Order Number 9222328

Clinical decision-making in neuropsychology: Bootstrapping the neuropsychologist utilizing Brunswik's Lens Model

Gaudette, Marc Donald, Psy.D.

Indiana University of Pennsylvania, 1992

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

CLINICAL DECISION MAKING IN NEUROPSYCHOLOGY:
BOOTSTRAPPING THE NEUROPSYCHOLOGIST UTILIZING BRUNSWIK'S LENS MODEL

A Dissertation
Submitted to the Graduate School
in Partial Fulfillment of the
Requirements for the Degree
Doctor of Psychology

Marc D. Gaudette
Indiana University of Pennsylvania
May 1992

Indiana University of Pennsylvania
The Graduate School
Department of Psychology

We hereby approve the dissertation of

Marc D. Gaudette

Candidate for the degree of Doctor of Psychology

April 3, 1992
(date of signature)

Donald U. Robertson
Donald Robertson, Ph.D.
Professor of Psychology, Committee Chair

April 3 1992
(date of signature)

Graham Ratcliff
Graham Ratcliff, D. Phil.
Harmarville Rehabilitation Center; Department of Psychiatry, WPIC, and University of Pittsburgh. Committee Advisor

April 3 1992
(date of signature)

Mario Sussmann
Mario Sussmann, Ph.D.
Professor of Psychology, Committee Advisor

April 15, 1992
(date of signature)

Virginia L. Brown
Virginia L. Brown, Ph.D.
Associate Dean for Research
The Graduate School and Research

Title: Clinical Decision Making in Neuropsychology: Bootstrapping the Neuropsychologist Utilizing Brunswik's Lens Model

Author: Marc D. Gaudette

Dissertation Chairperson: Donald U. Robertson, PhD

Dissertation Committee Members: Graham Ratcliff, D. Phil.
Mario Sussmann, PhD

This study utilized the Brunswik Lens Model and its corresponding mathematical indices to examine the judgment and decision making components of neuropsychologists (Brunswik, 1955; Goldberg, 1970; Hammond, Hursch & Todd, 1964). Specifically, the "bootstrapping" model (Dawes, 1971) of judgment and decision making research was employed. Bootstrapping is a combined human – statistical judgment model whereby a judge makes decisions and a mathematical or statistical linear model of that judge's decisions is computed (Kleinmuntz, 1990).

The purpose of this study was twofold: (a) Apply the bootstrapping model to the analysis of judgments made by neuropsychologists. It was hypothesized that the linear model of the judge (based on nine a priori chosen predictor cues) would be equal to or superior to the judge (using all cues) in judgmental accuracy consistent with previous rcscarch findings in clinical psychology. (b) Examine for differences between expert and novice neuropsychologists in their judgmental accuracy and decision making processes. Consistent with previous expert – novice research (Garb, 1989), it was hypothesized that there would be no significant or consistent differences between expert and novice neuropsychologists.

Six neuropsychologists participated in the study: Three were classified as experts and three as novices based on professional training and experience criteria. All judges were provided with the same 50 neuropsychological protocols and asked to make two judgments: the presence vs absence of brain damage and the localization of brain damage (right hemisphere, left hemisphere or diffuse damage). The 50 protocols were comprised of: 10 normal, 10 right hemisphere, 10 left hemisphere and 20 diffuse brain damaged records. Each protocol contained 20 to 29 cues consisting of demographic information and test scores from selected neuropsychological instruments.

Results supported the two hypotheses. Namely, the linear model of the judge was equal to or superior to the judge in judgmental accuracy for the presence/absence and the localization judgments. Also, there were no meaningful differences between the experts and novices in terms of judgmental accuracy or decision making processes for the two judgments.

Given the superiority of linear judgment models found in this study, as well as in research over the past 35 years, neuropsychologists are encouraged to use mathematical models in making categorical judgments (e.g., localization of brain damage) from assessment data.

ACKNOWLEDGEMENTS

This study could not have been completed without a great deal of assistance from a number of people.

First, I extend great thanks to my dissertation committee chair, Dr. Donald Robertson. Don provided continued support, assistance and guidance. He was very accommodating to scheduled and impromptu meetings which made "life" easier for me. In addition, he allowed me ample freedom, sometimes for better and sometimes for worse, in making decisions about the many choices that needed to be made in a study of this complexity. His non-domineering style helped me to be a better researcher. Finally, he helped to secure needed funds for the study when other internal sources did not come through. Overall, it was a pleasure having Don serve as my dissertation chairperson.

Dr. Graham Ratcliff has my gratitude for his role in the study. Graham was very accommodating to meet with me when issues needed to be discussed, not an easy task for a very busy man. He was essential in the task of making contact with potential judges used in the study. In addition, his expertise in neuropsychology provided valuable guidance in the development of the methodology. I felt honored to have Graham on my committee.

I thank Dr. Mario Sussman for particpating on my committee. Reo provided useful editorial feedback and statistical advice.

My sincere appreciation is extended to Dr. Robert Bornstein. Bob graciously allowed me access to his neuropsychological laboratory in order for me to obtain the necessary neuropsychological protocols used in the study. His policy to open-up his lab to other researchers is interpreted as a sign of his leadership in the field of neuropsychology.

I am grateful to the six anonymous neuropsychologists who participated in this study. They provided a lot of their own professional time to the judgment task with minimal compensation. I sincerely appreciated their time, cooperation and effort.

I extend many thanks to Dr. Andrew Krouskop, Chairperson, Harmarville Rehabilitation

Center's IRB Committee for giving me permission to collect normative data. Also, I sincerely recognize the efforts and assistance I obtained from Pauline Egan, Director, Volunteer Services at Harmarville. Mrs. Egan, her staff and volunteers unselfishly offered their time so that I could collect normative data.

Similarly, I am grateful to Mr. Douglas Motter and Ms. Bernadine Robertson for their cooperation in the process of collecting normative data at St. Andrew's Village. Ms. Hazel Hill, Interim Director, Volunteer Services, and her volunteers were instrumental to the success of gathering the additional normative data I required.

I also provide a note of thanks to a fellow student, Chuck Turek, for his time and effort in helping me to collect normative data. Chuck saved me a lot of time and I appreciated his assistance.

Last, but certainly not least, I lovingly acknowledge the efforts of my wife, Cyndi. She probably read more drafts then anyone and provided good editorial guidance. In addition, she helped in the tedious and boring task of double checking the accuracy of the many values in the tables created in this study.

TABLE OF CONTENTS

LIST of TABLES

LIST OF FIGURES

INTRODUCTION

The controversy over the superiority of mathematical (e.g., actuarial) judgments versus human judgments in clinical decision making received its impetus from Meehl's (1954) book - "Clinical vs Statistical Prediction: A Theoretical Analysis and a Review of the Evidence." Meehl eloquently presents theoretical analysis and empirical data strongly supporting a position that a mathematical model of a clinician's judgment will consistently outperform the clinician's judgment. The controversy has lasted for over 35 years with most of the research published in the 1960's and 1970's. The overwhelming evidence is that an actuarial judgment or a mathematical model of the clinician's judgment is equal and often superior to the clinician's judgments in clinical decision making (Dawes, Faust, & Meehl, 1988).

Bootstrapping is a method of representing human judgment with a linear regression model. It is used in studies of clinical decision making and is primarily concerned with how clinicians use cues (e.g., test scores) to make judgments about an outcome criterion. The term "bootstrapping" (proverbially - to pull judges up by their bootstraps, Camerer, 1981) was first used in a published paper by Dawes (1971), although his colleagues at the University of Oregon and Oregon Research Institute also deserve credit for introducing and investigating this particular aspect of judgment. In bootstrapping research, the clinician's judgments (outputs) are compared to a mathematical model of the clinician's judgments. Essentially, a linear multiple regression equation is developed whereby the clinician's judgments are regressed against the cue values used in making the judgments. The regression model of the clinician's judgments most often outperforms the clinician (Dawes, Faust, & Meehl, 1988; Sawyer, 1966).

Wedding and Faust (1988) recently reviewed the judgment and decision making research in neuropsychology. The researchers reported that (a) relatively few studies have been published that are directly applicable to clinical judgment and decision making in neuropsychology. (b) The results of judgment research in neuropsychology are consistent with research in clinical psychology. For example, in clinical psychology, clinical training and experience are generally not significantly related to the validity of clinical judgments (Garb, 1989). These findings have been substantiated in the area of neuropsychology by Faust, Guilmette, Hart, Arkes, Fishburne, and Davey (1988). (c) They noted that there have been no bootstrapping studies conducted utilizing neuropsychological data, and that such studies would be especially useful in order to understand the judgement strategies of clinical neuropsychologists.

The purpose of this study was to bootstrap the clinical neuropsychologist. Experienced neuropsychologists made judgments concerning the presence vs absence of brain damage and localization of damage. A linear regression model of the judgment was developed and compared to the clinician's judgment.

This study was based on a controversy that started in the 1950's. Therefore, an appropriate starting point in this Introduction is an overview of the clinical vs statistical judgment debate. The next two sections relate the reasons for the superiority of statistical methods and their infrequent use in clinical practice. The fourth section introduces the Brunswik Lens Model as it provides the conceptual design to studying clinical inferences. Also, in this section, mathematical analyses accompanying the Lens Model are outlined. This will be followed by a detailed review of two empirical studies on bootstrapping research which bridge the conceptual and mathematical issues of clinical inference to real world empirical questions (e.g., are clinicians or linear models of the clinician more accurate about judging a psychotic profile on the MMPI). The sixth section addresses an interesting peculiarity in the use of optimal vs equal weighting coefficients in the regression equation. Next, the specialty of neuropsychology is introduced, a representative sample of clinical decision making research in neuropsychology are reviewed, and a rationale is provided as to the nature of this study. The final section delineates the hypotheses of this study.

Clinical versus Statistical Judgment

Paul Meehl was not the first researcher to examine the issue of clinical decision making or actuarial judgments, but his "little book" (Meehl, 1986) published in 1954 is credited with starting the debate of the superiority of statistical (e.g., actuarial, bootstrapping) over clinical judgment. Thirty-five years of research later, the fundamental conclusion remains the same (Dawes, Faust, Meehl, 1988; Sawyer, 1966). This finding stands regardless of the experience of the clinician or whether the clinician is novice or expert (Sawyer, 1966; Wedding,1983). In addition, the superiority of statistical methods over clinical methods is generally not affected when clinicians are provided with more information. So, even when the clinician is given access to information that is not incorporated in the actuarial method, the statistical method is often superior (Sawyer, 1966; Wiggins, 1981).

The superiority of statistical methods is, in part, a result of it's mathematical features. The multiple regression procedures which are the typical statistical analyses performed in these studies are based on a maximization procedure. The resulting linear combination of the variables squeezes out every bit of predictive power and, thus, correlates maximally with the criterion (Stevens, 1986). Therefore, only those variables which provide adequate predictive power will contribute to the judgment, while those variables which add little predictabilily will not be weighted significantly. In contrast, clinical judges may not weight the cues appropriately. Also, judgments made by clinicians are suspect to many potential judgment biases. For example, Arkes (1981), Dawes, et al. (1988), and Wedding and Faust (1988) have documented factors that contribute to inaccurate clinical judgments, e.g., hindsight bias, illusory correlation, confirmatory hypothesis testing, and overreliance on salient data. In addition, even if the clinician takes measures to avoid these always lurking impediments to judgment, the clinician may be hampered by fatigue, boredom, interpersonal distractions or attentional limitations (Einhorn, 1986; Goldberg, 1970).

Although the issue is often couched in terms of clinical _versus_ statistical prediction, the research comparing mathematical judgment models to clinical judgments was not intended as an attack on clinicians. On the contrary, mathematical models of clinical inferences has greatly assisted the field of clinical psychology to ascertain potential pitfalls to clinical decision making, as well as develop solutions to these impediments (Arkes, 1981).

Also, decision making research has identified judgments which are best handled using clinical methods. (a) The judge may be more versatile and flexible than the statistical method. This may be the case when the judgments are based on cues from new tests, when no tests are available to tap some judgment, or when data cannot be coded into a regression equation. (b) When rare or unusual events enter into the judgment process, regression models may be inadequate. For example, Meehl (1957) presents an amusing scenario whereby a scientist is mathematically predicting the probability of one of her colleagues attending a movie on a particular night. The scientist constructs a mathematical model based on factors or cues she believes are relevant to the prediction (e.g., colleague's age, academic specialty, and introversion score). The resulting model yields a probability of 0.90 that her colleague will attend a movie tonight. But, if the scientist learned that her colleague just suffered a broken leg, she probably would not base her prediction on the mathematical model, becasue the predictor "broken leg" was not part of the regression model. The scientist's sample from which the probability of 0.90 was obtained, plus her cross-validation sample did not contain a single instance of a broken leg. The scientist predicts that her colleague will not attend the movies tonight, and rightly so, because she knows that having a broken leg is a relatively immobilizing experience, while attending a movie is a relatively mobilizing experience. (c) If clinical judgments are based on firm theoretical underpinnings, then such judgments may be superior to a statistical model. But, given the state-of-the-art of theory in psychology, this appears to be an uncommon occurrence (Dawes et al., 1988; Meehl, 1957; Phares, 1979).

Sawyer (1966) offered an additional conceptual framework to the understanding of the clinical versus statistical debate. Specifically, he differentiated prediction and measurement.

Prediction (whether clinical or statistical) depends on how the cues are combined, while measurement (whether clinical or statistical) depends on how the cues are collected. Furthermore, he substituted the word "mechanical" for the word "statistical", with the premise that the word mechanical better captures the process of cue collection and cue combination. Sawyer developed a table for the classification of prediction methods which is composed of (a) modes of cue collections and (b) modes of cue combinations (see Table 1. Adapted from Sawyer, 1966, p. 181).

A description of the eight classifications will be provided to make Table 1 more understandable. (a) The first classification, pure clinical, is concerned with clinically collected and clinically combined cues only. The clinician formulates a prediction based on interview data and/or observational data only. There are no test data or other objective information available. (b) Trait ratings relate clinically collected cues that are mechanically combined. The data collected in this method are the same as in the first classification, but the data are combined mechanically. (c) Profile interpretation takes mechanically collected cues and clinically combines them. For example, a clinician is provided with a set of scaled scores from the MMPI and asked to make some kind of prediction about the individual. (d) The fourth classification, pure statistical, is concerned with mechanically collected cues that are mechanically combined. An example might involve the collection of test scores and biographical information that are combined in a multiple regression equation to predict some outcome. (e) Clinical composite makes use of both modes of cue collection methods and clinically combines them. Sawyer suggested that this is the most frequent clinical assessment strategy. Here, interview cues, test scores, and observations are integrated by the clinician who then makes a prediction. (f) In mechanical composite, both types of cue collection methods are employed and mechanically combined. Cues are consistent with those collected in the clinical composite, yet they are mechanically combined via multiple regression equations. (g) The seventh classification is clinical synthesis. A prediction based on a mechanical classification procedure is incorporated into other clinical data and a clinical prediction is then made. (h) In mechanical synthesis, a prediction based on a clinical combination of cues is used as a datum which is combined mechanically with other cues to yield a prediction. (It appeared inappropriate for Sawyer to have labelled his Table "Classification of Prediction Methods," because the Table involves an examination of both modes of data collection (measurement) with modes of data combination (prediction). It would appear more appropriate to have labelled the table – Classification of Judgment Methods. The word "judgment" might be better because it captures both issues of data collection and data combination and it does not confuse the reader.)

Table 1

Classification of Prediction Methods

	Mode of cue combination	
Mode of cue collection	Clinical	Mechanical
Clinical	1. Pure clinical	2. Trait ratings
Mechanical	3. Profile interpretation	4. Pure statistical
Both	5. Clinical composite	6. Mechanical composite
Either or Both	7. Clinical synthesis	8. Mechanical synthesis

Sawyer (1966) reviewed 45 studies resulting in 75 comparisons based on his structure of classification methods (see Table 1). Overall, Sawyer (1966) found the mechanical mode of cue combination to be superior or at least equal to the clinical mode whether the cues are collected clinically or mechanically. Cues collected by both modes that are clinically combined (i.e., clinical composite) offers inferior prediction to mechanically collected cues that are clinically combined (i.e., profile interpretation). Clinical combination that incorporates a mechanical prediction (i.e., clinical synthesis) is inferior to a lower-ranked method of mechanical composite. An implication of the data is that the clinician probably does not add to prediction by formulating a clinical judgment, but by providing objective cues that can be incorporated in a multiple regression equation.

Using Sawyer's classifications, this study involves a comparison of "profile interpretation" (i.e., human judgment) versus a combination of profile interpretation and "pure statistical" (i.e., linear model of the judge). This point will be elaborated in the fourth section of the Introduction.

It is very important to understand that actuarial proponents do not purport that the mathematical models fully explain the cognitive processes of the clinician. Hoffman (1960) borrowed the term "paramorphic" from mineralogy to relate mathematical models and clinical judgment. Paramorphism is defined as a structural alteration of a mineral without change of chemical composition (American Heritage Dictionary). Hoffman (1960) briefly discussed the use of mathematical models in science and suggested that mathematical models provide an objective formulation of a phenomenon. The usefulness or quality of the equation(s) is based on how well it accounts for the data, how much predictive value it has, and how much it contributes to a greater theoretical understanding of the phenomena under study. The model is not required to completely account for the internal operations of the organism. Statistical methods are a representation of the human judge at a description level and they also provide predictive value. The weighting of the variables, as in a multiple regression analysis, provides a mathematical model of judgment or a mathematical simulation of the judge, but does not purport to completely explain or account for all aspects of the human judgment process. Thus, statistical methods provide paramorphic, as opposed to isomorphic (implying an one-to-one correspondence), representations of clinical judgments.

Einhorn (1986) alluded to the issue of Hoffman's paramorphic representation. He stated that the statistical model has access to only a limited number of predictor variables that it will combine in some mechanical manner. Such a model can never capture the full richness and complexity of the judgment under study. But, neither can the clinician. That is, the clinician may not be aware of or be able to accurately relate how he/she precisely weighted and combined information that lead to the judgment (see Ericsson & Simon, 1980; Nisbett & Wilson, 1977 for a discussion of verbal reports of mental processes). Therefore, it becomes an empirical question as to what method is superior. Thirty-five years of research has supported the statistical model as the winner.

This section made it clear that statistical models of judgments are often superior to clinical judgments. The next section will delineate the reasons why the statistical models outperform the clinical judge.

Explanations to Account for the Superiority of Statistical Models

Dawes and Corrigan (1974) provided three reasons for the superiority of linear models. (a) Linear models have been used in areas where the relationship between the criterion and predictor variables tend to be "conditionally monotone" (p. 98). This is, predictor variables (independent variables) can be scaled in such a way that higher scores on the independent variables predict higher scores on the dependent variable (criterion) independently of the scores of the remaining variables. For example, no matter how an individual scores on other variables, the higher they score on subtests of the WAIS-R the more likely they will be predicted to have a higher IQ. (b) The weights achieved by the optimal linear combination of the variable are unaffected by the unreliability of the criterion variable. This is because error, due to unreliability, results in a constant reduction in the weights. (c) Measurement error in the predictor variables tends to enhance linearity.

Einhorn (1986) stated that the statistical models outperform the clinician, because the mathematical models tend to be based on simple rules that accept error. Accepting some error up-front may improve accuracy of judgments. Einhorn related findings from research conducted in the 1950's that demonstrate the probability matching phenomena. In these studies, subjects are to predict the occurrence of a red or green light illuminating. They are provided with money upon correct predictions. However, the lights are programmed to provide a random pattern of illumination that is in the proportion of 60% red and 40% green. The general findings of these studies is that the subject mirrors the programmed proportion. That is, the subject predicts 60% red and 40% green. Now, given that the subject predicts red on 60% of the trials and red occurs on 60% of the trials the subject will be correct on 36% of the trials. Similarly, in the prediction of green, the subject will be correct on 16% of the trials. So, the subject will be correct on 52% of the trials (36% + 16% = 52%). But, consider what would happen if the subject employed a simple rule of always predicting the most frequent color. It is important to understand that such a simple rule accepts error. The result of following this strategy will produce a correct prediction rate of 60% which is superior to the other more complex(?) method. Thus, simple linear models provide more accurate judgments by accepting some error and reducing some of the unreliability and spuriousness in the human judgment process.

Although the superiority of statistical models over clinical judgments have been demonstrated, it is puzzling that there are few statistical models in use in clinical practice. Why is such a powerful tool not being utilized? This is addressed in the next section.

The Use of Statistical Models in Everyday Clinical Practice: An Oxymoron?

It is quite clear that statistical models are equal and often superior to the human judge in accuracy of prediction. It is equally clear that statistical models are few and infrequently used in everyday clinical practice. What accounts for this contradiction? This section will first present methodological limitations of statistical models and, second, present reasons for the lack of acceptance of the models.

Methodological Limitations:

Statistical models may not maintain their high accuracy on data or samples from which they were not derived. Basic issues in regression analysis dictate that the predictability of a regression equation will experience shrinkage when it is cross-validated or applied to a different sample from which it was derived (Stevens, 1986). This is a result of the maximization procedure of the least squares criterion and maximum likelihood approaches used in regression analyses. The extent of shrinkage is an indication of the equation's generalizability. An equation that experiences little shrinkage has greater utility. A second basic regression analysis issue is the subject to variable ratio. The larger the subject to variable ratio is the more stable the regression equation. The larger ratio enhances stability by minimizing or reducing error. Making judgments about a client's suicidality from a number of cues or making judgments about the potential of developing psychosis requires a large data base. If eight cues are part of the judgment process then at least 40 protocols are required (Wampold, 1987). In addition, before the statistical model can be used it should be cross-validated. Preferably, a different sample from which the regression equation was derived should be used to cross-validate. If substantial shrinkage occurs, then the clinician must start all over again. If shrinkage is minimal the clinician can go ahead and use the equation to make judgments. But, the clinician probably cannot market the equation to other facilities until greater cross-validation is achieved. Therefore, statistical equations that are generalizable to the everyday clinician (wherever he/she is) need to be based on hundreds to thousands of protocols and have been cross-validated in several different settings - this is an arduous task.

Dawes and Corrigan (1974) presented another limitation to statistical models. Statistical models require that cues be codable in some form so that they can be entered into the regression analysis. Clinicians may use a cue(s) that is not codable, therefore utilizing information not available to the model. It is unclear that the presence of one or two uncodable cues in the regression analysis would significantly increase the predictive power of a linear model based on codable cues only. Also, there may be cues that are codable, but need to be "experienced" by the clinician in order to be assessed. Such cues are equivalent to Sawyer's (1966) "clinical" mode of data collection. Interestingly, Sawyer (1966) found that this mode of data collection was inferior to the mechanical mode of data collection.

