your mind.

(S): Okay, um, so a clock is to time as a yardstick is to...okay, a clock, a clock tells time, a yardstick measures, so measure?

(E): Okay, just one more before we begin. Whatever comes into your mind as you are thinking about the problem.

(S): Hair is to scalp, hair is on the scalp, a tooth is in the mouth. Mouth.

(E): Good. Okay now I'll just read you a few written instructions I have before we begin the actual scoring. And, what I want you to do is just to please sat your thoughts as you decide on what specific point value to award the response to the item. Your thoughts can be simple or complex. You may give reasons for your choices. Some responses have a (Q) beside them. The (Q) means that theses responses have bee probed. You may or may not agree with whether the probing was appropriate. You may make comments on this too. You have the freedom to verbalize, or make reference to whatever you feel may help you make the best choice. And, no matter how irrelevant it may seem, I am interested in all that you have to say.

*Experimenter
**Subject

(S): Sure.

(E): Okay, so you can begin when you're ready, whatever comes into you mind as you're scoring. I'll just follow along.

(S): So, I'll just talk out loud.

(E): Sure, yes.

(S): [Number one]: Okay. So, um, treat it, you treat it. Treat it is a cue, so that's right, you treat it with things at home. So...that doesn't really expand on what so that would be a one still.

[Number two]: Find a wallet, okay. Ah, what do you do, turn it into the lost and found, ah, yes. Turn it into the lost and found is a two points. So that's a two.

[Number three]: Smoke, okay. Um, what should you do you do if you see thick smoke? Ask an adult to help is a one, and go over myself...um...go see what's the matter. I'd give that one a two, go over myself, go see what's the matter...okay, so I'd give that a two...um.

[Number four]: Okay, number four. What are some reasons why we need policemen? So, we need two reasons, catch bad people...ah, is a one, is one of them. Arrest crooks is the same one...and enforce laws...is the same category so we need to ask for another response on that one. And, so that is a one as it is.

[Number five]: And, okay. What's the thing to do if you lose a ball that belongs to one of your friends? Okay, look all over for the ball is part of a one, yes and there's a (Q), pay for it if lost makes it into a two.

[Number six]: What's the thing to do if a boy much smaller than yourself starts a fight with you? Let him be is a two points, so that doesn't need to be cued, and if it's cued and he spoils it with get mad...you see he spoiled it and I'm just not too sure with what I would do with that, so what do I know about spoiling responses, I think what would happen is that he gets a zero when he spoils it like that...yah, I think that he spoiled it and it's going to have a zero on that one. Actually, I'm curious now, I'm curious...[looks for rules about spoiling in the manual], a zero because he spoiled it, yah okay so that must have been into, into my repertoire here. Things to do...

[Number seven]: Okay, seven. In what ways is a house built of brick or stone better than one built of wood? Okay, it's cooler, that's one, and I guess (Q) is the same as an (R) here is it, I'm not sure, there's no R's here.

(E): You would ask for another response.

(S): Yes, so would that mean, are you signalling that by a (Q)?

(E): You can change it if you want.

(S): No, it doesn't matter, is that a signal?

(E): That's a (Q), yes.

(S): So, and it's fireproof is good too, so that's two. So that's two points.

[Number eight]: License plates. So the government can keep track of cars, um, um...that's one, and won't go,to jail isn't one. So, it's a one. But that's interesting you see because keeping track of cars you might even cue 'cause it really, what it says here is, um, the way the government keeps a record of the vehicles is different, so to clarify that I might cue and then say, now tell me another way, so that's what I'm just wondering about there. So, anyways that would be a one.

[Number nine]: Bad people are criminals, or bad people aren't nice is zero...and that's just a zero with no question, it's, that's a zero.

[Number ten]: Stamps, it's the law is a zero. Oh no, yes it is, it's a zero, and it should be questioned, it should be questioned. And, but as it is there it's a zero.

[Number eleven]: Okay, so people can get meat. No...it's just a zero, and that didn't need to be, it shouldn't have been questioned. When it's questioned and it might be bad, what do I do there? So the student wasn't really allowed to be questioned there under the standardized way, so I don't think I can score that, I think that's a zero. And, I'm going to check again though because this is good review for me. So,if you don't mind me taking the time...

(E): Sure...

(S): ...that would be my thinking on that one, and then I'll just check out my thinking...see where my thinking is coming from...[checks manual]...this is actually an interesting, an interesting exercise to do.

(E): A few people told me they're enjoyed it.