Lack of Acceptance:

In terms of reasons for the lack of acceptance of mathematical models regardless of methodological limitations, Meehl (1986) presents legitimate potential reasons, in a satirical manner, of why statistical models are not used in clinical practice. (a) "Sheer ignorance." There are countless clinicians of all persuasions who are not only unaware of the robustness of statistical predictions and the consistent finding over the course of 35 years, but who also do not know of this classical statistical vs clinical controversy. (b) "The threat of technological unemployment." That is, doctoral level clinicians take great pride in administering, interpreting and relating their interpretations of test scores (e.g., the Rorschach) and do not like to believe that a person trained in biometry could do at least an equivalent job making predictions. (c) "Theoretical identifications." The clinician who maintains a traditional orientation to psychotherapy (e.g., psychoanalytic) hates to admit that his/her theory permits very few predictions of importance, but nonetheless maintains his/her theoretical orientation. Admitting that statistical judgments outperform the judge would probably contribute to the clinician's theoretical insecurity. (d) "Dehumanizing flavor." Using an equation to make predictions about a human is dehumanizing, degrading, mechanical, and lifeless. (e) "Computer phobia." Many clinicians and social scientists have anxiety reactions, emotional blocks and cognitive blocks grappling with the idea that using computers can lead to results that exceed human performance.

Dawes (1971) addressed the issue that utilizing statistical models to make predictions about humans is dehumanizing. He countered by arguing that if the clinician or scientist is presented with 35 years of studies which consistently observe a positive and useful phenomenon in making judgments, and the clinician or scientist neglects these data and continues to make predictions based on the seat-of-his/her-pants, then that behavior is certainly irresponsible and unethical. Using reliable and valid statistical models to make judgments is a responsible, ethical and "human" action.

Thus, there are weaknesses and limitations with statistical models of judgments, but their lack of acceptance, based on personal "feelings" as opposed to scientific objectivity, is a significant contributory factor to their lack of use. So, this may, in part, explain their infrequent use, but does not justify it. There are many everyday clinical judgments that would be better made by a statistical model than a clinician. A clinician's time is usually at a premium. It would be advantageous for the clinician to have access to and use statistical models to make judgments when appropriate, therefore, saving time, and allowing more time performing other tasks (e.g., meeting with staff, doing psychotherapy) (Goldberg, 1970).

The previous sections have laid the groundwork about the clinical vs statistical controversy. Now more detailed information can be provided as to the specific conceptual and mathematical principles utilized in formulating the components used in clinical inference.

Brunswik's Lens Model

Brunswik (1955) introduced the Lens Model within the context of explicating a representative design and probabilistic theory in experimental psychology with particular emphasis in perceptual size constancy. Hammond, Hursch, and Todd (1964) applied Brunswik's Lens Model to the problem of clinical inference. Figure 1 diagrams Brunswik's Lens Model.

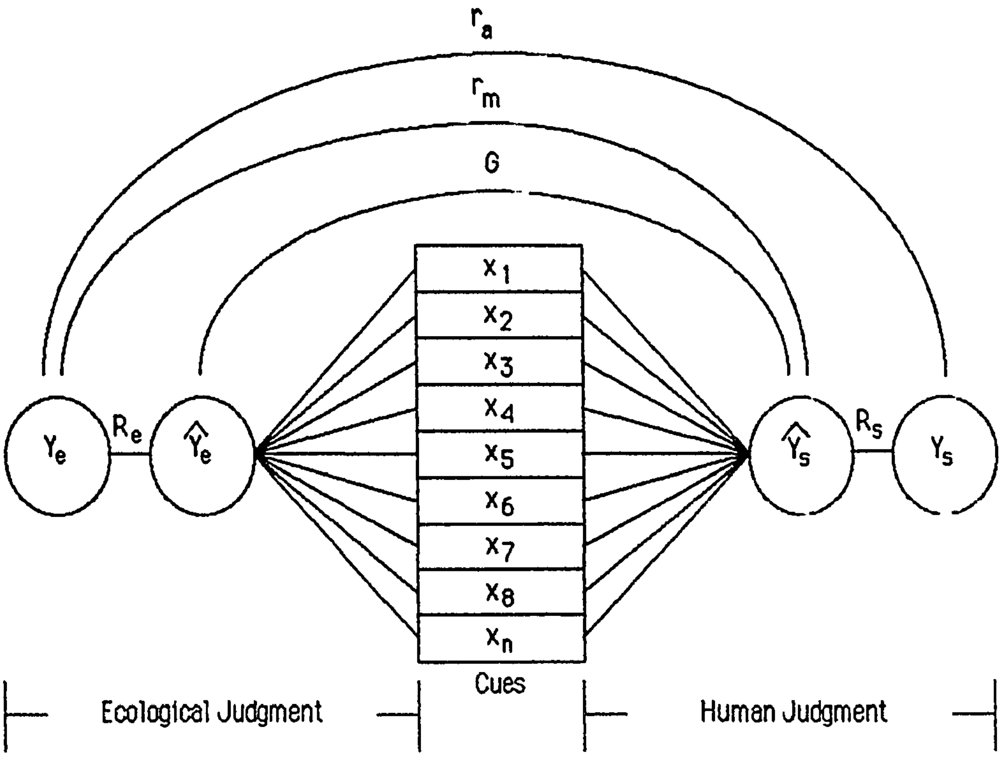

Figure 1. The Brunswik Lens Model. (Adapted from Hammond et al. 1964, p. 439).

A brief description of the diagram will be useful before a more detailed discussion ensues. The circles in the middle of the diagram (i.e., x1, x2,..) represent cues or predictor variables (e.g., test scores). The right side of the diagram is concerned with human judgment, and the left side relates ecological or environmental judgment. Y_S refers to the judgment(s) made by a judge, while Y_e refers to the actual criterion. The relationship between the cues and the human judgment (i.e., R_S) is known as the linear predictability of the judge. The relationship between the cues and the actual outcome or criterion (i.e., R_e) is known as the linear predictability of the criterion. The over arching line between the human judgment and the criterion is labelled the achievement index or validity coefficient of the judge. Human judgment research is concerned with how the judge uses cues to make a judgment (Y_s) (or, in other words, to make a prediction about the criterion, Y_e). Bootstrapping research represents a linear model ($\hat{Y}_s$), a mathematical abstraction, of the human judge which can be used to predict the criterion, Y_e. Actuarial research identifies a linear model of the ecology which can be used to predict the criterion, Y_e. Thus, bootstrapping research is concerned with generating a linear model of the judge ($\hat{Y}_s$), while actuarial research generates a linear model of the ecology ($\hat{Y}_e$).

Hursch, Hammond, and Hursch (1964) and Hammond et al. (1964) provided a detailed conceptual and mathematical formulation of the Brunswik Lens Model applied to clinical inference research. Hursch's et al. (1964) and Hammond's et al. (1964) mathematical proofs and resulting equation provided a benchmark for structuring a mathematical model to clinical inference. Tucker (1964) provided an alternative formulation of the Hammond et al. (1964) equation. Because Tucker's equation is somewhat more parsimonious and interpretable, it is typically utilized in statistical studies (e.g., Ebert & Kruse, 1978; Goldberg, 1970; Wiggins & Kohen, 1971). The reformulation is as follows (Goldberg, 1970, p. 424)

$$r_a = GR_eR_s + C\sqrt{1-R_e^2}\sqrt{1-R_s^2} \qquad [1]$$

where:

r_a: the achievement index or the validity coefficient of the judge: the correlation between the human judgment and the criterion ($r\ Y_s.Y_e$).

G: the linear component of judgmental accuracy: the correlation between the output from the linear model of the judge and the output from the linear model of the criterion ($r\ \hat{Y}_s.\ \hat{Y}_e$).

R_e: the linear predictability of the criterion: the multiple correlation between the cues and the criterion value ($r\ Y_e.\hat{Y}_e$).

R_s: the linear predictability of the judge: the multiple correlation between the cues and the judge's prediction ($r\ Y_s.\hat{Y}_s$).

C: the nonlinear component of judgmental accuracy: the correlation between the residual values of the criterion and the residual values of the judge's predictions after the linear components in both the criterion and the judge have been removed.

Goldberg (1970, p. 425) defines two more terms that are useful variables to be considered when performing actuarial studies:

r_m: the validity coefficient of the linear model of the judge: the correlation between the predicted scores from the judge's model and the actual criterion values ($r\hat{Y}_s.Y_e$).

Δ: the differential validity of model over the human judge: the difference in the validity coefficient between the model (r_m) and the human judge's achievement index (r_a).

Judgments, on both sides of the Brunswik Lens Model (i.e., ecological judgments and human judgments), may be conceptualized as being composed of three sources of variance: error variance, linear variance, and nonlinear variance. As is evident, the equation provides a number of indices that have direct implications in terms of assessing the degree of linearity and nonlinearity, as well as the accuracy of judgments. For example, if there is much nonlinear variance in the ecology and the mathematical model (assuming a linear model) of the clinician is unable to capture that variance, then the clinician should be more accurate. If there is mostly linear variance in the ecology, then the mathematical model of the judge will be more accurate to the extent that it eliminates error variance and nonlinear variance components from the clinician's judgment. In addition, the value of R_s has implications for the paramorphic process of the judgments regardless of the variance comprising the ecology. So that as R_s approaches its maximum value of 1.00, the clinician's judgments will become less distinguishable from the linear model. That is, as the clinician becomes more linearly predictable, the difference between the clinician and the model subsides.

Dawes (1974) has pointed out that if the difference between the human judgment and the linear model of the judgment is reliable, then the human judge is responding in a consistently nonlinear way. That is, the linear model is not accounting for all the systematic variance in the human judgment, therefore suggesting that the judge was utilizing nonlinear or configural processes. If the difference between the human judgment and the linear model of the judgment is not reliable, then the judge is responding linearly with an error component. That is, the judge is not consistently applying the linear porcess and, therefore, error must be intruding.

Next, two empirical studies will be examined that employed Tucker's equation. The various indices in the equation will be made explicit so that the reader will understand how differences between the human judge and the linear model are determined.

Bootstrapping Research

Bootstrapping studies are concerned with developing a linear model of the judge. It is not a purely statistical judgment method without any human interface in the judgment process. Instead, Kleinmuntz (1990) referred to bootstrapping as a combined use of judge's judgments and mathematical equations. First, a judge provides judgments, and second the judge is modeled, typically via regression analyses. The term bootstrapping was first used in a published paper by Dawes (1971), he and his colleagues at the Oregon Research Institute are credited with coining this term. Criterion information is not necessarily required to bootstrap. In relation to the Brunswik Lens Model, if criterion information is not available, $\hat{Y}_s$ can be compared to Y_s, and the value of R_s $(Y_s.\hat{Y}_s)$ can be assessed. A high value for R_s would mean that there was much linearity in the judgment, while a low value for R_s could mean that the human judgment was comprised of a large error component or the judge used more nonlinear or configural process than the linear model could account. When criterion information is unavailable no statements can be made about accuracy.

When criterion information is available more relationships and implications can be examined and the comparative accuracy of the human judge vs a linear model of the judge can be assessed. Specifically, the validity coefficient of the judge (i.e., r_a), the validity coefficient of the linear model of the judge (r_m), the linear model of judgmental accuracy (G), the nonlinear component of judgmental accuracy (C), the linear predictability of the judge (R_s), and the linear predictability of the criterion (R_e) can be computed.

Goldberg's (1970) study is frequently cited in the area of clinical vs statistical decision making. He examined the issue of judges vs linear model of the judge by re-analyzing Meehl's

(1959) data which involved judgments based on MMPI profiles. The cues consisted of the scale scores of the MMPI from 861 individuals from seven facilities who were categorized as psychotic or neurotic. The clinical judges were 29 clinical psychologists of varying levels of experience and training. The clinicians made judgments on a scale from least to most psychotic for all protocols within the seven samples. The clinician's judgments were used as the dependent variable or criterion. The 11 MMPI scales were used as independent variables or predictors. In addition, judgments from the 29 clinicians were combined and averaged to create a "composite judge."

The results showed that in five of the seven samples the "typical judge" (the typical judge was defined as the mean value of any index in Tucker's and/or Goldberg's equations across the 29 clinicians) produced a positive Δ value indicating that the linear model of the judge was more valid. Only one of the typical judges produced a value over the linear model. The linear component of judgmental accuracy (G) ranged from 0.24 to 0.77 with an average of 0.68, suggesting a high linear component. The nonlinear component of judgment accuracy (C) ranged from -0.16 to 0.19 with an average of 0.08. This means that only minimal nonlinear or configural processes contributed to the accuracy of the judgment.

The data from the composite judge (based on the average judgments of all 29 clinicians to each MMPI protocol) were similar to the typical judge. A positive Δ value was found in three of the seven samples (one of these was minimal, 0.003), and a negative value was obtained in four of the samples. In two of the samples where a negative value was obtained, the Δ value was negligible (-0.001 & -0.006). The overall Δ value was -0.017. This suggests that the composite judge and the linear model are about equal. The linear component of judgmental accuracy (G) ranged from 0.27 to 0.66, with an average of 0.72. The nonlinear component of judgmental accuracy (C) ranged from -0.28 to 0.33 with an average of 0.13. These latter two pieces of data indicate that the linear model accounts for the vast majority of judgmental accuracy, and only a small amount of the accuracy of judgement is from a nonlinear component.

In all indices of the equation, the composite judge outperformed or performed as least as well as the typical judge. It was found that the largest differences between the composite judge and the typical judge occurred on the R_s index. On this index, the composite judge outperformed the typical judge suggesting that the averaging technique used in constructing the composite judge removed the unreliability in the typical judge.

The achievement index (a.k.a., validity coefficient) of the most accurate human judge (r_a), across the seven samples, ranged from 0.32 to 0.56 with an average of 0.39. The achievement index of the most accurate linear model (r_m), across the seven samples, ranged from 0.32 to 0.60 with an average of 0.43. Therefore, the validity coefficients of the linear model of the judge were higher than the validity coefficients of the human judge. The validity coefficient of the typical judge was outperformed by the validity coefficient of the typical linear model (0.28 and 0.30, respectively). Interestingly, the validity coefficient of the composite human judge was higher than the validity coefficient of the composite linear model (0.35 and 0.33, respectively).

The rank order of the indices resulting in the most accurate judgment is as follows (Table 2).

Table 2

The Rank Ordering of the Validity Coefficients

The linear predictability of the criterion (R_e)	0.46
Actuarial formula	0.44
Most accurate model	0.43
Most accurate judge	0.39
Composite judge	0.35
Model of composite judge	0.33
Typical model	0.31
Typical judge	0.28
Least accurate model	0.16
Least accurate judge	0.14

The robustness of these findings was tested through a cross-validation procedure. The original sample of 861 MMPI profiles were reorganized in several different samples. For example, Goldberg examined the difference in the achievement indices between the judge and the linear model of the judge when smaller samples were employed to construct the linear model. When the linear model of the judge was constructed on one-half, one-seventh, and one-tenth of the original sample, the linear model outperformed the judge (the validity coefficient of the judge was based on the remaining portion of the original sample) in 86%, 79% and 72% of the comparisons, respectively.

A number of conclusions can be drawn from these data. (a) When criterion information is available and the validity coefficients of the judge and the linear model of the judge can be compared, the linear model of the judge consistently outperformed the judge. This was made clear by the rank ordering of the validity coefficients presented in Table 2. (b) When the generalizability of the findings are examined via cross-validation the results are generally maintained. (c) The composite judge is slightly more accurate than the composite linear model of the judge and the typical linear model of the judge. Therefore, combining and averaging judgments results in a level of accuracy that cannot be improved by a linear modelling technique. The prominence of a composite judge is not surprising and has been found in other studies (Wedding, 1983), but it is not a guarantee (Wiggins & Kohen, 1971). Although the composite judge has been found to be superior in some studies, its practicality is questionable. That is, the pooling of clinicians in everyday clinical practice to make judgments is extremely inefficient and costly (Goldberg, 1970).

Wiggins and Kohen (1971) examined the accuracy of predicting graduate students grade point average (GPA) from a standardized set of cues. Ninety-eight graduate psychology students at the University of Illinois volunteered to participate in the study. The sample represented all four years of student status and each was paid for his/her participation.

The graduate students were asked to predict the GPA of first year psychology students from the years 1965-1968. They based their judgments on ten cues: (a) GRE-Verbal; (b) GRE-Quantitative; (c) GRE-Advanced; (d) cumulative undergraduate GPA for the last two years of college; (e) ratings of the selectively of the undergraduate school; mean peer ratings received on a 5-point scale for need (f) Achievement, (g) Extraversion, and (h) Anxiety; (i) self-rating on conscientiousness, and (j) gender of student. The first five cues were, in fact, the cues used in the selection process of graduate applicants. The judges were provided with norms and averages for the cues when available, and asked to make prediction on 110 protocols (90 originals and 20 repeated protocols). The predictions were on an eleven point scale, ranging from 3.0 ("C") to 5.0 ("A") in increments of 0.2.

The results showed that the validity coefficient of the judge (r_a) was 0.33 with a range of 0.07 to 0.48. The mean validity coefficient of the linear model of the judge (r_m) was 0.50 with a range of 0.10 to 0.64. Therefore, the accuracy of the linear model was superior to the judge. The higher validity (r_m) found in Wiggins and Kohen's study as compared to Goldberg's (1970) study is, in part, a result of the higher value of R_e in Wiggins and Kohn's study. That is, Wiggins and Kohen obtained a value of 0.69 for R_e, while Goldberg obtained a value of 0.46. The much higher value of R_e in the Wiggins and Kohen's study indicates that there is a higher linear relationship between the cues and the actual criterion which is obviously best captured when a linear judgment process is employed. (The difference between the validity coefficients of the linear model (r_a) and the judge (r_m) will be minimized as the value of R_s approaches 1.00.) Wiggins and Kohen also constructed a composite judge. The mean accuracy of prediction of the composite judge was 0.47. This index was superior to the prediction made by the typical judge (0.33). The mean linear model of the composite judge was 0.58. This value was notably higher than the composite judge value (0.47), and unlike Goldberg's finding (again, Goldberg found a mean value of 0.35 for the composite judge, and a mean value of 0.33 for the linear model of the composite judge). This discrepancy suggests that the combining and averaging of predictions to form the composite judge did not result in reducing the unreliability of the individual judges.

The rank ordering of the indices resulting in the greatest accuracy (i.e., validity coefficients) was as follows (Table 3).

Table 3

The Rank Ordering of the Validity Coefficients.

Linear predictability of criterion	0.69
Most accurate model	0.64
Model of composite judge	0.58
Typical model	0.50
Most accurate judge	0.48
Composite judge	0.47
Typical judge	0.33

Thus, the data in this study are more striking than those obtained in Goldberg's study. Wiggins and Kohen found that the most accurate linear model outperformed the most accurate judge, the linear model of the composite judge outperformed the composite judge, and the linear model of the typical judge outperformed the typical judge.

The regression coefficients used in the above studies were optimal weights achieved through the mathematical operations of the regression analysis. A curious finding, that is not often addressed in applied research in this area or in papers that address general issues of clinical vs statistical judgment, is that substituting equal weight coefficients for optimal weights achieves the same results, and in some cases the equal weights outperform the optimal weights. This issue will be explored next.

Optimal Weighting Coefficients Versus Unit/Equal Weighting Coefficients in Linear Models

As has been outlined earlier, bootstrapping models are typically superior to clinical judgments because, in part, the linear mathematical model [being an abstraction of the clinical judgment process] is perfectly reliable and disregards the often spuriouness of the nonlinear or configural processes used by the clinician (e.g., Goldberg, 1965, found that the semi-partial correlation between the human judge and the criterion partialling out the variance of the linear model from the human judge leaves an association between these two variables at about 0.05). Assuming that the clinical judge is following valid principles in the decision making process, but follows them inaccurately, the mathematical model will abstract the valid principles and eliminate the inaccuracies (Dawes & Corrigan, 1974). The optimal weights achieved in these mathematical models is the reason the statistical approach outperforms the clinician. Or is it? What if unit/equal weighting was employed? Unit/equal weighting can be defined as beta weights that are a priori chosen by the researcher to be used in the regression equation based on theory and not on conventional least squares and maximum likelihood approaches. (In this section, the words "unit" and "equal" are used interchangeably.)

It certainly is an empirical question as to whether beta coefficients achieved through an optimal linear combination of the variables produce a more predictive equation than unit weighting coefficients. Dawes and Corrigan (1974) carefully examined this issue and concluded that the unit weighting scheme was equal to and often superior to optimal weighting coefficients (Table 4). They relate two reasons for this seemingly peculiarity. (a) In many studies, there are too many predictor variables and too few samples resulting in unstable beta weights. (b) In addition to the three reasons why linear models perform so well (see page 9), they are also robust to deviations from optimal weighting coefficients. That is, weights that nearly approximate optimal weights produce about the same effects; and actually, Dawes and Corrigan (1974) have shown that unit weights are often superior to optimal weights. It is important to note that Dawes and Corrigan did not make a blanket statement indicating that in all cases unit weighting will be superior to optimal weighting. But, they provided unequivocal evidence that in many cases unit weighting can be equal to or outperform optimal weighting procedures.

Einhorn and Horgarth (1975) also addressed the issue of unit/equal weighting in linear models. They provided four reasons for considering unit weighting schemes. (a) The important issue may not at all be the problem of what type of weighting to employ, but specifying the most predictive variables into the model to begin with. That is, once the most predictive variables are in the model and the less predictive variables are excluded, the weighting scheme may not be especially relevant. (b) The function form (e.g., linear, curvilinear) of the regression equation may be more important than the weighting scheme used. (c) In the production of beta coefficients through the optimal linear combination of the variables, there will always be some amount of sampling error. Therefore, the resultant weights are produced within the context of sampling error. The use of unit weights (e.g., equal weighting) contains no sampling error. Thus, a trade off ensues between estimation of accuracy vs estimation without error. Einhorn and Horgarth (1975) argue that since most real data includes both sampling and measurement error, the apparent superiority of standard regression procedures over unit weighting schemes may be unfounded. (d) Einhorn and Horgarth (1975) cited several empirical studies that have shown unit weighting schemes to be equal to (i.e., as predictive) standard regression procedures. An important presumption to these four factors is that the sign of the unit weight can be made a priori. But, this is not usually a concern, because one can discern the sign based on the hypothesized product-moment correlation between the predictor and the criterion.

Table 4

Correlations Between Predictions and Criteria Values

Example	Average validity of judge	Average validity of judge's model	Validity of equal weighting model
Prediction of neurosis vs psychosis	.28	.31	.34
Illinois students' prediction of GPA	.33	.50	.60
Oregon students' prediction of GPA	.37	.43	.60
Prediction of later faculty ratings at Oregon	.19	.25	.48
Yntema & Torgerson experiment	.84	.89	.97

Einhorn and Horgarth (1975) present an equation to determine the superiority of weights obtained through standard regression procedures vs a unit weighting scheme when both are applied to the same set of data. The superiority of one model over the other will depend on the number of predictor variables employed, the sample size, the clarity and accuracy in which the criterion is defined, and the degree of intercorrelations among the predictor variables. For example, if six predictor variables are employed and the sample size is 30, then the unit weighting scheme will probably be at an advantage, because the unit weighting scheme is not affected by the subject to variable ratio (it is not influenced by the peculiarities of the data) as is the optimal weighting scheme. In addition, if a relatively large number of predictor variables are included in a regression analysis, then the unit weighting scheme also may be at an advantage because of the problem of multicollinearity. In psychological research, there is often modest to substantial intercorrelations (multicollinearity) among predictor variables. The more predictor variables incorporated into the regression analysis the greater the chance that some important variable(s) may not receive a corresponding high beta coefficient because of the influence of multicollinearity in the set of predictor variables. In a unit weighting scheme, the beta coefficients are chosen a priori. Therefore, hypothesized important predictor variables will receive an appropriate beta coefficient uninfluenced by the intercorrelations of the other predictor variables. Standard regression analysis (i.e., optimal weighting) will probably be superior to a unit weighting scheme when there is a large subject to variable ratio and the measurement of the criterion is highly reliable (see Einhorn & Horgarth, 1975 for an extended discussion). In addition, if the sign (i.e., positive or negative) of a weight(s) cannot be determined for a cue(s) a priori, then optimal weighting may be better. Although, if the sign cannot be determined for a cue(s), then it appears questionable as to why the cue(s) is being employed (Camerer, 1981). That is, if the researcher cannot theoretically or logically assess the sign of a cue(s), then the contribution of the cue is suspect and may negatively interfere (e.g., increase multicollinearity) in the resulting analysis.

Also, Einhorn (1986) briefly addressed the issue of equal weighting. He purported that using equal weights deliberately introduces error into the model. The error may offset optimal weights, achieved in a standard regression analysis, that are the result of poor data (e.g., a low subject to variable ratio). Einhorn provided a simple example, if predictor variables x1 and x2 have a true relative weighting of 2:1, then using equal weights in the regression analysis prevents the standard regression analysis from producing a weight for x2 that is greater than that of x1 when a poor data sample is used. Therefore, introducing a known error may prevent spurious error.

Unit weighting schemes are not only a technical concern, but an important theoretical concern (Dawes & Corrigan,1974; Einhorn & Horgarth,1975; Camerer,1981; Wiggins,1981). The implication is that the unit weighting scheme allows for a more parsimonious prediction model. The extreme view is that there is no need to go through the tedious process of developing a linear model of the judge, but simply weight the cues accordingly and the resulting predictions will be at least as accurate as the judge. In addition, simple implementation of equal weighting in regression equations do not have the potential contaminations in standard regression analysis (e.g., multicollinearity) (Wiggins, 1981). Also, unit weighting schemes make cross-validation of the regression model less critical (Wiggins, 1981).

The conceptual and mathematical issues of bootstrapping research have been examined. Empirical studies of bootstrapping research were presented to make the conceptual and mathematics issue more understandable and concrete. Now it is time to bridge the areas of bootstrapping research with that of a relatively new specialty in psychology, i.e., neuropsychology.

Clinical Decision Making in Neuropsychology

Clinical neuropsychology is a recent specialty in the field of psychology. Fundamentally, neuropsychology is concerned with the study of brain-behavior relationships (Horton & Puente, 1986). Neuropsychology evolved from multiple influences, research areas and disciplines during the late nineteenth and early twentieth centuries. Specifically, Hartman (1991) advocates that neuropsychology emerged from the contributions of the mental testing movement, experimental and clinical psychology, medicine and neurology.

A few of the major figures in the history of neuropsychology over the past 130 years include: Paul Broca whose work in the 1860's and 1870's discovered that lesions in a specific area of the left frontal lobe resulted in difficulties in expressive speech (i.e., nonfluent aphasia) (Horton & Puente, 1986). John Hughlings Jackson, father of British neurology, made significant contributions in the mid- to late nineteenth century in the areas of epilepsy, aphasia, and the understanding of the central nervous system (Zangwill, 1987). Pierre Flourens, a French physiologist, advanced the techniques of ablation in the understanding of brain functioning in the mid nineteenth century. In experimental psychology, the works of Shepherd Ivery Franz (ablation, frontal lobe studies), Carl Lashley (equipotentiality, law of mass action), Roger Sperry (split-brain studies), Donald Hebb (cell assembly), and Karl Pribram (cortical functioning, memory) have made direct or indirect contributions to experimental and/or clinical neuropsychology in the early to mid- twentieth century. For a more in-depth account of the history of neuropsychology, Hartman (1991) provides a scholarly and comprehensive narrative.

Its formal clinical development in the United States can be traced to the World War II period (Matarazzo, 1972). Pioneers at this time included Arthur Benton at the University of Iowa, Kurt Goldstein, Ward Halstead at the University of Chicago, A. R. Luria in Russia, Brenda Milner, Ralph Reitan, and Hans-Lukas Teuber (Horton & Puente, 1986; Hamsher, 1984).

The histories of clinical versus statistical judgments and the developments of clinical neuropsychology closely approximate one another. Meehl's (1954) book is credited with igniting the fervor of clinical versus statistical judgments, and as stated above, clinical neuropsychology in the United States began around the World War II era. Meehl's book lead to an enormous number of published articles in regard to clinical judgment, while, in neuropsychology, only a handful of clinical judgment studies have been published. Three of the clinical judgments studies will be reviewed

Goldstein, Deysach, and Kleinknecht (1973) examined the accuracy of judgments of experienced clinicians, inexperienced clinicians, and the Impairment Index of the Halstead-Reitan Battery in the determination of cerebral impairment. Five clinicians (four were Board certified in clinical psychology by the American Board of Professional Psychology) with nine to eighteen years of experience comprised the experienced group. The inexperienced group consisted of five doctoral students, three students were completing their internship and two were in the advanced years of their program.

The clinicians were asked to make a judgment about the presence or absence of cerebral impairment in two groups of patients. One group was composed of ten patients with unequivocal evidence of brain impairment, while the second group of ten patients were evaluated to show no organic impairments. The groups were matched for gender, age, handedness, occupation, and education.

All judges were presented with the same set of data on each patient. The data consisted of test scores from the Halstead-Reitan Battery, WAIS, MMPI, and Bender Gestalt Test.

Three of the five experienced clinicians were not trained in the Halstead-Reitan Battery and, therefore, could not utilize these data in making judgments. The inexperienced clinicians were provided with 15 hr of training in the Halstead-Reitan Battery. Because of this discrepancy, the two sets of judges were given the 20 protocols containing data from the WAIS, MMPI and Bender. After the initial classification, the inexperienced clinicians were given data from the Halstead Reitan Battery and asked to classify the protocols again. A cut-off level of 0.4 on the Impairment Index was used to demarcate presence from absence of cerebral impairment.

The results showed that there was no significant difference on judgments between the two sets of clinicians on data from the traditional battery (i.e., WAIS, MMPI, and Bender). The Impairment Index was significantly more accurate in judgment than the experienced clinicians, and better, although not significantly, than the inexperienced clinicians utilizing the traditional battery. When the inexperienced clinicians were provided with the Halstead-Reitan Battery data they greatly improved their judgments. In fact, they bettered the judgment made by the

Impairment Index, although not significantly.

Wedding (1983) compared Russell's taxonomic key approach (actuarial), discriminant function analysis, and clinical judgments to classify five diagnostic groups. The clinical judges included ten practicing doctoral psychologists, three pre-doctoral interns, and one expert neuropsychologist. They averaged 12.6 years of post-doctoral experience (range=0-35 years), and had interpreted 20 to 900 Halstead-Reitan batteries. A number of Halstead-Reitan neuropsychological protocols were selected for the classification procedure. Protocols were classified into left hemisphere damage, right hemisphere damage, and diffuse damage. Also, Halstead-Reitan records from individuals with schizophrenia without medical documentation indicative of brain damage were included as the fourth group. Protocols from neurologically intact individuals were collected in the fifth group.

Discriminant functions analyses were used to classify the individuals into the five groups. In addition, prediction was made as to the etiology (vascular, neoplastic, traumatic, or degenerative), and chronicity (greater than or less than one year). Finally, all classifications were made under two levels of information. The high level of information condition included the individuals age, gender, handedness, education, and all Halstead- Reitan and WAIS summary scores. The low level of information condition contained data concerning the individuals age, gender, handedness, education, scores of Trails A and B, Block Design and Digit Symbol scores for the WAIS, number of errors on Speech sounds and Rhythm and the number of errors on the Aphasia Screening Exam. The discriminant functions were cross-validated (i.e., tested) on a random selection of six individuals from each of the five groups (i.e., the 30 cases that the judgments were based on).

Following the discriminant function analysis, the data on the cross-validated sample were analyzed by Russell's taxonomic key approach. Since Russell's key approach was not designed to predict psychiatric status, schizophrenics classified as non-brain damaged were considered correctly classified.

Finally, two regression equations (for the high and low information conditions) were developed from the larger data pool (as opposed to the cross-validation sample) to predict performance on the Wechsler Memory Scale.

Clinical judges were asked to classify the 30 randomly sampled protocols into one of the five groups, predict etiology (vascular, neoplastic, traumatic, or degenerative), predict chronicity (greater than or less than one year), and estimate the individual's Wechsler Memory Quotient. Judges were informed about the characteristics of the sampled protocols and given base rate information (e.g., six protocols were from individuals with schizophrenia, evenly divided between the high and low information conditions). Finally, judges recorded the amount of time spent performing these judgments, and estimated their judgment confidence.

Overall, the Russell key approach accurately classified 60% of the records, the discriminant function analysis accurately classified 63% of the records. Two clinical judges outperformed the statistical approaches (each at a rate of 70%), and one tied the discriminant function's level. All other clinical judges performed more poorly (range=33% to 57%). The accuracy of the clinical judgments were not significantly related to the amount of time spent on the judgments, clinical experience, or experience with the Halstead-Reitan Battery. In addition, there was no significant relationship between a clinician's confidence and his/her judgments about localization, etiology, and chronicity.

Clinical judges expressed greater confidence in their decisions made under the high information conditions than in the low information condition. Unfortunately, they were more likely to be inaccurate in this condition, while the discriminant function analysis improved by 7%. The clinical judges and the discriminant function analysis were equal in the low information condition.

In terms of estimating the Wechsler Memory Quotient, the two regression equations outperformed all of the clinical judges.

Recently, Faust et al. (1988) examined how training and experience in neuropsychology affected judgment accuracy. Neuropsychologists were solicited who were listed as a diplomate in Clinical Neuropsychology, a member of Division 40 of the APA, indicated clinical neuropsychology as a major field or area of specialization in the American Psychological Association directory, or indicated neuropsychology as a specialized health service in the National Register of Health Service Providers in Psychology. A random sample of 600 were chosen from the larger population pool. The 600 were randomly divided into 10 groups of 60, and each group received one of the 10 judgment cases.

The 10 protocols contained eight abnormal cases and 2 normal cases. The abnormal cases were chosen to be representative of common neurological disorders. The neuropsychological measures included scaled scores from the WAIS-R, all Halstead-Reitan Battery measures, portions of the Wechsler Memory Scale (i.e., semantic and figural memory for both immediate and delayed recall), and demographic information (i.e., age, education, employment, gender, and handedness). The requested judgments involved: (a) presence vs absence of brain impairment, (b) static vs progressive disorder, (c) area of cortex involved (localization general and exact), and (d) etiology.

Neuropsychology respondents also completed information concerning their background, training, and experience. Specifically, background questions included: (a) years practicing neuropsychology, (b) amount of pre-degree experience in neuropsychology, (c) percentage of predoctoral internship time in neuropsychology, (d) completion of a postdoctoralship, (e) number of formal neuropsychology courses completed, (f) percentage of professional time in neuropsychology, (g) presence or absence of publication in neuropsychology, and (h) a question as to whether or not the Halstead-Reitan Battery is the preferred assessment instrument.

The results showed that the overall accuracy of distinguishing presence vs absence of brain impairment was 80% (which is the base rate), a 60% accuracy rate of distinguishing static from progressive conditions, and an accuracy rate of 54% for general localization and a 29% accuracy rate of exact localization. Overall the background factors were found to be unrelated to type of judgments. Only two of the 48 correlations were found to be significant, but accounted for minimal variance (trainee experience and exact location = .21; years practicing and general location = -.22). The authors next separated the judges on various background variables to create more extreme groups in which to evaluate the issues of training and experience on judgment accuracy. But, overall the same conclusion was found. That is, training and experience had essentially no significant effect on accuracy of judgments. It would be erroneous to conclude that all clinicians are equal regardless of training and experience, but, when background data is grouped, training and experience do not significantly mediate accuracy of judgment.

The findings from the three clinical decision making studies in neuropsychology presented above are consistent with previous research in the area of clinical psychology (Dawes et al., 1988; Garb, 1989; Wedding & Faust, 1989). Specifically, training and experience do not significantly differentiate accuracy of judgments (Faust et al., 1988; Goldstein et al., 1973; Wedding, 1983), and statistical judgment models are superior to clinical judgments (Goldstein et al., 1973; Wedding, 1983).

Neuropsychological assessment lends itself very easily to the bootstrapping model. That is, in neuropsychological assessment multiple tests are used (providing multiple cues or predictors to form the bases of judgments) to make judgments about presence vs absence of brain impairment and location of impairment (these criteria are easily verifiable by brain imaging techniques).

Hypotheses

Neuropsychologists were asked to make judgments about the presence vs absence of brain damage and localization of brain damage (right, left, or diffuse) based on 20 to 29 cues (i.e., selected neuropsychological tests and demographic information). A linear model of each neuropsychologist's judgments was developed via regression equations involving both optimal weights and unit weights. Because of the subject to variable ratio issue in multiple regression analysis, the linear model was based on a subset of the cues. Specifically, the linear model was based on 9 cues (this is addressed further in the Method section). Thus, each judge's validity coefficient of judgmental accuracy was based on all the available cues, while the linear model of the judge's validity coefficient was based on nine cues. It was hypothesized that the linear model of the judge, regardless of the type of weight employed, will be equal to or outperform the judge. Also, it was hypothesized that the most accurate linear model, regardless of the type of weight employed, will be equal to or outperform the most accurate neuropsychologist's judgments.

Also, the extent of judges' experience was examined. Specifically, judges were classified as novice or expert based on criteria described in the Method section. It was hypothesized that there would be no notable or meaningful differences in the accuracy of judgments (i.e., hit rate) between novice and expert judges nor in their validity coefficients (i.e., r_a and r_m).

METHOD

Judges

Six neuropsychologists (i.e., judges) participated in the study: three experts and three novices. The criteria of expert and novice reflect the recent educational and training guidelines publsihed by joint INS/Division 40 (Report of Task Force, 1984; 1986) and Divison 40 (Report of The Executive Committee, 1989) Task Force Committies on Education, Accreditation and Credentialing. Specifically, experts consisted of three psychologists who attained the diplomate status (i.e., ABPP/ABCN) in clinical neuropsychology. In addition, experts #1, #2 and #3 have 13, 30 and 14 years of experience, respectively, in neuropsychology. Also, experts #1 and # 3 indicated that they completed a formal post-doctoral fellowship/program in clinical neuropsychology.

The criteria for novice status consisted of completion or partial completion of a post-doctoral program in neuropsychology and less than 3 years of full time experience as a neuropsychologist. In this study, two of the novices were in the process of completing a formal post-doctoral fellowship/program in clinical neuropsychology under the supervision of a diplomate in clinical neuropsychology. The third novice completed a formal post-doctoral fellowship/program in clinical neuropsychology about 2 years previously under the supervision of a diplomate in clinical neuropsychology. Novices #1, #2 and #3 reported that they have less than 1, 2.75 and 1.5 years of experience, respectively, in clinical neuropsychology. All judges were paid $100.00 for participating.

Judgments and Cues

Judges were asked to make up to two decisions on each neuropsychological protocol on the basis of the cues (i.e., test scores and demographic information) provided. The judgments were: (a) The presence vs absence of brain damage, and (b) the localization of brain damage (i.e., right hemisphere, left hemisphere or diffuse). If the protocol was judged as indicating the absence of brain damage, no judgment was made as to localization. These two judgments have been used in previous clinical decision making research (Faust et al., 1988; Wedding, 1983). It is important to point out that these judgments are historically the major clinical decisions made by neuropsychologists, but not so in recent years (Chelune & Moehle, 1986). Most recently neuropsychologists tend to use their skills to make more complex judgments, e.g., can this person return to his/her former occupation, can this person return to independent living following rehabilitation, how can rehabilitation be structured to maximize the patient improving (Chelune & Moehle, 1986). The reason these latter judgments were not used as the criteria was because of the great difficulty in objectively and operationally defining them as well as coding them into a regression analysis.

Professional time constraints limited neuropsychological judges to making judgments on no more than 50 protocols. The judges predictions were based on up to 29 cues. Because the linear model was produced using a regression analysis, the subject (or in this case - protocol) to variable (or in this case - cue) ratio constrained the number of cues chosen to construct the linear model of the judge. That is, in order to produce a reliable equation, Stevens (1986) suggests a subject to variable ratio of 15:1 and Nunnally (1978) suggests a 5 - 10:1 ratio. Although the subject to variable ratio is important, such simple rules of thumb have only limited utility (Wampold & Freund, 1987), and ratios as low as 5:1 are not unreasonable. In general, a subject to variable of 5:1 is probably the lowest ratio allowable to produce a stable regression equation.

A maximum number of 50 protocols suggest that at most 10 cues could be used. Given that there is probably a modest intercorrelation among many neuropsychological tests, including 10 cues (predictors) in a regression equation would probably produce a multiple correlation coefficient that would not increase substantially even if more cues were used. In fact, using fewer cues (e.g., six cues) may well have the same predictive value as an equation based on 10 cues. Having redundant cues tends to add little, if anything, to the size of the multiple correlation (R). In addition, in many cases the squared multiple correlation has most of it's variance accounted for by a smaller number of cues than those actually included in the study. For example, in a study that employed 10 cues, most of the variance in the squared multiple correlations would likely be explained by only 4 or 5 cues.

Thus, because of time constraints, and, given that there are probably modest intercorrelations among neuropsychological tests and that the value of the multiple correlation will not significantly change just because more cues are provided, it was decided to approximate a 5:1 ratio. Given a maximum of 50 protocols on which to make judgments, the number of cues should be 10.

The neuropsychological tests (i.e., cues) used to comprise each judgment protocol were selected for the following reasons: (a) Tests that are commonly employed and understood by neuropsychologists were included. (b) Tests that are purported to be sensitive measures of brain impairment were included. (c) Tests which aid in localization were included. (d) Cues which tap different neuropsychological functions (e.g., memory, motor, visual- spatial, language) were included.

Judges were provided with up to 29 cues (see Appendix A). The cues consisted of scores from: WAIS-R (Verbal IQ, Performance IQ, Full Scale IQ and age equivalents scaled scores from the 11 subtests); Category test (number of errors); Wisconsin Card Sort test (number of categories completed); immediate and delayed recall trials Logical Memory subtest, Wechsler Memory Scale-Revised; immediate and delayed recall trials Visual Reproduction subtest, Wechsler Memory Scale-Revised; Trail Making Test, Parts A and B; Controlled Word Association Test (FAS); Finger Tapping Test (right and left hands); occupation status; age; education; and gender. The majority of the tests selected to be included in the protocols are among the 11 neuropsychological instruments most frequently employed by practicing neuropsychologists (Guilmette, Faust, Hart, & Arkes, 1990. See also Butler, Retzlaff, & Vanderploeg, 1991).

Neuropsychological cues used in the regression equation to construct the linear model of the judge were chosen on the following basis: (a) Tests that are purported to be sensitive measures of brain impairment were included. (b) Tests which aid in localization were included.

(c) Cues which tap different neuropsychological functions (e.g., memory, motor, visual-spatial, language) were included. The following tests comprised the 9 cues used to generate the linear model:

The Controlled Oral Word Association Test (i.e., FAS Test) (Borkowski, Benton, & Spreen, 1967). This task required that the individual rapidly generate words beginning with the letters F, A & S (others letters are used in alternative tests, cf. Benton & Hamsher, 1978). It has been shown to be sensitive to left hemisphere functioning, left frontal lobe functioning, and generally to be a sensitive measure of global brain functioning (Benton, 1968; Miceli, Caltagirone, Gainotti, Masullo, & Silveri, 1981; Parks, Loewenstein, Dodrill, Barker, Yoshii, Chang, Emran, Apicella, Sheramata, & Duara, 1988).

Block Design subtest - (age equivalent scaled score). (Wechsler Adult Intelligence Scale-Revised, (WAIS-R); Wechsler, 1981). This task measures visual-spatial skills in the reproduction of abstract designs. It is considered the best indicator of visual- spatial functioning among the WAIS R subtests (Lezak, 1983). Block design correlates most highly with Performance IQ. Impaired scores on this subtest tend to be associated with right hemisphere dysfunction (Black & Strub, 1976).

Similarities subtest - (age equivalent scaled score). (WAIS R; Wechsler, 1981). This task is concerned with verbal concept formation. It tends to be the most sensitive of the Verbal subtest to brain dysfunction (Lezak, 1983). It tends to be sensitive to left hemisphere injury, especially involving the anterior left hemisphere (McFie, 1975).

Digit Symbol subtest - (age equivalent scaled score). (WAIS-R; Wechsler, 1981). This task measures motor speed, sustained attention, and symbol learning. It has been found to be the most sensitive subtest to cortical dysfunction of the WAIS (Hirschenfang, 1960). It is a non-localizing task (Lezak, 1983).

Finger-Tapping Test - (average number of taps per 10 sec - score for the right and left hands). (Halstead, 1947. Part of the Halstead-Reitan Battery; Reitan & Davison, 1974). This task measures fine motor speed and control with the hands. Tapping speed tends to decrease when there is brain impairment (Lezak, 1983). Because the test is performed by both hands, there is the potential for contralateral differences to emerge which have implications for lateralization of brain damage (Finlayson & Reitan, 1980).

Trail Making Test, Part B - (time to complete test)(U.S. Army Individual Test Battery. Reitan, 1955; 1958. Part of the Halstead-Reitan Test Battery; Reitan & Davison, 1974). This task measures cognitive flexibility and the ability to execute a sequential plan. It is considered one of the most sensitive tests of brain functioning (Lewinsohn, 1973).