(S): Yes, it is interesting, because it is interesting to look at why do I think that, you know...um....so if...um, [reads manual out loud] the score may stay at zero or it may be raised

to one or even two depending on the quality of the child's elaboration - so, it's to the child advantage if you make this kind of mistake because you can then raise the score. Okay, so he will get a two because it might be, no it might be bad is a one, that's a one isn't it...I think it's one point for that one.

[Number twelve]: Charity needs it more. More, charity needs it more I think, well that's another one that's scorable. That's just a zero, and you just leave it at that [checks manual].

[Number thirteen]: Okay, secret ballot. So people cannot catch you what does that mean...so people do not catch you...so the police don't catch you, so people do not catch you... I would question that one just like they did, and um it's the democratic way, doesn't give you a...it doesn't give you what any of those criteria...it's the democratic way, now that's a one there, okay so, yes, that was a good question because then you get a one point out of that...now, I would have questioned that one. That was not clear.

[Number fourteen]: Is it okay to...okay, paperbacks. It's okay to bend them however you want..um,um, that's one to question because...that's under the cheaper...it's okay to bend or fold it, it's okay to bend or fold it, doesn't have a question though, so you can bend it however you want, so, that one shouldn't have been questioned 'cause that's one that's already there, that's already worth one point. And, then, oh I'm sorry the (Q) is an R here as well, that's confusing for me. So when that becomes an R, so then that's when I would say tell me another way, so that's an R would be tell me another way. And, don't know so it's a one.

[Number fifteen]: People are depending on you to keep your word...it's a one, and now we have to ask for another, no this isn't a two, this is the promise. So the promise is good, that's a one point, people are depending on you to keep you word is a one. So, now we've stopped, and we don't have a ceiling. So, I'm not sure why we've stopped here. It needs to go on to sixteen and seventeen. Yes.

(E): So you would go on.

(S): I would go on.

(E): Okay, just to summarize your experience, what are any general strategies you use in the face of very difficult type responses?

(S): Yah, I'm just aware of what I did do. First of all, I think about it, give it my...I come up with, well I come up with

an idea well I think of what that is, and then I would go back and have a look. If I'm really, if it's one I'm really stuck on, look back and doublecheck the criteria first the ones right by the answers, and then back to the book, yah, yah. You don't get very many where, I don't have experience questioning where, where it's not a questioning response. So, that's why that's an unusual experience because for me I always have that part of my book open to, even after all these years. I always use this, I don't take anything for granted. I always have this open. So, I would check it then. So, that's a difficult one, um, but if I did, then I would have to go through just what I did, I would have to go through and see if that's the way that it was right. Like I mean this was interesting for me, number eleven, but then because it's the opposite, the other way if you question it they can lose it, but if you also, if you question it they can also gain it so really that's interesting for me because there's nothing to lose by questioning, really in that case, is there, if you're unclear. If you question and they've given you a response that's accurate, you question they can below [lower] that. But, if they aren't or they're not clear or they've given you an inaccurate response and you question and then get it, what that suggests to me is that you're better off to question than not to question. Yes, so that's interesting.

APPENDIX D

Table D1

Frequencies (Percentages) of Verbalizations for Non-Problematic Items

Item 1

Category	Frequency (Percent)
Memory	7 (29)
Self-explanations	5 (21)
Monitoring Statements	3 (13)
General Metastrategic Statements	3 (13)
Manual	3 (13)
Recommendations/Evaluations	3 (13)
Planning	0
	n=24

Item 2

Memory	6 (56)
Manual	3 (33)
Monitoring Statements	1 (11)
Planning	0
Self-explanations	0
General Metastrategic Statements	0
Recommendations/Evaluations	0
	n=9

Item 9

Manual	5 (29)
Memory	4 (24)
Monitoring Statements	4 (24)
Recommendations/Evaluations	2 (12)
Monitoring Statements	1 (6)
Planning	1 (6)
General Metastrategic Statements	0
	n=17

Table D2

Frequencies (Percentages) of Verbalizations for Difficult Items

Item 4

Manual	8 (30)
Planning	5 (19)
Self-explanations	5 (19)
Recommendations/Evaluations	4 (15)
Monitoring Statements	3 (11)
General Metastrategic Statements	1 (4)
Memory	1 (4)
	n=27

Item 6

Recommendations/Evaluations	8 (36)
Memory	5 (23)
General Metastrategic Statements	2 (9)
Monitoring Statements	1 (5)
Manual	6 (27)
Planning	0
Self-explanations	0
	n=22

Item 11

Recommendations/Evaluations	6 (33)
Manual	5 (28)
Monitoring Statements	4 (22)
Planning	1 (6)
Self-explanations	1 (6)
General Metastrategic Statements	1 (6)
Memory	0
	n=18