Logical Memory Subtest, delayed trial- (total number of details recalled for both stories) (Wechsler Memory Scale-Revised; Wechsler, 1987). This task measures the ability to recall verbal material in short paragraph form. Typically, immediate and delayed (30 min) recall trials are given. The test has been found to be sensitive to left hemisphere functioning during the delayed trial, and not to be especially lateralizing during the immediate recall trial (Delaney, 1980).

Visual Reproduction Subtest, delayed trial - (total score for all figures). (Wechsler Memory Scale-Revised; Wechsler, 1987). This task measures the ability to reproduce simple designs from memory. There are immediate and delayed recall trials. As with the logical memory subtest, the Visual Reproduction subtest seems to aid in lateralizing brain damage during the delayed recall trial as opposed to the immediate recall trial (Delaney, 1980). Impaired delayed recall is usually associated with right hemisphere dysfunction.

There are three reasons that the protocols contained a different number of cues (range 20 to 29). (a) Many neuropsychologists follow an individualized approach to neuropsychological assessment. That is, depending upon the particular circumstance of the client, selected tests will be administered. Therefore, not all neuropsychological assessments have the exact same tests administered and same number of cues. (b) Selecting protocols with a range of cues significantly aided the researcher in filtering through potential records to employ in this study. That is, records that were appropriate were not discarded just because one, two or three tests were not administered. (c) Twenty to 30 cues approximates the amount of data incorporated in many neuropsychological assessments.

Protocol Selection

Neuropsychological records from right-handed adults (18 to 65 years) were used. Only right handers were included because the great majority of people who are right-handed have speech dominance primarily lateralized in the left hemisphere and have nonverbal, visual-perceptual and spatial functions primarily lateralized in the right hemisphere (Kolb & Whishaw, 1990). Therefore, in order to maximize the probability of the optimal and unit weight regression coefficients (beta weights) associated with each cue to follow the theoretical division of tests or functions primarily associated with the left and right hemispheres, only right-handers were used. In addition, the four sets of protocols to be described below (i.e., normals, right hemisphere, left hemisphere, and diffuse brain damage) were matched, as closely as possible, for age and educational attainments. Matching was used to prevent judges from differentiating the protocols simply on the basis of systematic differences in the age and education cues across the four sets of protocols. For example, diffuse brain injuries (e.g., traumatic brain damage) usually involve young people while brain injuries from a cerebral vascular accident tend to be associated with older persons.

Forty of the 50 neuropsychological protocols were from brain-injured individuals. Thirty-eight protocols were obtained by reviewing records from a neuropsychology laboratory at a major university hospital center in the mid-west. Two additional protocols were obtained from a hospital in western Pennsylvania for a total of 40 brain damaged protocols.

Ten of the 50 protocols were from "normal" individuals. These individuals were recruited from the Volunteer Services Department in two hospital settings in western Pennsylvania. These individuals did not have a self-reported history of head injury, neurological disease (e.g., epilepsy, strokes), major psychiatric disorders (e.g, organic mental disorders, psychotic disorders), learning disabilities or drug and alcohol abuse. All ten were right-hand dominant, had at least 12 years of education and were not paid for participating. Testing of the controls was completed by this author and an advanced doctoral student.

Three of the four groups of protocols, consisting of a total of 40 neuropsychological records, were from people who sustained a brain injury. Group one consisted of ten protocols from individuals who had a brain lesion apparently confined to the right hemisphere, group two was composed of ten protocols from individuals who had a brain lesion apparently confined to the left hemisphere, and group three consisted of twenty protocols from individuals who sustained diffuse brain injury (i.e., brain lesions involving both the right and left hemispheres). All of the individuals in each of these groups, were right-handed, did not have a self-reported history of learning disability, drug and alcohol abuse or a major psychiatric disorder.

The criterion of right hemisphere injury and left hemisphere brain injury was based exclusively on reports from brain imaging scans and, in some cases, neurological examinations which revealed some type of brain insult ostensibly localized to the right or left hemisphere in the absence of significant herniation, raised intracranial pressure or other mass effect. The right hemisphere group was composed of the following etiologies: tumors, gun shot wound, strokes, brain abscess, infarcts, and a contusion. The left hemisphere group was composed of the following etiologies: tumors, AVMs, strokes and brain abscesses. Individuals who sustained right or left hemisphere injury from a motor vehicle accident were not included because such injuries usually result in diffuse damage which may be undetected by brain scans.

The criterion of diffuse injury was based on reports from a patient's medical record that the patient experienced significant neurological sequelae, ostensibly resulting in bilateral lesions, following a motor vehicle accident or some other type of closed head injury. This group was composed of the following etiologies: traumatic head injuries from motor vehicle accidents, motorcycle accidents and falls. The neuropsychological data were not used as a determinant in the establishment of the criterion of right, left and diffuse brain injury.

Procedure

Six judges (three experts and three novices), who met the criteria for participation in the study as described above, were sent an introductory letter that outlined the purpose and nature of the study, and asked them to return an attached form indicating whether or not they were interested in participating (see Appendix B). These six judges voluntarily agreed to participate. Subsequently, they were sent a three-ring binder containing the 50 protocols, general information about how the neuropsychological protocols were obtained and how the criterion variables of right, left and diffuse damage were defined, base rate information, and instructions as to how to complete the task (see Appendix C).

Regarding base rates, judges were informed that 40 of the 50 protocols were cases of brain damage and 10 were non-brain damaged. In addition, they were informed that 20 of the 40 brain-damaged protocols were from individuals who sustained a diffuse brain injury, 10 who sustained a brain injury ostensibly confined to the right hemisphere and 10 who sustained a brain injury ostensibly confined to the left hemisphere (see Appendix C).

In addition, test norms were provided to participating neuropsychologists (Appendix D). They were not required to use these exact norms in the formation of their judgments. The norms were provided as a convenient aid.

Judges were requested to complete the task in four weeks. If, at the end of four weeks, the materials were not returned, a friendly phone contact was made to the judge as a reminder and as a way to ascertain when the materials might be returned.

Once the materials were returned, a thank you letter and a set of follow-up questions were sent to each judge (see Appendix E). The follow-up questions were designed to obtain information about the amount of time to complete the judgment task, a subjective estimate of the judges' degree of confidence in his/her judgments, a subjective estimate of protocols correctly predicted, and the degree of importance of each of the tests in relation to the presence vs absence judgment and the localization judgment. In regard to this latter point, judges were provided with a seven-point scale (Not at all important - Very important) and asked to provide a rating on each of the tests in terms of its importance to the judge in the making the presence vs absence and localization judgments, separately. Judges were also invited to make comments about the nature of the study and materials provided.

Methodological and Data Analysis Issues

Preliminary Analyses.

Eight of the 50 neuropsychological protocols contained missing data on one to three predictor test scores (apparently as a result of administrative constraints at the time of testing). Regression analysis was used to generate predicted test scores to be used in the place of the missing value(s). Specifically, if a neuropsychological protocol contained a missing value for one of the nine predictor scores, a regression equation was computed using the remaining eight predictor scores and their corresponding regression coefficients to generate a predicted score to be used in the place of the missing value. If two of the nine predictor scores were missing, a regression equation was computed using the remaining seven predictor scores and their corresponding regression coefficients to generate predicted scores to be used in the place of the missing values, and, similarly, if three of the nine predictor scores were missing.

Although the judges obviously did not have access to these predicted values for the missing data; it was hypothesized that the judges predicted or estimated values for missing data. That is, it was hypothesized that during the decision making process for each protocol, judges probably made a subjective prediction or estimate concerning the values of missing data (Levine, Johnson & Faraone, 1984). Therefore, the fact that values for the missing data were statistically computed for data analysis purposes is not a confound, and probably is not dissimilar to how judges dealt with the missing data.

Terminology.

A number of terms and indices were associated with the data analysis, and this section will list and provide a brief definition of the symbols. The goal of this section was to provide the reader with a useful explanation of the terms and indices, and a page of text to which the reader may refer for clarity. The definitions of the mathematical indices were quoted from Goldberg (1970, p. 424-425).

R_e: The linear predictability of the criterion: The multiple correlation between the cues and the criterion values ($rY_e.\hat{Y}_e$).

R_s: The linear predictability of the judge: The multiple correlation between the cues and the judge's predictions ($rY_s.\hat{Y}_s$).

r_a: The validity coefficient of the judge: The correlation between the judge's predictions and the actual criterion values ($rY_s.Y_e$). In this study, the judge's predictions were based on all cues (compare to r_m).

r_m: The validity coefficient of the linear model of the judge (also known as"bootstrapping"): The correlation between the predicted scores from the judge's model and the actual criterion values ($r\hat{Y}_s.Y_e$). In this study, the judge's model was based on nine cues (compare to r_a).

Δ : The differential validity of model over judge: The difference in validity coefficients between the model (r_m) and the judge (r_a). A positive value for this index favors the linear model over the judge.

G: The linear component of judgmental accuracy: The correlation between the predicted scores from the linear model of the judge and those from the linear model of the criterion ($r\hat{Y}_s.\hat{Y}_e$).

C: The nonlinear component of the judgmental accuracy: The correlation between the residual values of the criterion and the residual values of the judge's predictions after the linear components in both the criterion and the judge have been removed.

Majority: A Majority judge and Majority linear model of the judge were created for the novice and expert groups. Specifically, it was created by examining the judgments made by the judges and incorporating a "majority rules" decision criterion. For example, if the ecological judgment for a neuropsychological protocol indicated right hemisphere brain damage, then at least two [of the three] judges needed to indicate right hemisphere damage in order for this particular protocol to receive a Majority judgment of right hemisphere brain damage. If the ecological judgment for a neuropsychological protocol indicated right hemisphere brain damage, and two judges indicated left hemisphere brain damage (or diffuse brain damage), then whatever the majority ruled was used as the Majority judgment. In the case where each of the three judges provided a different judgment (e.g., right hemisphere, diffuse and left hemisphere brain damage; or no brain damage, right hemisphere brain damage and diffuse brain damage), it was decided that diffuse brain damage be used as the Majority judgment. (This occurred in three of the 50 cases for the experts and in one of the 50 cases for the novices.) The rationale for this procedure was that if one judge inferred ample evidence for a diffuse judgment, one for a left hemisphere judgment and one for a right hemisphere judgment, then the judges, as a majority, have found data to support damage throughout the brain (i.e, diffuse brain damage). Previous research showed that Wedding (1983) employed a majority rule type aggregate judge.

Composite: A Composite judge and Composite linear model of the judge were created for the novice and expert groups. This index was created by taking the arithmetic mean of the judgments. In the presence vs absence judgment, all judges provided a judgment; therefore, making the Composite judge a simple index to compute.

For the localization judgment, some complexity was inherent in computing the Composite judge. That is, not all of the judges provided a localization judgment for each and every protocol, because each may not have judged a protocol as demonstrating the presence of brain damage. In the case where only one judge out of the three made a localization judgment for a particular protocol, it was decided not to use this judgment in the creation of the Composit index.

The Composite index used in this study is congruent with the term "composite" used by Goldberg (1970) and Wiggins and Kohn (1971).

RESULTS

Judgment Task Analysis

Before the presentation of the myriad of data analyses ensues, it is important to consider whether the judges thought the test scores comprising each protocol were adequate to the purpose of making the two judgments. Table 5 shows the judges' subjective ratings (1=not at all important to 7=very important.) for each of the nine predictor test's relative importance in the decision making process for the presence vs absence and localization judgments (see Appendix F for a listing of the judges' subjective ratings for all the test scores)

For the presence vs absence judgment, only the gender cue received a mean rating lower than 3 for both the experts and novices. The majority of the cues received a rating of 4 or higher. The mean rating for the nine predictor cues ranged from 4 to 5 for the experts and from 3 to 6 for the novices. Overall, the cues provided per protocol as well as the nine predictor cues used to comprise the linear model were at a level to indicate that they were at least relevant to the judgment of presence vs absence of brain damage.

For the localization judgment, the majority of the cues received a mean rating of 4 or higher for the expert and novice groups. Appendix F shows that the experts provided a mean rating of less than 3 for the gender cue, while the novices provided a mean rating of less than 3 for the age, gender, Trail A and Trail B cues. Eight of the nine predictor cues received a mean rating of 4 or higher for the experts and novices. Only the Trail B cue received a relatively low mean rating (mean rating was 1) by the novices. Overall, the judges subjective estimate of the relative value of the cues in the determination of the localization judgment suggest that the cues were at least relevant.

Table 5

Judges' Subjective Weighting of the Nine Predictor Cues for the Two Judgments

	Presence/Absence Judgment		Localization Judgment	
Predictor Cues[a]	Experts	Novices	Experts	Novices
Similarities	4 (2-5)[b]	5 (4-6)	5.7 (5-6)	4.7 (3-6)
Block Design	5 (4-6)	5.3 (4-6)	5.3 (5-6)	5.7 (5-6)
Digit Symbol	4.7 (4-5)	5.7 (5-6)	4 (2-5)	4.3 (3-6)
Trail B	5 (3-6)	5 (2-7)	4 (3-6)	1.3 (1-2)
Finger Tapping right	4.3 (3-6)	3.7 (3-4)	5.3 (4-6)	6 (5-7)
Finger Tapping left	4.3 (3-6)	3.7 (3-4)	5.3 (4-6)	6 (5-7)
DRVERB	4.7 (3-6)	3.7 (1-6)	5.3 (4-6)	3.7 (1-7)
DRVIS	4.3 (3-5)	3.7 (1-6)	5 (4-6)	3.7 (1-7)
FAS	5 (4-6)	5 (4-6)	5.7 (5-6)	6 (6)

[a] Predictor cues. DRVERB=delayed recall trial, Logical Memory subtest; DRVIS=delayed recall trial Visual Reproduction subtest, WMS-R.

[b] Mean (range).

Analysis of the Cues in Relation to the Ecological Criteria

Before presenting the mathematical indices of the Brunswik Lens Model, it is important to examine the test scores in relation to the ecological side of the Model. Specifically, the mean values of the cues will be examined in relation to the actual criteria (i.e., normals, right hemisphere, left hemisphere and diffuse brain damage), and the intercorrelations of the cues will be explored.

As was stated in the Method section related to protocol selection, neuropsychological protocols chosen to represent the four groups were matched, as best as possible, for age and education. Results from one-way ANOVAs showed that the cues of age and education did not significantly differ among the the four sets (see Appendix G), therefore, validating that the four sets of protocols were adequately matched for age and years of education.

Appendix G provides a listing of the means and standard deviations for the 25 test scores for each of the four sets of protocols (i.e., normals, right hemisphere, left hemisphere and diffuse brain damage). In addition, 25 one-way ANOVAs were computed to determine if there were any significant differences between the four sets of protocols on the 25 cues. The results showed that there were significant differences between the four groups on ten cues: PIQ, Arithmetic subtest, Block Design subtest, Digit Symbol subtest, Trail B, Finger Tapping right hand, immediate recall trial Logical Memory subtest, delayed recall trial Logical Memory subtest, immediate recall trial Visual Reproduction subtest and delayed recall trial Visual Reproduction subtest. When significant differences were obtained (i.e., $\underline{p} < 0.05$), a Tukey post-hoc analysis was computed to determine what pair(s) of means were significantly different. All but one of the significant differences involved the normal group contrasted with one or more of the brain damaged groups, with the normals obtaining a significantly better score. In addition, of the remaining 15 cues that were not found to be significantly different among the groups, seven of the cues resulted in higher scores for the normal group compared to the brain damaged groups (higher scores, except for Trail B and the Category Test, are associated with better performance). Overall, 17 of the 25 cues (68%)

resulted in higher or better scores for the normals. Given the low number of subjects per group as well as the robustness (i.e., lack of sensitivity) of some tests to brain damage (e.g., Vocabulary subtest), the mean differences among the four groups is what was expected and consistent with previous research on patient vs control differences (e.g., Van Gorp, Satz, Hinkin, Evans, & Miller, 1989).

Tables 6 presents the intercorrelations among the nine predictor cues. As is apparent, the majority of the predictor cues significantly correlated with each other. One predictor, left hand Finger Tapping test, significantly correlated with only two other cues (i.e., RHFI and BD). The absolute value of the correlations were not so high as to cause linear dependence in the predictor cues (meaning that a row(s) or column(s) of a matrix is a linear combination of other vectors in the matrix, Pedhazur, 1982). The absence of linear dependence among the predictor cues suggests that multicollinearity did not interfere with the estimation of regression coefficients (Pedhazur, 1982).

In addition, Table 6 shows the correlations among the predictors and the two criteria. Four of the nine predictors significantly correlated with the presence/absence judgment, while two of the nine significantly correlated with the localization judgment. Overall, out of the total set of 27 cues presented to the judges, seven correlated with the presence/absence judgment (PIQ - r - 0.37, Digit Symbol subtest - r = 0.62, right hand Finger Tapping - r = 0.37, immediate recall Logical Memory subtest - r = 0.35, delayed recall Logical Memory subtest - r = 0.40, immediate recall Visual Reproduction subtest - r = 0.33, and delayed recall Visual Reproduction subtest - r = 0.34); and five correlated with the localization judgment (Information subtest - r = 0.33, Digit Span subtest - r = 0.35, immediate recall Logical Memory subtest - r = 0.36, delayed recall Logical Memory subtest - r = 0.43, and FAS - r = 0.35).

Table 6

Intercorrelations Among the Nine Predictor Cues[a], and Correlations Between the Predictor Cues and the Ecological Criteria[b]

	Sim.	BD	DS	TrailB	RHFT	LHFT	DRVERB	DRVIS	FAS	PORA[b]	LOC
Sim.	--	0.40**	0.30*	-0.28*	0.29*	0.19	0.58**	0.40**	0.30*	-0.24	-0.05
BD	--	--	0.50**	-0.53**	0.29*	0.31*	0.37**	0.51**	0.16	-0.17	0.28
DS	--	--	----	0.58**	0.53**	0.24	0.45**	0.53**	0.36*	-0.62**	-0.08
TrailB	--	--	----	---	-0.49**	-0.12	-0.40**	0.60**	0.41**	0.27	0.06
RHFT	--	--	----	---	---	0.54**	0.38**	0.35*	0.37**	-0.38**	-0.25
LHFT	--	--	----	---	---	---	-0.00	0.24	0.07	-0.09	0.17
DRVERB	--	--	----	---	---	---	---	0.40**	0.30*	-0.40**	-0.43**
DRVIS	--	--	----	---	---	---	---	---	0.34*	-0.30*	0.14
FAS	--	--	----	---	---	---	---	---	---	-0.10	-0.35*

*= $p < 0.05$.

**= $p < 0.01$.

[a] Cues=Similarities subtest, Block Design subtest, Digit Symbol subtest WAIS-R. Right hand Finger Tapping Test left hand Finger Tapping Test. Delayed trial, Logical Memory subtest, delayed trial Visual Reproduction subtest WMS-R.

[b] Ecological criteria. PORA=Presence vs absence criterion. Loc=Localization criterion.

Analysis of the Presence vs Absence Judgment

The presence vs absence judgment represented a categorical criterion in the regression equation, and judgments of presence were coded with "1s" and absence were coded with "0s." Ten sets of analyses (primarily consisting of regression and correlation analyses) were computed for each of the seven mathematical indices: three sets of equations for the individual novice judges, three sets of equations for the individual expert judges, one set of equations to represent the Majority novice judge, one set of equations to represent the Majority expert judge, one set of equations for the Composite novice judge and one set of equations for the Composite expert judge. In addition, one regression and correlation analysis was computed to produce r_m using unit weights. Specifically, eight of the predictor cues were weighted + 1/9 and one (Trail B) was weighted - 1/9 in the equal weights regression analysis. All of the least squares regression analyses were computed by combining all nine cues at the same time (often referred to as a simultaneous procedure).

Hit Rate and Validity Coefficient of the Judge (r_a).

One of the hypotheses of this study was that there would be no significant or notable differences between the two sets of judges (i.e., experts and novices) in terms of hit rate and judgmental accuracy (r_a).

Table 7 provides a listing of the judges' success in correctly identifying the presence vs absence protocols (i.e., normals and brain damaged). Recall that the judges were given the base rates of these groups as part of the background materials for the judgment task. As is evident, the expert group correctly identified an average of 50% (range=40% to 60%) of the normal protocols, while the novice group correctly identified an average of 40% (range=30% to 50%). For the protocols reflecting brain damage, the experts correctly identified an average of 87% (range=85% to 87.5%), while the novices correctly identified an average of 84% (range=82.5% to 85%).

Overall, the expert group achieved a slightly higher hit rate (average hit rate = 79%, range = 76% to 80%) than the novice group (average hit rate = 75%, range = 72% to 78%). Therefore, the experts slightly outperformed the novices in the presence vs absence judgment, but not to an extent to disprove the hypothesis. Relative to the base rates, however, none of the expert or novice judges exceeded the base rates. Two of the expert judges matched the base rates (i.e., 80%), but none of the novices. In terms of aggregate judgments, the Majority expert index did exceed the base rates (82% vs 80%, respectively).

Table 7 shows that the validity coefficient of the judge (r_a) was somewhat higher for the experts ($\overline{r_a}$=0.34, range=0.25 to 0.38) than for the novices ($\overline{r_a}$=.24, range 0.12 to 0.34), suggesting a higher or stronger relationship between the experts' judgments and the actual ecological criterion, as compared to the novices' judgments and the actual ecological criterion. The Majority index for the experts were higher than for the novices, suggesting that combining the policy of this group of experts will lead to a much higher relationship between the judges' predictions and the actual criterion values as compared to the novices. Also of note, was that the Majority and Composite indices outperformed most of the judges in their respective groups, supporting the idea that combining judgments tends to eliminate error (this statement is supported by the generally higher hit rate value for the Majority judge as compared to the hit rate value of most of the individual judges).

It is important to note that the rank order of r_a perfectly corresponds to the hit rate rank order in Table 7. Specifically, the higher r_a values correspond to the better hit rates, and vice versa (this is intuitive given that r_a is defined as the correlation between the judge's prediction and the actual criterion values).

Table 7

Hit Rate[a] and Validity Coefficients for the Presence vs Absence Judgment

	Protocols[b]			Validity Coefficient[c]
	Normals	BD	Hit Rate	r_a
Experts				
#1	5/10	35/40	40/50	0.38
#2	5/10	35/40	40/50	0.38
#3	4/10	34/40	38/50	0.25
Majority	6/10	35/40	41/50	0.46
Novices				
#1	4/10	34/40	38/50	0.25
#2	3/10	33/40	36/50	0.12
#3	5/10	34/40	39/50	0.34
Majority	4/10	34/40	38/50	0.25

[a] Hit rate= Ratio of correct to total judgments.

[b] BD= Brain damaged.

[c] Validity coefficient: r_a= validity coefficient of the judge.

Linear Model of the Judge and the Differential Validity of the Model Over the Judge.

The second hypothesis was that the linear model of the judge (i.e., validity coefficient, r_m) will be equal to or superior to the judge (r_a). Recall that r_a is the correlation of the judge's judgment and the criterion, with the judge's judgment based on all cues; while r_m is the correlation between the predicted scores of nine cues from the judge's model (computed via regression analysis) and the criterion.

The validity coefficient of the linear model of the judge (r_m, a.k.a. the validity of bootstrapping) was notably higher for the experts ($\overline{r_m}$=0.39, range=0.36 to 0.41) as compared to the novices ($\overline{r_m}$=0.26, range=0.23 to 0.30) (see Table 8). This suggests that a linear model of an expert judge will lead to more accurate predictions than a linear model of a novice judge. The Majority and Composite linear models were generally higher than most of the individual judges' linear models.

Of greater importance, was the difference between r_a and r_m (i.e., Δ =the differential validity of model over judge) (see Table 8). The data shows that the validity coefficients for the linear model vs the judge were essentially equal to each other in four of the five comparisons for each group. In one of the comparisons in each group, the linear model outperformed the judge (see Δ for expert #3 and novice #2). These results are consistent with thirty-five years of research that has supported a conclusion that a simple linear model of the judge will generally be equal to or superior to the judge.

Table 8

Judge versus Linear Model of the Judge: Mathematical Indices of the Brunswik Lens Model for the Presence Versus Absence of Brain Damage Judgment

	Mathematical Indices of the Brunswik Lens Model[a]						
	R_e	r_a	r_m	Δ	R_s	G	C
Experts							
#1	0.69	0.38	0.40	0.02	0.61	0.58	0.24
#2	0.69	0.38	0.36	-0.02	0.68	0.53	0.24
#3	0.69	0.25	0.41	0.16	0.62	0.60	-0.02
Majority	0.69	0.46	0.44	-0.02	0.67	0.64	0.36
Composite	0.69	0.40	0.41	0.01	0.73	0.59	0.20
Novices							
#1	0.69	0.25	0.23	-0.02	0.66	0.34	0.17
#2	0.69	0.12	0.25	0.12	0.63	0.36	-0.06
#3	0.69	0.34	0.30	-0.04	0.72	0.43	0.26
Majority	0.69	0.25	0.25	0.00	0.67	0.37	0.15
Composite	0.69	0.27	0.26	0.01	0.74	0.38	0.17

[a] R_e=The linear predictability of the criterion. r_a=The validity coefficient of the judge. r_m=The validity coefficient of the linear model of the judge. Δ=The differential validity of model over man; A positive value favors the model. R_s=The linear predictability of the judge. G=The linear component of judgmental accuracy. C=The nonlinear component of judgmental accuracy.

Ecological Side of the Brunswik Lens Model

Linear Predictability (R_e). The linear predictability of the criterion ($R_{c.}$ The multiple correlation between the cues and the ecological criterion) was 0.69 (see Table 8). The value of 0.69 indicated that a large proportion of the variance (i.e., approximately 48%) between the nine predictor cues and ecological criterion judgment was accounted for by a simple linear model. In addition, note that in this study, R_e is equivalent to the actuarial model.

Human Side of the Brunswik Lens Model

Linear Predictability (R_s). The linear predictability of the judge ($R_{S.}$ The multiple correlation between the cues and the judge's judgments) was slightly higher for the novices ($\overline{R_S}$=0.67, range=0.63 to 0.72) as compared to the experts ($\overline{R_S}$=0.64, range=0.62 to 0.68). The Majority and Composite R_S indices were generally higher than the R_S index for most of the individual judges, suggesting that combining the policy of each set of judges (i.e., novices and experts) contributed to higher linear predictability.

Linear and Nonlinear Components of Judgmental Accuracy (G and C). The expert judges utilized a much higher linear component ($\overline{G}$=0.57, range =0.53 to 0.60)(see Table 8) in their judgmental accuracy than the novices ($\overline{G}$=0.38, range=0.34 to 0.43). The size of the nonlinear component to judgmental accuracy was only slightly higher for the experts ($\overline{C}$=0.15, range -0.02 to 0.24) as compared to the novices ($\overline{C}$=0.12, range=-0.06 to 0.26). Overall, across both sets of judges, the linear component to judgmental accuracy was much greater than the nonlinear component.

The Majority and Composite linear component indices were generally higher than most of the judges considered individually which suggests that combining judgments tended to enhance linear judgmental accuracy. Regarding the Majority's and Composite's nonlinear component to judgmental accuracy, a different picture emerged. That is, for the expert group, only the Majority's "C" index produced a greater value than for the individual judges; similarly, for the novice group, the Majority's and Composite's "C" index produced a value greater than only one of the judges. Therefore, unlike the findings of the linear component (G), combining the judgments from the expert and novices groups tended not to enhance the nonlinear component (C) of judgmental accuracy. This pattern of results suggest that judges were using similar linear policies, but dissimilar nonlinear policies.

Analysis of Mathematical Indices Associated with the Relative Magnitude and Rank Ordering of r_a.

It is useful to examine the relative contribution of the linear component of judgmental accuracy (G) and the nonlinear component of judgmental accuracy (C) independent of the value of r_a. That is, given Tucker's (1964) decomposition of the validity coefficient,

$$r_a = G R_e R_s + C \sqrt{1-R_e^2}\sqrt{1-R_s^2} \qquad [1]$$

it is not possible to determine the relative contribution of G and C independent of r_a, because a high G and C (along with the other indices) are mutually dependent on r_a, and vice versa. But, by dividing each side of equation [1] by r_a gives:

$$1 = \frac{G R_e R_s}{r_a} + \frac{C\sqrt{1-R_e^2}\sqrt{1-R_s^2}}{r_a} \qquad [2]$$

The first term will be referred to as the "relative linearity coefficient" of judgmental accuracy and the second term as the "relative nonlinearity coefficient" of judgmental accuracy. Equation [2] allows for the assessment of the relative contributions of G and C independent of r_a. If judgmental accuracy, whatever its level, is exclusively based on a linear policy, then the first term on the right hand side of equation [2] (i.e., the relative linearity coefficient) would be 1 and the second term (i.e., the relative nonlinearity coefficient) would be 0. Similarly, a judge who derives his/her accuracy exclusively from a nonlinear policy would have the relative nonlinearity coefficient equal to 1 and the relative linearity coefficient equal to 0.

Results showed that independent of the value of r_a, the relative linearity coefficient was much higher than the relative nonlinearity coefficient. Specifically, the mean relative linearity coefficient for the three experts was 0.77 (range=0.642 to 1.03) and for the three novices it was 0.76 (range=0.619 to 1.30). (Note that a value greater than 1 was obtained in some cases for the relative linearity coefficient, because some judges employed a negative nonlinear component. Regardless, the relative linearity and nonlinearity coefficients summed to 1.) The mean relative nonlinearity coefficient for the three experts was 0.217 (range=-0.045 to 0.362) and for the three novices it was 0.158 (range=-0.281 to 0.384).

Analysis of Mathematical Indices Associated with the Relative Magnitude of r_m.

In terms of understanding the magnitude of r_m, it is useful to analyze an equation offered by Goldberg (1970, p. 425):

$$r_m = GR_e. \qquad [3]$$

Simply, the greater the linear component of judgmental accuracy (i.e., G) and the linear predictability of the criterion (i.e., R_e), the greater the value of r_m (see Table 8).

Analysis of Mathematical Indices Associated with Δ.

Δ was previously defined as simply the difference in validity coefficients between the model (r_m) and the judge (r_a). It is useful to examine in greater detail the mathematical indices of the Brunswik Lens Model that when combined in an equation predict the differential validity of model over judge or judge over model (i.e., a positive or negative Δ).

Goldberg (1970, p.425) formulated an equation that predicts the differential validity of the linear model over the judge, and vice versa:

$$\Delta = GR_e(1-R_s) - C\sqrt{1-R_e^2}\sqrt{1-R_s^2}. \qquad [4]$$

This equation indicates that the model will outperform the judge when:

$$GR_e(1-R_s) > C\sqrt{1-R_e^2}\sqrt{1-R_s^2}. \qquad [5]$$

Briefly, this equation suggests that, all other indices about equal, the higher the value of G relative to the value of C the greater likelihood that the linear model will outperform the judge (i.e., $r_m > r_a$). Also, all other indices about equal, a value for R_e of 0.71 or higher will contribute to a progressively larger value on the left side relative to the right side of the equation; while, a value for R_e of 0.70 or smaller will contribute to a progressively larger value on the right side of the equation relative to the left side. Therefore, all other indices being equal, the greater the linear predictability of the ecology (i.e., $R_e > 0.71$), the greater the probability that the linear model will outperform the judge (i.e., $r_m > r_a$). In addition, a high value of R_s will favor the judge over the linear model, because a high value of R_s will contribute to a relatively low value on the left side of the equation and a relatively high value on the right side of the equation. This latter finding is supported by a general rule of thumb offered by Goldberg (1970); essentially, all other indices being equal, as the judge becomes more linearly predictable (i.e., as R_s approaches unity) his/her validity coefficient will become less distinguishable from the linear model (i.e., $r_a = r_m$, or $r_a > r_m$).

The data show that the value of R_e=0.69, therefore, only a very slight advantage is given to the right side of the equation, favoring the judge over the model. Equation [5] indicates that the extent to which the multiplicative values on the left side of the equation (i.e., G, R_e and $1 - R_s$) exceeds the right side of the equation (i.e, C, ($\sqrt{1-R_e^2}$), ($\sqrt{1-R_s^2}$)) will produce an outcome value that favors the linear model over the human judge. Table 8 shows that this relationship (i.e., equation [5]) occurred in 5 of the 10 comparisons. Therefore, it is possible to simply obtain a value for Δ by (a) simply subtracting r_a from r_m (a positive value favors the model over the judge) or (b) by examining and computing in greater detail the mathematical indices of Goldberg's equation (i.e., equation [5]).

Rank Ordering of Validity Coefficients for the Presence versus Absence Judgment

First, in order to compute the validity coefficient for the equal weight model, all of the predictor cues and the criterion were converted into standardized Z-scores.

Table 9 displays the rank ordering of validity coefficients for the presence vs absence judgment. Four major findings were suggested: (a) The actuarial formula (i.e., R_e) was far superior to any other judge or model. (b) A simple unit weighting formula outperformed most rival judges or models. (c) In general, the most accurate linear model of a judge outperformed its respective human judge (the only exception was for the Majority judge which outperformed the Majority linear model). (d) Aggregate judges (i.e., the Most Accurate Majority or Composite judges and linear models) outperformed all of the individual judges, and all but one of the linear models of a judge.

Table 9

Rank Ordering of Validity Coefficients for the Presence versus Absence Judgment

Presence versus absence judgment	Validity Coefficient[a]
Linear predictability (R_e)[b]	0.69
Most accurate Majority judge (expert)	0.46
Most accurate Majority model judge (expert)	0.44
Unit Weight model[c]	0.43
Most accurate Composite model judge (expert)	0.41
Most accurate model (expert #3)	0.41
Most accurate Composite judge (expert) -	0.40
Most accurate judge (experts # 1 & 2)	0.38
Least accurate Composite judge (novice)	0.27
Least accurate Composite model judge (novice)	0.26
Least accurate Majority model judge (novice)	0.25
Least accurate Majority judge (novice)	0.25
Least accurate model (novice # 1)	0.23
Least accurate judge (novice #2)	0.12

[a] Validity coefficient refers to r_a and r_m.

[b] In this study, R_e is equivalent to an actuarial formula (i.e., the criterion is equal to the linear combination of the nine predictor cues).

[c] Given the coding of the presence vs absence judgment, as expected a negative correlation resulted for the unit weight model, but for clarity of comparison purposes a positive value was displayed in the table.

Ecologically Valid Policy.

The final data analysis for the presence vs absence judgment involves comparing the standardized beta weights on the ecological side and the human judgment side of the Brunswik Lens Model. This comparison will determine the relative importance of the cues in relation to the actual criterion and in relation to each judge's judgments, therefore examining how closely the judge captured the ecologically valid policy. The standardized beta weights of the nine predictor cues from the ecological side of the Brunswik Lens Model were computed on the actual criterion (i.e., hit rate of 100%). The standardized beta weights from the human side of the Brunswik Lens Model were computed on each judge's judgments, therefore, with varying hit rates (see Table 7). In principle, as the judge's judgments approach a hit rate of 100%, his/her standardized beta weight values will mirror those of the ecology.

Table 10 shows the ecologically valid policy. It is important to recall that the standardized beta weights are based on scores from an extreme base rate sample (i.e., 80% brain damage and 20% normal). Digit Symbol was far and away the most important cue to be weighted in the determination of the presence vs absence judgment (beta = -.670). The next two cues of importance were Block Design and FAS with weightings of 0.2 or greater. Three cues had a weighting of 0.1 or greater: Finger Tapping right, DRVERB and Trail B. The three cues with the least importance, i.e., beta weight less than 0.1 were: Finger Tapping left, Similarities and DRVIS.

In terms of the three most important cues in the ecology, experts and especially novices notably underweighted the importance of Digit Symbol, but provided relatively high ratings for Block Design and FAS. For the three least important cues in the ecology, experts and novices appropriately estimated the relative insignificance of DRVIS and somewhat so for Similarities. Both groups of judges overestimated the relative insignificance of Finger Tapping left hand.

Finally, judges' subjective weightings of the nine predictor cues (see Table 5) were compared to their weightings obtained from regression analysis (see Table 10). Overall, there was a low association or relationship between a judge's subjective weighting of the relative importance of a cue in relation to the judgment and the cue's beta weight. In general, most of the judges subjectively overestimated the importance of the cues (evidence by ratings of 4 to 6) in comparison to the beta weights (where there were high, moderate and low beta weights). Therefore, judges' discrepancy between their subjective weights and beta weights of cues indicated that, in particular, judges did not weight cues as they subjectively estimated, and, in general, were not fully and accurately aware of their cognitive processes in relation to the judgment task (Nisbett & Wilson, 1977).

Table 10

Standardized Beta Weights for the Nine Predictor Cues for the Ecological Side and Human Judgment Side of the Brunswik Lens Model for the Presence vs Absence Judgment

	Ecology	Experts					Novices				
Cues		1	2	3	Maj.	Comp	1	2	3	Maj.	Comp
Similarities	-.074	-.120	-.243	.040	-.061	-.129	-.009	-.126	-.109	-.046	-.090
Block Design	.205	-.155	-.267	-.204	-.144	-.250	-.378	-.289	-.379	-.346	-.387
Digit Symbol	-.670	-.186	-.225	-.136	-.248	-.219	.047	-.093	-.065	-.023	-.042
Trail B	-.126	.023	-.078	-.094	.036	-.059	-.159	-.041	-.172	-.111	-.138
Tapping Right	-.163	.223	.162	-.127	.249	.103	-.119	-.036	-.089	-.124	-.090
Tapping Left	.090	-.175	-.217	-.191	-.218	-.233	-.188	-.273	-.168	-.207	-.232
DRVERB	-.153	-.330	-.181	-.129	-.444	-.256	-.112	.003	-.083	-.063	-.071
DRVIS	-.072	.025	.079	-.139	.058	-.014	-.060	.122	-.044	.038	.006
FAS	.202	-.136	-.167	-.136	-.083	-.175	-.322	-.295	-.311	-.338	-.344

Analysis of the Localization Judgment

Some complexity was inherent in analyzing the localization judgment. That is, judges could have rendered four localization judgments for any given protocol: right hemisphere brain damage, left hemisphere brain damage, diffuse brain damage, or no brain damage. A no brain damage judgment could have occurred when a judge incorrectly decided a protocol demonstrated the absence of brain damage (i.e., false negative); since there was brain damage, a localization judgment should have been declared, but no entry was made.

The matrix in Table 11 presents the different ways in which the judge could have responded. The top, horizontal part of the matrix reflects the ecology, and the vertical side of the matrix reflects the human judgment. Recall that there were 10 protocols that were associated with nonbrain damage, 10 associated with right hemisphere brain damage, 10 associated with left hemisphere brain damage, and 20 associated with diffuse brain damage (see Totals on the bottom, horizontal portion of the Table). It is possible for the judges to have judged a protocol indicating nonbrain damage even though it is from the sample associated with one of the three localization judgments in the ecology (i.e., a false negative error. See FNs in Table 11). It is possible for the judges to have judged a protocol with one of the localization judgments, given that this protocol comes from one of the three localization categories in the ecology (i.e., a true positive. See TPs). Finally, it is possible for the judges to have judged a protocol with one of the localization judgments when, in fact, this protocol was from a nonbrain damaged protocol in the ecology (i.e., a a false positive. See FPs).

It was decided that only localization protocols from the ecology in which the judge made one of the three localization judgments (TPs) would be used in the regression analyses to assess judgmental accuracy. The implication of this decision is that this will probably result in an artificially inflated validity coefficient for the judge, because each judge's error associated with the FNs and FPs were removed.

Table 11

Possible Types of Judgments Made By The Judge In Relation to the Ecology

	Ecology[a]				
	NBD	RHBD	LHBD	DBD	(Totals)
Judge					
NBD		FN[b]	FN	FN	??
RHBD	FP[c]	TP[d]	TP	TP	??
LHBD	FP	TP	TP	TP	??
DBD	FP	TP	TP	TP	??
(Totals)	10	10	10	20	

[a] Ecology: NBD=Nonbrain damaged. RHBD=Right hemisphere brain damaged. LHBD=Left hemisphere brain damaged. DBD=Diffuse brain damaged.

[b] FN= False negative. A clearly incorrect judgment. Actual brain damage was present, but no localization judgment is made.

[c] FP= False positive. A clearly incorrect judgment. A protocol associated with nonbrain damage was given a localization judgment.

[d] TP= True Positive. A correct judgment was made as to the presence of brain damage, but was it localized correctly?

Localization judgments were coded by assigning a "-1" for right hemisphere brain damage, "0" for diffuse brain damage, and "+1" for left hemisphere brain damage.

The same ten sets of analyses (primarily consisting of regression and correlation analyses) that were done for the presence vs absence judgment were computed for each of the seven mathematical indices for the localization judgment. In addition, one regression and correlation analysis was computed for a unit weight model. Recall that the localization judgment was based on, at most, 40 protocols (10 of the 50 protocols were non-brain impaired). All of the least squares regression analyses were computed by entering all nine cues at the same time.

Hit Rate and Validity Coefficient of the Judge (r_a).

One of the hypothesis of this study was that there would be no significant or notable difference between the two sets of judges in terms of protocols accurately judged (i.e., hit rate and r_a).

Table 12 shows that the novices did slightly better than the experts in correctly identifying the localization of brain damage. Specifically, novices correctly identified an average of 60% (range=50% to 70%) of the right hemisphere brain damage protocols, while experts correctly identified an average of 52.5% (range=50% to 60%). For the left hemisphere brain damage protocols, novices correctly identified an average of 42.5% (range=20% to 60%), while experts correctly identified an average of 32.5% (range=20% to 50%). Finally, for the diffuse brain damage protocols, novices correctly identified an average of 41% (range=35% to 50%), while experts correctly identified an average of 41% (range=35% to 55%). Overall, the novices achieved an average hit rate of 47% (range = 40% to 50%), while the three experts achieved an average hit rate of 42% (range = 37.5% to 50%). Therefore, although the novices slightly outperformed the experts, the two sets of judges were not notably different from each other.

The base rate prediction is diffuse because it is the most frequent protocol in the localization sample. If a judge simply employed the base rate prediction, he/she would have achieved a hit rate of 50% (20 diffuse protocols/total localization sample equalled 40). The novice group came the closest to achieving the base rate prediction level, but neither group outperformed the base rate. It is important to note that comparing the judges' hit rates to the the base rate prediction for the localization judgment was somewhat confounded. The judges did not know which 40 out of the sample of 50 protocols were the true localization protocols, while the base rate prediction was computed assuming such knowledge. Therefore, in this analysis, judges' hit rates were compared to perhaps an unfair base rate prediction level.

Table 12

Hit Rate[a] and Validity Coefficients for the Localization Judgment

		Protocols[b]			Validity Coefficients[c]
	RHBD	LHBD	DBD	Hit Rate	r_a
Experts					
#1	5/10	2/10	8/20	15/40	0.34
#2	5/10	3/10	7/20	15/40	0.48
#3	6/10	5/10	9/20	20/40	0.54
Majority	5/10	3/10	11/20	19/40	0.51
Novices					
#1	6/10	6/10	8/20	20/40	0.63
#2	5/10	5/10	10/20	20/40	0.53
#3	7/10	2/10	7/20	16/40	0.50
Majority	6/10	4/10	10/20	20/40	0.59

[a] Hit rate=Ratio of correct to total judgments.

[b] RHBD=Right hemisphere brain damage. LHBD=Left hemisphere brain damage. DBD=Diffuse brain damage.

[c] Validity coefficients: r_a=validity coefficient of the judge.

In addition, Table 12 shows that the validity coefficient for the judge (r_a) was higher for the novices (r_a=0.55, range=0.50 to 0.63) than for the experts (r_a=0.45, range=0.34 to 0.54). The validity coefficient for the Majority judge for each group tended to be higher than most of the r_a value for any individual judge.

It is important to note that the rank order of r_a generally corresponded to the hit rates' rank order in Table 8. Specifically, the higher r_a values generally corresponded to the better hit rates, and vice versa.

Linear Model of the Judge and the Differential Validity of the Model Over the Judge

The second hypothesis of this study was that the validity coefficient of the linear model of the judge (r_m) would be equal to or superior to the validity coefficient of the judge (r_a).

The validity coefficient of the linear model of the judge (r_m) ranged from 0.51 to 0.61 ($\overline{r_m}$=0.57) for the experts, and ranged from 0.52 to 0.74 ($\overline{r_m}$=0.59) for the novices (see Table 13). Although the mean values for both groups were similar, there was much greater spread or variability associated with the novices' r_m values, suggesting that this group of judges used a variable level of linear component to judgmental accuracy (given that r_m = R_eG). The Majority and Composite linear models of the judges tended to be equal to or greater than most of the r_m values of any individual judge.

Of most importance was the value of the differential validity of the model over the judge (Δ). A positive Δ value means that the model outperformed the judge. The linear model outperformed the judge in all of the five comparisons in the expert group. In the novice group, the linear model outperformed the judge in one comparison (novice #1), while in the remaining four comparisons the linear model and the judge were about equal (see Table 13). These results indicate that the linear model is equal to or better than the human judge, and are consistent with the findings found for the presence/absence judgment.

Table 13

Judge versus Linear Model of the Judge: Mathematical Indices of the Brunswik Lens Model for the Localization of Brain Damage Judgment

	Mathematical Indices of the Brunswik Lens Model[a]						
	R_e	r_a	r_m	Δ	R_s	G	C
Experts							
#1	0.70	0.34	0.51	0.16	0.76	0.81	-0.18
#2	0.70	0.48	0.60	0.14	0.86	0.91	-0.19
#3	0.70	0.54	0.61	0.07	0.76	0.80	0.24
Majority	0.70	0.51	0.58	0.07	0.83	0.86	0.03
Composite	0.70	0.52	0.61	0.09	0.89	0.91	-0.15
Novices							
#1	0.70	0.63	0.74	0.11	0.83	0.97	0.18
#2	0.70	0.53	0.52	-0.01	0.78	0.79	0.23
#3	0.70	0.50	0.52	0.02	0.72	0.76	0.25
Majority	0.70	0.59	0.61	0.02	0.76	0.85	0.30
Composite	0.70	0.65	0.66	0.01	0.85	0.91	0.29

[a] R_e=The linear predictability of the criterion. r_a=The validity coefficient of the judge. r_m=The validity coefficient of the linear model of the judge. Δ=The differential validity of model over man; A positive value favors the model. R_s=The linear predictability of the judge. G=The linear component of judgmental accuracy. C=The nonlinear component of judgmental accuracy.

Ecological Side of the Brunswik Lens Model

Linear Predictability (R_e). The linear predictability of the criterion (R_e) was 0.70 (see Table 13). The value of 0.70 indicated that a large proportion of the variance (i.e., 49%) between the nine predictor cues and ecological criterion judgment was accounted for by a simple linear model. In addition, note that in this study, R_e is equivalent to the actuarial model.

Human Side of the Brunswik Lens Model

Linear Predictability (R_s). The linear predictability of the judge (R_s) was similar for both the expert ($\overline{R_s}$=0.79, range=0.76 to 0.86) and novice groups ($\overline{R_s}$=0.78, range=0.72 to 0.83) (see Table 13).

The Majority judge, for each group, was higher than two out of the three expert judges, while it was higher than one of the three novice judges. The Composite judge, for each group, was higher than all of the judges in their respective groups.

Linear and Nonlinear Components of Judgmental Accuracy (G and C) Expert and novice judges tended to employ a similar level of linear component to judgmental accuracy (G) (Experts: $\overline{G}$=0.84, range=0.80 to 0.91. Novices: $\overline{G}$=0.84, range=0.76 to 0.97) (See Table 13). Each group had one judge with a very high value of G (expert #2 and novice # 1). Considering the value of G with these two judges removed, it was evident that the experts (0.80 & 0.81) had a slightly greater level of linear component to judgmental accuracy than the novices (0.76 & 0.79).

As was found in the presence/absence judgment, the Majority and Composite judges tended to have a higher linear component to judgmental accuracy than most of the individual judges in their respective groups.

A much different picture emerged when analyzing the data from the nonlinear component to judgmental accuracy (C). Two of the experts employed a negative C value, while one had a positive C value ($\overline{C}$=-0.04, range=-0.19 to 0.24) (see Table 13). (A negative C value essentially means that the nonlinear component to judgmental accuracy was zero or inconsequential. Therefore, a C value of -0.20 is considered essentially equal to 0.00) The novices employed a modest amount of nonlinear component to judgmental accuracy ($\overline{C}$=0.22, range=0.18 to 0.25).

The Majority and Composite judges tended to enhance the nonlinear component to judgmental accuracy for most of the novice judges, but not for the experts.

Analysis of Mathematical Indices Associated with the Relative Magnitude and Rank Ordering of r_a

It is useful to examine the relative contribution of the linear component of judgmental accuracy (G) and the nonlinear component of judgmental accuracy (C) independent of the value of r_a as was done in the presence/absence judgment (see equation [2])

Results showed that independent of the value of r_a, the relative linearity coefficient was much more important than the relative nonlinearity coefficient. Specifically, the mean relative linearity coefficient (i.e., value on the left side of the summation sign) for the three experts was 1.07 (range=0.788 to 1.27) and for the three novices it was 0.825 (range=0.766 to 0.895).

The mean relative nonlinearity coefficient (i.e., the value on the right side of the summation sign) for the three experts was -0.061 (range=-0.246 to 0.206) and for the three novices it was 0.185 (range=0.114 to 0.248). Integrating the data from the experts' and novices' relative linear and nonlinear coefficients, it is inferred that the novices' higher hit rate for the localization protocols was in part a reflection of a slightly larger non-linear component.

Analysis of Mathematical Indices Associated with the Relative Magnitude of r_m.

In terms of understanding the magnitude of r_m, it is useful to analyze an equation offered by Goldberg (1970, p. 425):

$$r_m = GR_e \qquad [3]$$

Simply, the greater the linear component of judgmental accuracy (i.e., G) and the linear predictability of the criterion (i.e., R_e), the greater the value of r_m.

Analysis of Mathematical Indices Associated with Δ.

Δ was previously defined as simply the difference in validity coefficients between the model (r_m) and the judge (r_a). It is useful to examine in greater detail the mathematical indices of the Brunswik Lens Model that when combined in an equation predict the differential validity of model over judge or judge over model (i.e., Δ).

Goldberg (1970, p.425) formulated an equation that predicts the differential validity of the linear model over the judge, and vice versa:

$$\Delta = GR_e(1-R_s) - C\sqrt{1-R_e^2}\sqrt{1-R_s^2}. \quad [4]$$

This equation indicates that the model will outperform the judge when:

$$GR_e(1-R_s) > C\sqrt{1-R_e^2}\sqrt{1-R_s^2}. \quad [5]$$

(The reader may refer to the Presence/absence section for a descriptive analysis of this equation).

The data in this study shows that the value of R_e= 0.70 in all cases, therefore, only a very slight advantage is given to the right side of the equation, favoring the judge over the model. Equation 5 indicates that the extent to which the multiplicative values on the left side of the equation (i.e., G, R_e and $1-R_s$) exceeds the right side of the equation (i.e, C, ($\sqrt{1-R_e^2}$), ($\sqrt{1-R_s^2}$)) will produce an outcome value that favors the linear model over the man. Table 13 shows that this relationship (i.e., equation [5]) occurred in 9 of the 10 comparisons. Therefore, it is possible to simply obtain a value for Δ by (a) simply subtracting r_a from r_m (a positive value favors the model over the judge) or (b) by examining and computing in greater detail the mathematical indices of Goldberg's equation (i.e., equation [5]).

Rank Ordering of Validity Coefficients for the Localization Judgment

In order to compute the validity coefficient for the unit weight model, all of the predictor cues and the criterion had to be converted into standardized Z-scores. The nine predictor cues were weighted as follows: Trail B and Digit Symbol were coded "0"; Similarities, right hand Finger Tapping, DRVERB and FAS were coded "+1/4"; and Block Design, left hand Finger Tapping and DRVIS were coded "-1/3."

Table 14 displays the rank ordering of validity coefficients for the localization judgment. The rank ordering of the validity coefficients were not as orderly as found in the presence/absence judgment (see Table 10). The major difference between the rank ordering of the validity coefficients for the presence/absence judgment and the localization judgments was due to the performance of novice #1 who used a high linear policy for localization judgments (although this judge exceeded the ecological policy, the judge's validity coefficient is inflated because it is based on only 34 protocols. See Table 7, and Table 11 which explains the constraints as to which judgments were used in the localization analysis). Nonetheless, the major findings were as follows: (a) In all cases, the most accurate linear model outperformed its respective most accurate judge. (b) In all cases, the least accurate linear model outperformed its respective least accurate judge. (c) There was a general tendency for the Majority and Composite judges to outperform any individual judge (the exception was novice #1) (see Tables 13 & 14). (d) The unit weights model scored in the top middle tier, suggesting that simply weighting the stated predictor cues with + 1/4 (i.e., Similarities, right hand finger tapping, DRVERB and FAS) and - 1/3 (i.e., Block Design, left hand finger tapping, DRVIS) produced an r_m that outperformed most of the linear models and judges.

Table 14

Rank Ordering of Validity Coefficients for the Localization Judgment

Localization judgment	Validity Coefficient[a]
Most accurate model (novice # 1)	0.74
Linear predictability (R_e)[b]	0.70
Most accurate Composite model judge (novice)	0.66
Most accurate Composite judge (novice)	0.65
Most accurate judge (novice # 1)	0.63
Unit weight model[c]	0.62
Most accurate Majority model judge (novice)	0.61
Least accurate Composite model judge (expert)	0.61
Most accurate Majority judge (novice)	0.59
Least accurate Majority model judge (expert)	0.58
Least accurate Composite judge (expert)	0.52
Least accurate Majority judge (expert)	0.51
Least accurate model (expert # 1)	0.51
Least accurate judge (expert # 1)	0.34

[a] Validity coefficient refers to r_a and r_m.

[b] In this study, R_e is equivalent to an actuarial formula (i.e., the criterion is equal to the linear combination of the nine predictor variables).

[c] Given the coding of the localization judgment, as expected a negative correlation resulted for the unit weight model, but for clarity of comparison purposes a positive value was displayed in the table.

Ecologically Valid Policy

The final data analysis for the localization judgment involves comparing the standardized beta weights on the ecological side and the human judgment side of the Brunswik Lens Model. This comparison will determine the relative importance of the cues in relation to the actual criterion and in relation to each judge's judgments, therefore examining how closely the judge captured the ecologically valid policy. The standardized beta weights of the nine predictor cues from the ecological side of the Brunswik Lens Model were computed on the actual criterion (hit rate of 100%). The standardized beta weights from the human side of the Brunswik Lens Model were computed on each judge's judgments, therefore, with varying hit rates (see Table 12). In principle, as the judge's judgments approach a hit rate of 100%, his/her standardized beta weight values will mirror those from the ecology.

Table 15 shows the ecologically valid policy. The delayed recall trial of the Logical Memory subtest was the most important cue to be weighted in the determination of the localization judgment (beta= -0.445). The next cue of importance was Block Design with a beta weight of -0.351. Two cues received weightings of -0.2 or greater: FAS and Finger Tapping right hand. Three cues had a weighting of -0.1 or greater: DRVIS, Finger Tapping left and Similarities. The two cues with the least importance, i.e., beta weight less than -0.1 were: Digit Symbol and Trail B.

In terms of the four most important cues in the ecology, experts and novices generally approximated the relative importance of DRVERB, Block Design and Finger Tapping right hand. Novices, more so than experts, matched the relative importance of FAS. For the three least important cues, the experts overestimated the importance of Trail B, while novices tended to provide lower weightings. Alternatively, novices generally overestimated the importance of Digit Symbol and Similarities, while experts generally matched the low weighting of these cues found in the ecology.

Finally, judges' subjective weightings of the nine predictor cues (see Table 5) were compared to their weightings obtained from regression analysis (see Table 15). Overall, there was a low association or relationship between a judge's subjective weighting of the relative importance of a cue in relation to the judgment and the cue's beta weight. In general, experts subjectively overestimated the importance of the cues (evidence by ratings of 4 to 6) in comparison to the beta weights (where there were high, moderate and low beta weights). Novices tended to provided a greater range of subjective weights to the cues than the experts, but their subjective weights generally did not reflect the cues beta weight based on their mathematical judgment policy. Therefore, judges' discrepancy between their subjective weights and beta weights of cues indicated that, in particular, judges did not weight cues as they subjectively estimated, and, in general, were not fully and accurately aware of their cognitive processes in relation to the judgment task (cf. Nisbett& Wilson, 1977).

Table 15

Standardized Beta Weights for the Nine Predictor Cues for the Ecological Side and Human Judgment Side of the Brunswik Lens Model for the Localization Judgment

	Ecology	Experts					Novices				
Cues		1	2	3	Maj.	Comp	1	2	3	Maj.	Comp
Similarities	.125	-.033	.011	-.312	-.157	-.092	.056	-.377	-.300	-.233	-.207
Block Design	.351	.379	.541	.164	.328	.412	.172	.275	.407	.269	.336
Digit Symbol	-.019	-.176	.038	-.025	.150	-.024	.086	.301	-.030	.346	.150
Trail B	.003	-.176	-.075	-.115	-.224	-.142	-.176	.051	-.210	.056	-.108
Tapping Right	-.265	-.337	-.362	-.144	-.429	-.320	-.575	-.434	-.134	-.361	-.399
Tapping Left	.145	.227	.341	.245	.395	.311	.324	.315	.115	.226	.285
DRVERB	-.445	-.467	-.446	-.197	-.353	-.440	-.421	-.184	-.169	-.348	-.354
DRVIS	.174	.227	.095	.476	.158	.335	.231	.233	.168	.200	.268
FAS	-.270	.168	-.020	-.349	-.084	-.112	-.251	-.096	-.194	-.118	-.209

Judges' Subjective Ratings of Judgment Task

Upon receipt of the each judge's judgments, a brief follow-up questionnaire was sent to the judge to assess a few aspects of the task (see Appendix E). Typically, the questionnaire was mailed-out 1 to 2 days after receipt of the judge's judgments.

The first item in the questionnaire requested that judges estimate the time it took to complete the judgments on the 50 protocols. The three novices estimated that it took 3, 3 and 4 hours respectively, while the three experts reported 1.5, 3 and 4 hours, respectively.

The second and third items requested that the judge use a 7-point scale (1=not at all confident to 7 =very confident) to provide a mean rating and a range rating of how confident he/she was making the presence vs absence judgment and the localization judgments. As Table 16 shows, there appears to be no notable differences among the two groups in terms of the subjective level of confidence in making the judgments. In addition, each judge's mean rating of confidence across the four judgments was relatively stable (the mean rating was within a two-point range across the four judgments for each judge). This finding suggests that judges did not experience one of the judgments as notably more difficult than another.

Of note was the wide range in the level of confidence (judges had a 2 to 7- point spread in confidence ratings). That is, regardless of the judgment being rendered, all of the judges noted thinking and/or feeling very confident, moderately confident and not at all confident making a judgment depending on the protocol.

Table 17 displays the judges' subjective estimate for correctly judging the protocols (in percentages). Expert judges tended to more accurately estimate their actual hit rate for the presence/absence judgment than novices (see Tables 7 and 17). All of the expert judges and two of the novice judges notably over estimated their ability to correctly judge right, left and diffuse hemisphere brain damage (see Tables 12 and 17). Novice #2 most closely estimated his actual ability to correctly identify the three brain damaged protocols.

Table 16

Judges' Subjective Confidence Ratings in Making the Four Judgments[a]

	Experts			Novices		
	1	2	3	1	2	3
Judgment						
Presence vs Absence						
Mean (range)	5.5 (4-7)	3 (1-5)	5 (1-7)	5 (1-7)	3 (1-7)	5 (2-7)
Localization						
Mean (range)						
RHBD	5.5 (4-7)	2 (1-5)	4 (1-7)	6 (4-7)	3 (1-5)	4 (2-7)
LHBD	5.5 (4-7)	3 (2-6)	6 (1-7)	6 (4-7)	2 (1-4)	5 (2-7)
DBD	5.5 (5-7)	4 (2-6)	5 (1-7)	4 (1-7)	2 (1-4)	5 (2-6)

[a]Confidence scale=1 (not at all confident) to 7 (very confident).

Table 17

Judges' Subjective Estimate of Protocols Correctly Judged

	Experts			Novices		
	1	2	3	1	2	3
Judgment						
Presence vs Absence						
	90%	75%	75%	90%	75%	90%
Localization						
RHBD	80%	50%	70%	90%	60%	80%
LHBD	80%	60%	60%	90%	40%	80%
DBD	80%	75%	75%	80%	60%	80%

DISCUSSION

The purpose of this study was to utilize the Brunswik Lens Model in order to compare the accuracy of the judge versus the accuracy of the linear model of the judge in decisions regarding the presence/absence of brain damage and the localization of brain damage. Such a study has not been published in the area of neuropsychology; although, in clinical psychology, a wealth of experiments have been published which have provided strong support for the phenomenon that a simple linear model of the judge typically is equal to or superior to the judge in many types of clinical decision and judgment tasks (Dawes, Faust, & Meehl, 1989; Sawyer, 1966).

The results from this study clearly showed that the linear model of the judge was equal to or superior to the judge in the presence/absence and in the localization judgments. In fact, in none of the comparisons (i.e., r_a vs r_m) was the judge's accuracy notably or meaningfully higher than its counterpart linear model (the greatest disparity, in favor of the judge, was -0.04). Therefore, the relative superiority of statistical models (i.e., a linear model of the judge) over the judge so staunchly demonstrated in clinical psychology over the past 25 years has been extended into the area of clinical neuropsychology. In other words, the method of bootstrapping did "pull the judges up by their bootstraps" and enhanced their accuracy. Thus, the first hypothesis was supported.

The second hypothesis purported that there would be no meaningful differences between expert and novice judges in their accuracy regarding the two judgments. The data showed that the experts slightly outperformed the novices in their overall hit rate for the presence/absence judgment, while the novices slightly outperformed the experts in the overall hit rate for the localization judgment. Overall, there were no outstanding differences between the two groups. Therefore, the hypothesis of no meaningful differences between the expert and novices groups was supported. The finding of no meaningful differences between the two groups in judgmental accuracy is congruent with a recent review of the literature on training and experience in association with clinical judgment (Garb, 1989). That is, Garb reported that professional training and experience tend not to be associated strongly with judgmental accuracy. This issue will be explored in more detail later.

Now that a few brief statements have been advanced concerning the hypotheses of this study, a more detailed discussion will ensue. First, the neuropsychological data comprising the protocols will be examined. Specifically, it will be shown that the neuropsychological protocols that represented each of the four groups were consistent with conventional neuropsychological principles and empirical data concerning hemispheric specialization. In addition, the correlations between the predictor cues and ecological criteria will be evaluated. Second, judges' accuracy, hit rates and overall achievement are reviewed. Third, the validity coefficients will be explored. The validity coefficients of the judge, linear model of the judge, the unit model judge and the aggregate judges will be examined. Fourth, the expert - novice issue and professional training implications based on this study's findings will be discussed. The fifth section considers the limitations of this study. Finally, directions for future research will be provided.

Neuropsychological Protocols.

Appendix G presents the means and standard deviations of 27 of the 29 possible cues (occupation and gender were not quantified). Examining the 25 test scores for the normal group it is clear that this so called normal group generally achieved test scores in the average range of cognitive functioning on the majority of the tests. In addition, the normal group's mean scores outperformed (not always statistically) the other three groups on 17 of the 25 tests.

As was reported in the Method section, rather stringent criteria were used in selecting the three brain damaged sets of protocols. The obvious goal was to provide judges with representative protocols of right, left and diffuse brain damage in which to render judgments. Examining the neuropsychological data from each of the three sets of brain damage protocols it is apparent that the right hemisphere and left hemisphere brain damaged groups' test scores generally conformed to conventional neuropsychological principles and empirical data concerning hemispheric specialization (see Lezak, 1983; Kolb and Whishaw, 1990).

The test scores from the diffuse brain damaged group appeared to be more similar to the normal group than to the right or left hemisphere groups. That is, the diffuse group did not seem as impaired as the other two brain damaged groups. Specifically, the normal group outperformed the right and left hemisphere brain damaged groups in 19 of the 25 comparisons, while the diffuse brain damaged group outperformed the right and left hemisphere brain damaged groups in 15 of the 25 comparisons. Intuitively, this finding indicates that the protocols used to represent the diffuse group were not as severely brain damaged as the protocols used to represent the other two brain damage groups. Nonetheless, the normal group outperformed the diffuse brain damaged group on 17 of the 25 comparisons.

A subjective severity appraisal of neurological insult for the three different brain damaged protocols indicate that the individuals comprising these groups probably sustained a "mild," "mild to moderate" or "moderate" neurological insult ("mild" brain injury can be conceived of as scores 1 SD below the mean and "moderate" brain injury can be considered as scores 2 SDs below the mean). These levels of neurological and neuropsychological impairments are probably most frequently seen by neuropsychologists (people who sustain severe impairment either receive a superficial screening evaluation or the neuropsychologist waits for the person to recover to a level that allows for a more comprehensive assessment).

Overall, the four sets of neuropsychological protocols upon which the two judgments were rendered were generally representative of conventional neuropsychological theory in terms of how right, left and diffuse brain injury typically affects neuropsychological test scores. The neuropsychological test scores across the three brain damaged groups suggest that the individuals sustained mild to moderate brain damage (based on their test score at the time of the neuropsychological assessment). (A more detailed discussion of these issues appear in Appendix H).

Correlation Between the Cues and the Criteria:

A complete discussion of the correlations between the cues and the criteria is not a vital issue regarding the purpose of this thesis. Therefore, this topic will be reviewed briefly below and the reader is referred to Appendix I for a more in-depth examination.

Four of the nine predictor cues significantly correlated with the presence/absence criterion: Digit Symbol, right hand Finger Tapping, delayed trial of the Logical Memory subtest of the WMS-R (DRVERB) and the delayed trial of the Visual Reproduction subtest of the WMS-R (DRVIS). Four other cues significantly correlated with the criterion: Trails B, PIQ and the immediate recall trials of the Logical Memory subtest and the Visual Reproduction subtest.

The delayed recall trial of the Logical memory subtest of the WMS-R and the FAS test were the only two of the nine predictor cues that significantly correlated with the localization criterion. The Information and Digit Span subtests, and the immediate recall trial of the Logical Memory subtest also significantly correlated with the localization criterion. The Information subtest has not been found to be especially sensitive to brain injury (unless the person is aphasic), while the Digit Span subtest and immediate recall trial of the Logical Memory subtest are moderately sensitive measures.

An important variable to consider in the interpretation of the correlation matrix of the predictor cues and the two criteria is the neuropsychological data on which the correlations were based. As was stated in the section above, because of the nature of the design elements in this study, neuropsychological data associated with "severe" brain injuries were probably not consistent with the protocols used. Therefore, in theory, tests scores associated with progressively more severe brain insults were not indicative of the neuropsychological data in this study. Thus, because the brain damaged groups were not representative of a full range of severity (i.e., mild, moderate, severe), the neuropsychological test scores were restricted and the resulting correlations were probably attenuated (Nunnally, 1978).

Analysis of Judgmental Accuracy

Judges Hit Rates:

Experts' hit rates for the presence/absence averaged 79% (range = 76% to 80%) while novices averaged 75% (range = 72% to 78%). Given that forty of the fifty protocols represented the presence of brain damage, the base rate judgment equaled 80%. The expert and novice groups did not achieve the base rate level. Although, individually, two of the experts did achieve the base rate level. No judge surpassed the base rate. Meehl and Rosen (1955) mathematically proved that as the base rate departs from 0.50 it will become more and more difficult for the judge to make judgments that are more accurate then the base rate. The base rate level in this study was somewhat extreme (i.e., 10 absence and 40 presence; or 80%), nonetheless judges were provided with explicit information about how the groups were formed, the etiologies of the brain damaged protocols and the base rates themselves.

Experts' hit rate for the localization judgment averaged 42% (range = 37.5% to 50%), while novices averaged 47% (range = 40% to 50%). The base rate judgment equalled 50% (see page 71 of the Results section). Neither group matched or surpassed the base rate judgment. Individually, one expert and two novices equalled the base rate level, but no judge surpassed it.

Faust et al. (1988) found that their neuropsychological judges average an overall hit rate of 80% (ranging from 50 % to 94%) distinguishing normal from brain damaged protocols. Judges achieved an overall accuracy rate of 54% for identifying the general location (defined as judges ability to note any quadrant (i.e., right, left, anterior or posterior) where the brain lesion occurred) and an accuracy rate of 29% judging the exact localization (defined as judges ability to note only the primary lesion site and not note adjacent lobes where involvement was possible). (It is important to note that because each judge provided judgement(s) on only one protocol, the accuracy of judges' decisions could not be compared to a base rate value.)

Wedding (1983) requested that judges (i.e., psychologists, graduate students and one expert neuropsychologist) classify 30 protocols into five diagnostic groups: left damage, right damage, diffuse damage, schizophrenic and normal. Judges were provided with information concerning how the samples were drawn and base rates. Judges overall hit rate averaged 55% and ranged from 33% to 70%. Re-analyzing Wedding's data assuming that judges were simply requested to rate the presence vs absence of brain damage per protocol, the judges would have achieved an 88% hit rate (range=73% to 93%). (The base rate would have been 80%: 24 brain damage protocols and 6 normal protocols.)

Heaton et al. (1978) evaluated neuropsychological judges' ability to distinguish neuropsychological protocols from a malingered vs actual brain damaged sample. Judges rated 32 protocols and were not provided with base rates (the base rate was 0.50: 16 brain damage protocols and 16 malingered protocols), although they were told that some of the protocols were from malingerers and some were from actual brain damaged individuals. Results showed that the judges correctly classified 50% to 69% of the protocols.

It is not possible to directly compare the accuracy of judges hit rate in this study to the other published studies, because in two of these studies the judges were not provided with base rate data. But examining judges hit rate across studies, regardless of base rates, it appears that the judges in the present study were comparable to judges in Faust et al.'s (1988) study, were outperformed by Wedding's judges (assuming that the judges were simply requested to make presence vs absence judgments), and were outperformed by Heaton's judges (although Heaton's judges were not given the base rates, the judges' hit rate were at or above the base rate level of .50).

Analysis of Validity Coefficients:

Judge versus Linear Model of the Judge

This study has demonstrated the equality to superiority of the linear model of the judge compared to the judge, documented across 35 years of research, in the area of clinical neuropsychology. It is important to remember that bootstrapping is not a purely statistical decision making strategy as in discriminant function analysis. Rather, bootstrapping is inherently and directly tied into the judge's judgments. As Kleinmuntz (1990) put it, bootstrapping is "a combined use of head and formulas" (p. 301), meaning that first the judge ("head") supplies a judgment and second a model ("formula") of that judge is mathematically created. It is equally important to recall, as was pointed out in the Introduction section, that the linear model is not to be confused with an isomorphic representation of the judge. Rather, the linear model is conceived of as a paramorphic representation of the judge. This means that the linear model is not completely accounting for all of the internal operations and human judgment processes of the judge but, instead, the linear model is best conceived of as a mathematical simulation of the judge's decision making process. In this study, the mathematical simulation is not especially sophisticated, that is, a few test scores are used in a simple regression equation to predict a criterion.

The equality to superiority of the linear model becomes more apparent when considering the fact that the judge made his/her judgments based on all cues, while the linear model used only nine cues. It can be inferred that the fact that judges had access to three times as many cues as the linear model in the judgment process gave judges a clear advantage and may well have resulted in higher r_a relative to r_m values (assuming, of course, that the additional cues were valid and provided relatively nonredundant information). Alternatively, having to integrate data from so many cues may have confused judges' judgments and/or led to inconsistent decision making strategies. Therefore, perhaps so many cues disadvantaged the judge relative to the linear model. The fact that the linear model was equal to superior to the judge demonstrates the "judgmental power" (so to speak) of a linear model and suggests that judges may have committed several errors: (a) Judges may have subjectively overweighted or underweighted some cues which would have lowered r_a as compared to the cues weighting based on beta weights from regression analysis (which produced r_m); (b) Judges may have used too much of a configural component in the decision making process; (c) Judges may have generated a correct strategy to their decision making, but inconsistently utilized this strategy. Each of these explanations will be explored below.

The first explanation suggests that r_m was equal to or greater than r_a because judges over- or under-emphasized cues in relation to their actual beta weighting via regression analysis. Due to the nature of the study, beta weights were computed only on the nine cues that were apriori chosen as the predictors in the linear model, while judges provided subjective weights to all cues (see Appendix H). Therefore, a complete examination of this issue is not possible. In principle, the extent to which the judges' subjective weights over- or under-emphasized the cues in relation to their beta weighting (assuming that beta weights were computed on all cues) explains, in part, the superiority of the linear model over the judge.

The second explanation as to why the linear model outperformed the judge considers the extent of linear and nonlinear processes employed. In the data analysis section of this paper, the "absolute" and the "relative" contributions of linear and configural processes to judgmental accuracy were examined. It was shown that the absolute and relative contributions of the linear component contributed notably more than the absolute and relative nonlinear (i.e., configural) component to the value of the validity coefficient of the judge. Therefore, the extent to which the judges employed a greater linear component to judgmental accuracy relative to the nonlinear component strongly influenced their validity coefficient (i.e., r_a). This is not to imply that the nonlinear process was unimportant or meaningless. In fact, the data showed that a small or modest nonlinear component to judgmental accuracy contributed to the validity coefficient of the judge. But, overall, a substantial linear component in combination with a small nonlinear component contributed to the ranking of the judges' validity coefficients. Thus, the more the judges deviated from a substantial linear component and small nonlinear component to judgmental accuracy the more likely they lowered their validity coefficient.

A related issue is whether the apparent meaningfulness of the judges' nonlinear component (i.e., C) based on their r_a values adds a substantial component to judgmental accuracy over the linear component of judgmental accuracy (i.e., G) based on their "bootstrapped" model. The validity coefficient of the judge (i.e., r_a) is computed using equation [1] (see page 14) which considers linear and nonlinear components of judgmental accuracy. The linear model of the judge (i.e., r_m) is based on a much simpler formula (equation [3]. Page 61) and considers only the linear component of judgmental accuracy. As was previously described, the linear model of the judged was equal to or superior to the judge. Therefore, the apparent meaningfulness of the judges' nonlinear component does not add a significant component to judgmental accuracy over-and-above the linear component of the judge.

The third explanation of why the linear model was equal to or superior to the judge had to do with the possibility that the judges had the correct strategy (e.g., weighted cues similar to the ecologically valid policy, or employed a substantial linear component to judgmental accuracy), but inconsistently employed such a strategy. This issue can be explored in two ways. (a) The judges' weighting of the cues can be compared to the ecological policy. The extent to which the judges' weighting of cues differs from the weightings in the ecology indicates: that the judges employed the incorrect policy, or the judges employed the policy inconsistently. (b) The extent to which each judge's r_a value is lower than his/her r_m value indicates that the judge inconsistently employed his/her nonlinear component of judgmental accuracy. Specifically, if a judge obtained a lower r_a value compared to his/her r_m value then the judge must have inconsistently employed their nonlinear policy or employed an invalid nonlinear component, because a major difference between r_a and r_m in the contribution of the nonlinear component. Alternatively, when r_a was greater than r_m then the judge must have employed the correct nonlinear component and applied it somewhat consistently, again, because the major difference between r_a and r_m is the contribution of the nonlinear component.

The data showed that r_m outperformed r_a in 15 of the 20 comparisons. Therefore, judges either employed an invalid nonlinear component to judgmental accuracy, employed the nonlinear component inconsistently, or inconsistently used the correct linear policy. Dudycha and Naylor (1966) examined the issue of judges generating correct strategies, but applying them inconsistently. They concluded that judges are very capable of determining the proper strategy, but are notoriously inconsistent in applying their own correct rules. From their analysis, Dudycha and Naylor (1966) conclude that once the judge has determined the correct strategy, he/she should be replaced with a model that will follow his rules consistently.

Rank Ordering of Validity Coefficients:

Validity indices from the presence/absence and localization judgments were generally of the following pattern: (a) typically the most accurate linear model outperformed its respective most accurate judge, (b) the least accurate model outperformed its respective least accurate judge, (c) an aggregate judge tended to outperform any individual judge, and (d) the unit weight model scored in the upper tier of the validity indices. These findings are consistent with data presented by Goldberg (1970) and Wiggins and Kohn (1971).

The unit weight model was in the upper tier of the rank ordering of the validity coefficients for both judgments. The fact that the unit weight model outperformed most of its competitors supports Dawes and Corrigan's (1974) and Einhorn and Horgarth's (1975) review of the literature on optimal vs unit weighting as well as their rationale for this finding. These issues were explored in the introductory section of this paper and will not be repeated here. The implication is that a unit weight model can be constructed based on previous empirical data and neuropsychological theory, substituted for many of the judges and linear models of the judges and produce a higher validity index (i.e., more accurate judgments). In addition, the unit weight model is much easier to compute than a model based on a judge, because for a unit weight model the researcher simply substitutes the appropriate weights in the regression equation rather than requiring one or more judges to make predictions and then using regression analysis to produce optimal weights.

Aggregate judgments:

Two aggregate judges were created in this study: a Majority and Composite judge (the reader is referred to the Method section for a definition of these terms). Essentially, the Majority judge was created by using a "majority rules" decision criterion, while the Composite judge was created by constructing an arithmetic mean of the judges' judgments. For the presence/absence and localization judgments, the Majority and Composite indices (based on r_a and r_m values) typically outperformed the accuracy of a single judge in both the expert and novice groups. This finding is consistent with results from others studies (Goldberg, 1970; Wedding, 1983). In addition, a rather comprehensive review of group vs individual performance differences (Hill, 1982) indicated that groups typically outperformed individuals, qualitatively and quantitatively.

Although, there were exceptions to this general finding, nonetheless, Hill's conclusion about group vs individual performance was robust.

Although the validity coefficients (i.e., r_a and r_m) of the Majority and Composite indices were similar, the Composite index outperformed the Majority index in six of the eight comparisons (see Tables 8 and 13). This suggests that if the goal is to aggregate the judgments of several judges, a simple averaging procedure, rather than a majority rule procedure, will enhance accuracy. In principle, the Composite index by its computation may have an a priori advantage (also an a priori disadvantage) over the Majority index, because it [the Composite] allows for a wider range of "values" (i.e., 0, 1/3, 2/3, 1) in correlation and regression analyses than the Majority index (i.e., either 0 or 1).

The practicality of aggregate indices of judgments is tenuous in clinical settings. Goldberg (1970) and Wedding (1983) point out that because professional time is valuable, it is atypical that two or more neuropsychologists will confer on a case. Although, if the goal of a program was to develop an equation to use as a model to compare a clinician's judgment, then the data indicate that constructing a majority or composite index would be beneficial. Even if such a situation were possible, the resulting judgment may not be more valid. For example, if an inaccurate judge was vocal and/or adamant about his/her opinion which wrongfully swayed the views of others, then the outcome judgment would be less accurate (Wedding, 1983). Alternatively, if the judgments were made privately, without public discussion, then averaging the judgments should lead to more accurate judgments.

Experts vs Novices: In Search of Differences

This study found no meaningful differences between expert neuropsychologists (i.e., diplomates in clinical neuropsychology) vs novice neuropsychologists (i.e., postdoctoral psychologists in neuropsychology) in their accuracy in identifying presence/absence protocols and in identifying right, left and diffuse hemispheric brain damage. In addition, there were no notable differences between the two groups in their subjective reports of confidence in the judgments and their actual hit rate (see Tables 7, 12 and 14). Novices tended to subjectively over estimate the percentage of presence/absence protocols correctly identified compared to their actual hit rate as opposed to the experts who were on target in their estimate (see Tables 12 and 15). For the localization judgment, experts and novices notably over estimated their actual hit rate (see Tables 12 and 15).

Previous research in neuropsychology on expert vs novice differences support the findings in this study. That is, that training and experience have been found not to significantly correlate with accuracy of judgment. Faust et al. (1988) asked psychologists with varying amounts of training and experience in neuropsychology to make up to five judgments on a neuropsychological protocol (i.e., presence/absence of brain damage, functional vs cortical factors contributing to the abnormal data, the cortical areas involved in the brain damage, whether the brain damage was static or progressive, and a judgment of the disorder causing the brain damage) with a standard array of neuropsychological test scores and demographic information. The researchers then correlated trainee experience, completion of an internship in neuropsychology, completion of a fellowship in neuropsychology, supervision hours received in neuropsychology, courses taken in neuropsychology, years of practice in neuropsychology and clinical experience in neuropsychology with judges decisions on the five judgment variables. The resulting correlations were found to be low (i.e., less than 0.22) and had negligible influence on the judges' judgments. Next, Faust et al. created more extreme groups based on judges extent of training and experience in neuropsychology to determine if the results would change. Again, it was found that there were minimal differences in judgmental accuracy as a result of training and experiential factors in neuropsychology.

Wedding (1983) compared the performance of ten psychologists with experience in neuropsychology, three graduate students with training in neuropsychology and one expert in neuropsychology in making common neuropsychological judgments (i.e., localization of brain damage, etiology of brain damage and acuteness of the disorder). The results showed that the accuracy in judgments were not significantly correlated with clinical experience and experience with the Halstead-Reitan Battery. Judge's level of confidence also was not found to be significantly related to the judgments.

Heaton et al. (1978) asked 10 judges to rate 32 neuropsychological protocols as to whether or not the protocols represented data obtained from malingerers or from people who had real brain damage. Judges' experience in neuropsychology ranged from 8 weeks to 18 years. Results showed that experience and confidence ratings did not significantly correlate with accuracy of judgments.

Garb (1989) reviewed the literature on the impact of clinical training and professional experience in relation to clinical judgments. His literature review involved studies that compared experts vs novices, experienced clinicians vs inexperienced clinicians, graduate students vs clinicians, and graduate students with varying years of training and clinicians vs lay judges on a wide variety of judgments tasks. The studies examined: (a) how judges used projective tests, the MMPI and others tests, in making diagnostic judgments and in making personality ratings; and (b) how judges used neuropsychological data in the judgment of organic brain damage or to classify protocols into brain damaged categories. His conclusions, most relevant to this discussion, were that: (a) Overall, experience was not found to be significantly related to making valid judgments, while training was found to be somewhat related to valid judgments in some of the comparisons; (b) In the area of personality assessment (e.g., making diagnoses, personality ratings), experienced clinicians were not more accurate than less experienced ones; and (c) In the area of neuropsychology (e.g., judge organic brain damage, classify protocols into brain damaged categories), experts were more accurate than nonexpert psychologists, but experienced clinicians were no more accurate than inexperienced clinicians.

On the positive side, experienced clinicians tended to make more accurate confidence ratings than inexperienced clinicians. Also, training and experience tended to enhance a clinician's ability to more effectively and/or efficiently structure problems and identify important variables.

Overall, Faust et al. (1988), Garb's (1989), Heaton et al. (1978), and Wedding (1983) presented findings that strongly support a position that experience and training are not significantly related in accuracy of judgments. Probably most psychologists and neuropsychologists believe that years of experience and proper training will improve judgmental accuracy. But, the overwhelming evidence based on the research cited above suggests not.

Certainly, there is something to being an expert that may be expressed and measured in other domains not examined in this study. Certainly, there is value in training and experience beyond pedagogy. Cognitive psychology has studied expert vs novice differences in solving problems (e.g., physics problems, "mind teasers", chess problems, and making diagnoses from x-ray films). Research has clearly identified processes that experts have used that validate their expertness over novices in physics, chess and in medicine (see Lesgold, 1988). But, in psychological and neuropsychological judgments, research has not tended to objectify this concept of expertness, especially as it relates to judgmental accuracy. Roger C. Schank, a well renowned cognitive scientist, writes in his unusual but intriguing book, "The Connoissueur's Guide to the Mind" (Schank,1991), that "Real experts are just individuals with collections of experiences and the ability to find those cases when they need them to help them solve new problems" (p. 143). Schank assumes that experience is a valid teacher, but as will be explored below, experience can often lead to errors in learning.

Brehmer (1980) offers explanations from research findings to account for experience as a variable that is often unremarkable and at times lead to erroneous decisions in accuracy of judgment studies. He indicated that lay judges and professional clinicians utilize a number of biases in their decision making process that interferes with experience as a valid teaching tool. Specifically, judges of all kinds tend to employ confirmatory rather than disconfirmatory hypothesis testing. That is, it is difficult to test alternative explanations of phenomenon unless one is searching for the alternative and not just searching for one's pet explanation. A second error judges make, similar to the first, is using tests or instruments that tend to confirm the inference under study, instead of using procedures that may provide information contradictory to the inference. Therefore, if a neuropsychologist develops a hypothesis about a client's assessment data and seeks to validate that hypothesis using a confirmatory rather than disconfirmatory strategy, then the hypothesis may be confirmed not because it was valid, but because it happened to be right this time. Thus, experience teaches, but not always validly. A third bias that interferes with experience being a valuable learning tool, is the tendency for people to think about data or information in a deterministic or causal manner, rather than in a probabilistic manner. This predilection for determinism often results in the judge utilizing ineffective decision strategies, making inaccurate judgments or misunderstanding the phenomenon under study. Therefore, experience tends not to improve judgmental accuracy, because judges probably do not utilize the correct strategy of integrating information (i.e., a probabilistic rather than deterministic strategy).

Garb (1989) also presented explanations accounting for professional experience as a nonsignificant contributor to judgmental accuracy. Many of Garb's explanations are consistent with Brehmer's, but one factor of judgmental accuracy that Garb stressed was the use of feedback. A significant impediment to learning from experience is that feedback about judgments are often unavailable or biased. For example, a neuropsychologist may use assessment information to recommend that a client return to work. But, if the neuropsychologist does not follow up on the success or lack thereof of the client's performance at work, he/she cannot determine the validity of the assessment information in relation to this judgment. In addition, if the neuropsychologist does not know of the base rates for success in returning to work, then his/her judgment may be less valid then a base rate judgment.

There are also other classic biases (e.g., availability and representative heuristics) that have been identified by judgment and decision making researchers (see Fischhoff, 1988) that have been demonstrated in hundreds of experimental situations with lay and professional judges to distort or misrepresent information obtained from experience.

Garb (1988) showed that training is a variable that can, but not always significantly, correlate with accuracy of judgment. For example, psychologists usually outperform lay judges in judgments concerning diagnosis and personality variables. But, once the two sets of judges begin to move from the extreme ends of the continuum and involve somewhat more of an equitable comparison (e.g., novices in psychology vs experts in psychology; graduate students vs psychologist) the training variables tends to become unrelated to accuracy of judgments.

What does it mean to be an expert in neuropsychology? Division 40, APA, indicates that it is that person who has obtained the diplomate status (i.e., ABPP/ABCN); meaning that the person has passed a peer evaluated examination on several professional issues involved in neuropsychology. Probably most neuropsychologists would agree that experts as compared to novices have read more books in neuropsychology, know how to administer more tests, have completed more assessments, know more about neuroanatomy, have published more articles in neuropsychology and have attended more neuropsychological conferences. Yet, all of these variables do not seem to be related to making more accurate judgments, at least in terms of how judgmental accuracy has been studied.

Professional Training Implications:

The methodology and statistical analyses (i.e, Brunswik Lens Model and its mathematical formulations) can be used to study judgment and decision making strategies in order to train neuropsychologists in the use of linear components, nonlinear components and the relative importance of predictor cues to various decisions.

There was a tendency for experts to be slightly more accurate in the presence/absence judgment (they employed a higher linear component to judgmental accuracy and generally used a higher nonlinear component than novices), while novices were slightly more accurate in the localization judgment (they employed a small to modest size nonlinear component to judgmental accuracy, while most experts did not use a nonlinear component). Because neither groups of judges clearly and meaningfully demonstrated superiority in judgmental accuracy, it is not appropriate to focus on decision processes employed by experts or not employed by novices. One point is clear, that being that the linear model of the judge outperformed the judge in 15 of the 20 comparisons. Therefore, it is probably better to train neuropsychologists to makes models of themselves (e.g., utilizing the methodology and analyses described in this study) and then request that they use their linear model to make the judgments.

A second training implication of the data is that when possible an aggregate judgment will be likely more accurate than an individual judgment. Therefore, consultation to other neuropsychologists should be encouraged when making judgments. Related to the first point describe in the previous paragraph, the data as well as previous research (Hill, 1982; Goldberg, 1970) indicate that it is advantageous to have a few neuropsychologists make judgments and then build a model of them to use for future judgment purposes.

The results also suggest another training implication, that being to use a linear model rather than a nonlinear model. A linear component implies that a cue with respect to a criterion is constant. For example, as the score for Trails B gets higher and higher the probability of brain impairment is greater, regardless of other test scores. In other words, the judgment (i.e., output) is directly proportional to each cue's value (i.e, input), regardless of its relationship to the other cues. A nonlinear component implies that the value of a cue with respect to a criterion is not constant, but may vary in its "meaningfulness" and/or sign as a function of the other cues. That is, the configural or patterned nature of the cues are explored in combination. For example, a value of Trails B (or any other cue (or input)) does not directly lead to a judgment (output). A value of Trails B is always examined in combination with other cues (e.g., in a configural or patterned manner) in order to generate a judgment (i.e., output).

The eminent Paul Meehl has been examining the merits of linear and configural models to decision making for about 35 years (Dawes, Faust, & Meehl, 1989; Meehl, 1954, 1959, 1986). Overall, it has been consistently shown that a simple linear model is equal to more complex configural models and human judges (see Wiggins, 1981). This does not mean that linear models always outperform configural models or that linear models should always be used. There are clearly cases when a linear model should not be applied. Meehl (1954) presented a humorous example of a linear equation predicting whether a professor will attend a movie after just breaking his leg, signifying the importance of special cases and their effects on a linear equation.

Although it is important to differentially consider special cases in one's judgment strategy, Meehl implicitly states that the frequency of special cases are probably quite rare, and it would be unusual as well as a mistake to disregard the statistical model in the great majority of cases.

A detailed examination of the merits of linear and configural processes are beyond the scope of this thesis. Suffice to say that in this study, and in many other studies (Dawes, et al., 1989; Sawyer, 1966; Wiggins, 1981), a simple linear model is equal to or superior to the judge and configural models. Furthermore, clinicians or neuropsychologists who claim complex configural processes in their judgment process have the responsibility to: (a) demonstrate that the relationship between the cues and criterion is configural, (b) demonstrate that the clinicians can consistently model this configural relationship, and (c) demonstrate that the configural relationship is superior to a simplified linear model.

Limitations and Weaknesses of This Study

Some appropriate criticisms can be leveled against this study. For example, the judgment task was somewhat artificial and/or did not fully mirror how a neuropsychologist operates in the "real world." Judges did not have access to the qualitative aspects of test performance. In addition, judges may have preferred to have test scores from instruments that were not part of the data provided in the protocols.

In defense of the study, judges were provided with information as to how the protocols were gathered and classified. They were given information concerning the various etiologies that caused the brain damage, and they were provided with the base rates. In addition, they were given essential demographic information and neuropsychological test scores based on a relatively comprehensive assessment that tapped general intellectual, memory, motor, visual-perceptual, speech and abstract thinking functions. Furthermore, the two judgments were relatively fundamental or primary in neuropsychological assessment (as compared to a more sophisticated or substantive judgments: can this person safely drive a car?, or can this person successfully return to work?).

Rock, Bransford, Maisto and Morey (1987) reviewed the judgment and decision making literature within the context of the research's ecological validity. These scientists argue that the judgment tasks may have significantly differed on one or more clinical dimensions to which the therapist usually has access when functioning in the "real world." Therefore, the extent to which the experimental judgement task differed from similar tasks in actual clinical practice provides therapists with valid reasons as to why they may not have performed optimally or provides well-grounded explanations why differences have not been identified between experts and novices.

Jenkins (1979) developed the Tetrahedron Model which focuses on four contextual factors found in learning and memory experimental situations. Rock et al. (1987) applied this Model to the area of judgment research. The Tetrahedron Model stresses the importance of four variables in experimental situations: characteristics of subject (e.g., abilities, training, experience), criteria tasks (e.g., diagnosis, treatment plan), characteristics of learning materials (e.g., test scores, interview data, case history, level of difficulty), and information processing activities of subject (e.g., opportunity for feedback, opportunity to request additional information).

Rock et al. (1987) argue that although these four factors are present in probably all judgment studies, the mere presence of a factor does not necessarily enhance the study's ecological validity. The researchers contend that in order to optimize the ecological validity of judgment studies, researchers should ask the participating judges to rate the judgment task as to its ability to simulate real world activities on the four dimensions of the Tetrahedron Model. Conditions that judges perceive as not representative or as hindering their judgment should be corrected if possible. Therefore, the judge and the experimenter collaborate on the structure of the judgment task to optimize the ecological validity of the study and maximize the probability that the study's findings will be clinically meaningful and/or representative of judges' decision making processes.

How did this study do in relations to the Tetrahedron Model? This study clearly provided information about the characteristics of the judges. Level of knowledge was based on obtaining the diplomate status for experts and completing a postdoctoral program within the past 2 years for novices. Number of years of experience was also provided. In terms of the critical task, detailed information was provided as to how the protocols were gathered, selected and classified. Clear and explicit information about the two judgments to be rendered was given. The third factor in the Tetrahedron Model is characteristics of learning material. On the positive side, a relatively comprehensive set of neuropsychological data were provided on each protocol. Judges had access to base rate information. In addition, the data were obtained from real people and not fabricated. On the negative side, judges did not have access to the qualitative aspects of the assessment data, they were unable to observe or interview the person who represented a protocol of data. They may not have been presented with scores from their preferred tests. The fourth component of the model examines the information processing activities of the judges. The information processing activities of the judges were measured using the mathematical indices of the Brunswik Lens Model and requesting that judges make subjective rating as to how much they weighted each cue in their judgment process. But, judges were not asked to provide a real time account of their decision making processes and they were not allowed to follow-up on questions that they may have encountered while making a judgment.

Directions for Future Research

One argument as to why expert/novice differences were not found is that experts are typically expert in one area within neuropsychology. That is, some neuropsychologists work exclusively with epileptic disorders or cerebral vascular disorders. In this study, the 50 neuropsychological protocols were composed from a variety of neurologic disorders. Therefore, the variety of neurologic disorders may have attenuated the experts ability to demonstrate their expertness. There is some support for this position in problem-solving studies. Lesgold (1988) indicated that there is a reasonable amount of data to support a position that expertise is based on a rather specific store of knowledge. Given these issues, a useful study would be to gather neuropsychological protocols from one neurologic disorder (e.g., neuropsychological data gathered on temporal lobe epileptics or people who sustained CVAs) and ask neuropsychologists who are experts in this area to make a lateralization, localization, acuteness or some other kind of judgment and compare their judgments to novice neuropsychologists or expert neuropsychologists who have their expertise in some other area.

Another fruitful line of research would be to obtain an on-line account of what cognitive or decision making processes the judge is using in the process of making a judgment. These decision making strategies can then be compared to a model of the judge using regression analyses. Such a design would assess the validity of the judge's subjective decision making processes.

Clinicians often argue that the more assessment information the better. That is, it is not unusual for a clinician to state -"If only I administered Test X, I would have made a better judgment." While certainly some core amount of assessment information is required in order to respond to a referral question, greater amounts of test scores do not usually improve judgment accuracy (Faust, 1986; Wedding & Faust, 1989). Given these caveats, an interesting research study would be to give neuropsychologists a core set of assessment information consisting of medical, educational and occupational history, demographics, and tests scores from selected neuropsychological procedures. Next, ask them to make a judgement about the brain injury (e.g., static, progressive or localization of damage). Then give them additional assessment information, first asking them what they expect the additional assessment information will reveal and then asking if they would like to change their judgment based on the additional data. Therefore, the design involves providing judges with progressively greater amounts of information interspersed with judgments about the forthcoming data and the outcome judgment. Such a design would evaluate the extent to which more assessment information effects judgmental accuracy and evaluate the judge's hypothesis about what they expect the additional information to reveal.

Finally, continued research in clinical judgment and decision making should incorporate Rock et al.'s (1987) Tetrahedron Model (adaptive from Jenkins, 1979) and design judgment tasks that maximize the ecological validity of the study.

References

American Heritage Dictionary. (1976). Boston: Houghton Mifflin Company.

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. Journal of consulting and Clinical Psychology, 3, 323-330.

Benton, A. L. (1968). Differential behavioral effects in frontal lobe disease. Neuropsychologia, 6, 53-60.

Benton, A. L., & Hamsher, K. deS. (1978). Multilingual aphasia examination. Iowa City: University of Iowa.

Black, D. W., & Strub, R. L. (1976). Constructional apraxia in patients with discrete missile wounds of the brain. Cortex, 12, 87-93.

Borkowski, J. G., Benton, A. L., & Spreen, O. (1967). Word fluency and brain damage. Neuropsychologia, 5, 135-140.

Botwinick, J. (1984). Aging and behavior: A comprehensive integration of research findings. (3rd ed.). New York: Springer.

Brehmer, B. (1980). In one word: Not from experience. Acta Psychologica, 45, 223-241.

Brunswik, E. (1955). Representative design and probablistic theory in a functional psychology. Psychological Review, 62, 193-217.

Butler, M., Retzlaff, P., & Vanderploeg, R. (1991). Neuropsychological test usage. Professional Psychology: Research and Practice, 22, 510-512.

Camerer, C. (1981). General conditions for the success of bootstrapping models. Organizational Behavior and Human Performance, 27, 411-422.

Chelune, G. J., & Moehle, K. A. (1986). Neuropsychological assessment and everyday functioning. In D. Wedding, A. H. Horton, & J. Webster (Eds.), The neuropsychology handbook. New York: Springer Publishing Company.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. American Psychologist, 26, 180-188.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. Psychological Bulletin, 81, 95-106.

Dawes, R. M., Faust, D., & Meehl, P. E. (1988). Clinical versus actuarial judgment. Science, 243, 1668-1674.

Delaney, R. C., Rosen, A. J., Mattson, R. H., & Novelly, R. A. (1980). Memory function in focal epilepsy: A comparison of non-surgical, unilateral lobe and frontal lobe samples. Cortex, 16, 103-117.

Diplomates in clinical neuropsychology. The Clinical Neuropsychologist, 4, 390-391.

Dudycha, L. W., & Naylor, J. C. (1966). Characteristics of the human inference process in complex choice behavior situations. Organizational Behavior and Human Performance, 1, 110-128.

Ebert, R. J., & Kruse, T. E. (1978). Bootstrapping the security analyst. Journal of Applied Psychology, 63, 110-119.

Einhorn, H. J. (1986). Accepting error to make less error. Journal of Personality Assessment, 50, 387-395.

Einhorn, H. J., & Horgarth, R. M. (1975). Unit weighting schemes for decision making. Organizational Behavior and Human Performance, 13, 171-192.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports of data. Psychological Review, 87, 215-251.

Faust, D. (1986). Research on human judgment and its application to clinical practice. Professional Psychology: Research and Practice, 17, 420-430.

Faust, D., Guilmette, T. J., Hart, K., Arkes, H. R., Fishburne, J., & Davey, L. (1988). Neuropsychologists' training, experience, and judgment accuracy. Archives of Clinical Neuropsychology, 3, 145-163.

Finlayson, M. A., & Reitan, R. M. (1980). Effect of lateralized lesions on ipsilateral and contralateral motor functioning. Journal of Clinical Neuropsychology, 2, 237-243.

Fischhoff, B. (1988). Judgment and decision making. In R. J. Sternberg and E. E. Smith (Eds.), The Psychology of Human Thought. New York: Cambridge University Press.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. Psychological Bulletin, 105, 387-396.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence for a method of improving clinical inference. Psychological Bulletin, 73, 422-432.

Goldstein, S. G., Deysach, R. E., & Kleinknecht, R. A. (1973). Effect of experience and amount of information on identification of cerebral impairment. Journal of Consulting and Clinical Psychology, 41, 30-34.

Guilmette, T. J., Faust, D., Hart, K., & Arkes, H. R. (1990). A national survey of psychologists who offer neuropsychological services. Archives of clinical Neuropsychology, 5, 373-392.

Halstead, W. C. (1947). Brain and intelligence: A qualitative study of the frontal lobes. Chicago: The University of Chicago Press.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. Psychological Review, 62, 255-262.

Hammond, K. R., Hursch, C. J. & Todd, F. J. (1964). Analyzing the components of clinical inference. Psychological Review, 71, 438-456.

Hamsher, K. deS. (1984). Specialized neuropsychological assessment methods. In G. Goldstein & M. Hersen (Eds.), Handbook of psychological assessment. New York: Pergamon Press.

Hartman, D. E. (1991). Reply to Reitan: Unexamined premises and the evolution of clinical neuropsychology. Archives of Clinical Neuropsychology, 6, 147-165.

Heaton, R. K., Grant, I., & Matthews, C. G. (1986). Differences in neuropsychological test performance associated with age, education, and sex. In I. Grant & K. M. Adams (Eds.), Neuropsychological assessment of neuropsychiatric disorders. New York: Oxford University Press.

Hill, G. W. (1982). Group versus individual performance: Are N + 1 heads better than one? Psychological Bulletin, 91, 517-539.

Hirschenfang, S. A. (1960). A comparison of WAIS scores of hemiplegic patients with and without aphasia. Journal of clinical Psychology, 16, 351.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. Psychological Bulletin, 57, 116-131.

Horton, A. M., & Puente, A. E. (1986). Human neuropsychology: An overview. In D. Wedding, A. M. Horton, Jr., & J. Webster (Eds.), The neuropsychology handbook. New York: Springer Publishing Company.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-cue probability studies. Psychological Review, 71, 42-60.

Jenkins, J. J. (1979). Four points to remember. A tetrahedron model of memory experiments. In L. S. Cermak & I. M. Craik (Eds.), Levels of processing in human memory (pp. 429-446). New York: Erlbaum.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. Psychological Bulletin, 107, 296-310.

Kolb, B., & Whishaw, I. Q. (1990). Fundamentals of human neuropsychology. New York: W. H. Freeman and Company.

Lesgold, A. (1988). Problem solving. In R. J. Sternberg and E. E. Smith (Eds.), The Psychology of Human thought. New York: Cambridge University Press.

Levin, I. P., Johnson, R. D., & Faraone, S. V. (1984). Information integration in price-quality tradeoffs: The effects of missing information. Memory & Cognition, 12, 96-102.

Lewinsohn, P. M. (1973). Psychological assessment of patients with brain injury. Unpublished manuscript, Eugene, Oregon, University of Oregon.

Lezak, M. D. (1983). Neuropsychological assessment. New York: Oxford University Press.

Mandelberg, I. A., & Brooks, D. N. (1975). Cognitive recovery after severe head injury. 1. Serial testing on the Wechsler Adult Intelligence Scale. Journal of Neurology, Neurosurgery, and Psychiatry, 38, 1121-1126.

Matarazzo, J. D. (1972). Wechsler's measurement and appraisal of adult intelligence (5th ed.). Baltimore: Williams & Wilkins.

McFie, J. (1975). Assessment of organic intellectual impairment. London: Academic Press.

Meehl, P. E. (1954). Clinical vs Statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.

Meehl. P. E. (1957). When shall we use our heads instead of the formula. Journal of Counseling Psychology, 4, 268-273.

Meehl, P. E. (1959). A comparison of clinicians with five statistical methods of identifying psychotic MMPI profiles. Journal of Counseling Psychology, 6, 102-109.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. Journal of Personality Assessment, 50, 370-375.

Meehl, P. E., & Rosen, A. (1954). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychological Bulletin, 52, 194-216.

Miceli, G., Caltagirone, C., Gainotti, G., Masullo, C., & Silveri, M. C. (1981). Neuropsychological correlates of localized cerebral lesions in nonaphasic brain-damaged patients. Journal of Clinical Neuropsychology, 3, 53-63.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 34, 231-250.

Nunnally, J. C. (1978). Psychometric theory. New York: McGraw-Hill Company.

Pedhazur, E. J. (1982). Multiple regression in behavioral research. New York: Holt, Rinehart, and Winston, Inc.

Parks, R. W., Loewenstein, D. A., Dodrill, K. L., Barker, W. W., Yoshii, F., Chang, J. Y., Emran, A., Apicella, A., Sheramata, W. A., & Duara, R. (1988). Cerebral metabolic effects of a verbal fluency test: A PET scan study. Journal of Clinical and Experimental Neuropsychology, 10, 565-575.

Phares, E. J. (1979). Clinical psychology: Concepts, methods, and profession. Homewood, Illinois: The Dorsey Press.

Reitan, R. M. (1955). The relation of the Trail Making Test to organic brain damage. Journal of Consulting Psychology, 19, 393-394.

Reitan, R. M. (1958). The validity of the Trail Making Test as an indicator of organic brain damage. Perceptual and Motor Skills, 8, 271-276.

Reitan, R. M., & Davison, L. A. (1974). Clinical neuropsychology: Current status and applications. Washington, D.C.: V. H. Winston & Sons.

Report of the Division 40/INS Joint Task Force on Education, Accreditation and Credentialing. (1984). Newsletter 40, 2, 3-8.

Report of the INS-Division 40 Task Force on Education, Accreditation and Credentialing. (1986). Newsletter 40, 4, 4-5.

Report of the Executive Committee of Division 40. (1989). Definition of a clinical neuropsychologist. The Clinical Neuropsychologist, 3, 22.

Rock, D. L., Bransford, J. D., Maisto, S. A., & Morey, L. (1987). The study of clinical judgment: An ecological approach. Clinical Psychology Review, 7, 645-661.

Russell, E., Neuringer, C., & Goldstein, G. (1970). Assessment of brain damage: A neuropsychological key approach. New York: Wiley-Interscience.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. Psychological Bulletin, 66, 178-200.

Schank, R. C. (1991). The connoisseur's guide to the mind. New York: Summit Books.

Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum associates.

Tucker, L. R. (1964). A suggested alternative formulation in the developments of Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. Psychological Review, 71, 528-530.

Wampold, B. E., & Freund, R. D. (1987). Use of multiple regression in counseling psychology research: A flexible data- analytic strategy. Journal of Counseling Psychology, 34, 372-382.

Wedding, D. (1983). Clinical and statistical prediction in neuropsychology. Clinical Neuropsychology, 5, 49-55.

Wedding, D., & Faust, D. (1988). Clinical judgment and decision making in neuropsychology. Archives of Clinical Neuropsychology, 4, 233-265.

Wechsler, D. (1981). Wechsler Adult Intelligence Scale-Revised. The Psychological Corporation. New York: Harcourt, Brace, Jovanovich.

Wechsler, D. (1987). Wechsler Memory Scale-Revised. The Psychological Corporation. San Antonio: Harcourt, Brace, Jovanovich.

Weisberg, L. A. (1979). Computer tomography in the diagnosis of intracranial disease. Annuals of Internal Medicine, 91, 87-105.

Wheeler, L. E. (1964). Complex behavioral indices by linear discriminant functions for the prediction of cerebral damage. Perceptual and Motor Skills, 19, 907-923.

Wiggins, J. S. (1981). Clinical and statistical prediction: Where are we and where do we go from here? Clinical Psychology Review, 1, 3-18.

Wiggins, N. & Kohen, E. S. (1971). Man versus model of man revisited: Forecasting of graduate school success. Journal of Personality and Social Psychology, 19, 100-106.

Zangwill, O. L. (1987). John Hughlings Jackson. In R. L. Gregory (Ed.), The oxford companion to: The mind. Oxford: Oxford University Press.

Appendices

Appendix A
Protocol 01

1. Presence vs Absence Judgment (check one): Present_____ Absent_____

2. Localization Judgment (check one): Right____. Left_____ Diffuse_____

Age...
Education..................................
Occupation................................
Gender.....................................

Verbal IQ...............
Performance IQ......
Full Scale IQ..........

WAIS-R (age equivalent scaled scores)

Verbal subtests
Information ....
Digit Span..............:....
Vocabulary ....
Arithmetic...............
Comprehension ....
Similarities............

Performance Subtests
Picture Completion...........
Picture Arrangement........
Block Design.....................
Object Assembly...............
Digit Symbol....................

Trail Making Test (time in seconds) Part A ..... Part B.....

Wisconsin Card Sorting Test (number of categories completed)..................
Category Test (number of errors)...

Finger Tapping, Halstead-Reitan (average number of taps per 10 sec)
Right hand.... Left hand....

Immediate Recall Logical Memory subtest, Wechsler Memory Scale-R....................
Delayed Recall Logical Memory subtest, Wechsler Memory Scale-R..........................
(each score represents the total number of details recalled for both stories)

Immediate Recall Visual Reproduction subtest, Wechsler Memory Scale-R....................
Delayed Recall Visual Reproduction subtest, Wechsler Memory Scale-R.........................
(each score represents the total number of details recalled for all figures)

Controlled Oral Word Association Test (i.e., FAS Test)....................................... .
(total number of words produced for all three letters)

Appendix B

Name
title
address
address

Dear

I am kindly requesting your participation in my doctoral dissertation. The study examines clinical decision making in neuropsychology.

I am requesting your participation because you have achieved the diplomate status, i.e., ABCN, in neuropsychology. I am interested in understanding the decision making strategies you may use in making judgments about neuropsychological data. Specifically, I will be examining the accuracy of judgments, what tests may be weighted more heavily in judgments, and whether or not a linear or configural decision making process was used. These issues in clinical decision making will be examined utilizing the methodology of the Brunswik Lens Model. Many of these decision making issues in neuropsychology have received minimal, if any, research attention.

The methodology I am using does not require me to sample the judgments of a large number of neuropsychologists. In fact, I am only requesting six neuropsychologists to participate. Therefore, your consent to participate is all the more valuable.

If you choose to participate, you will be asked to make two judgments on each of 50 neuropsychological protocols. The judgments are presence vs absence of brain injury and localization of brain injury (i.e., right, left, or diffuse). I realize that these judgments are somewhat passe, but the methodology I am employing in this study has not been used in neuropsychology. Therefore, I chose to begin to examine the strengths and weaknesses of this methodology with basic judgments. If the methodology is shown to be a useful technique to understand decision making processes, then more complex judgments can be examined.

Judgments will be based on 20 to 30 cues, that is, pieces of data consisting of commonly employed neuropsychological tests and demographic information. (See next page for a listing of the neuropsychological tests used and see the last page for a copy of a protocol). Base rates will be provided about the ratio of presence vs absence of brain injury and the ratio of right, left and diffuse injury protocols. Neuropsychological protocols from people who have a documented brain injury were obtained from reviewing records from a neuropsychology laboratory. Protocols from individuals without a history of brain impairment were obtained from volunteers working in a hospital setting. All protocols are from right-handed adults (18 to 65 years of age). If you participate, you will be provided with more information concerning how the neuropsychological protocols were collected and selected and how the criteria (i.e., right, left and diffuse injury) were defined. There is no form of deception or trickery, of any kind, in this study.

Measures will be taken to maximize confidentiality for those participating. Specifically, all forms will be coded so that your name will not appear. In addition, your name, employment affiliation(s) and state residence will not appear in the dissertation or in any published article/presentation.

Participants will be asked to complete making the two judgments on the 50 protocols and return the materials within 4 weeks. Those completing the task will receive a nominal monetary sign of appreciation of $100.00. Following the completion of the dissertation, participants will be sent a summary of the study as well as data on their own judgment accuracy.

If you are interested in participating, please complete the form on the next page and return it as soon as possible. As soon as I receive your response, the protocols will be sent to you to complete.

Sincerely,	Sincerely,
Donald U. Robertson, PhD Professor of Psychology	Marc D. Gaudette, MA Doctoral Candidate

---

Neuropsychological data may include scores from the following test:

WAIS-R
Category Test
Wisconsin Card Sorting test
Finger Tapping Test
Trail Making Test (Parts A & B)
FAS Test
WMS-R (immediate and delayed recall trials of the Logical memory subtest and the Visual Reproduction subtest)

Protocols will contain the following demographic information: Age, education, gender, and occupation.

Please Return This Form

Name:________________________

Address:________________________

____

Check one:

I am interested in participating________**

I am NOT interested in participating________

**If you are interested in participating, please complete the following:

Subsequent to receipt of your doctoral degree, how many years (full-time years) of experience do you have in neuropsychological assessment?

____

Did you complete, or are you in the process of completing, a formal post doctoral program/fellowship in clinical neuropsychology?

(Circle one) YES NO

Appendix C

General Information

In this binder, there are 50 actual (as opposed to confabulated) neuropsychological protocols from four groups of individuals between the ages of 18 to 65 years.

Ten of the protocols are from "normal" individuals. These individuals were recruited from the Volunteer Services Department in two hospital settings in western Pennsylvania. These individuals do not have a self-reported history of head injury, neurological disease (e.g., epilepsy, strokes), major psychiatric disorders (e.g, organic mental disorders, psychotic disorders), learning disabilities or drug and alcohol abuse. All ten are right-hand dominant and have at least 12 years of education.

Thirty-eight of the 40 neuropsychological protocols from people who sustained a brain injury were obtained from a neuropsychological service in an university hospital in the mid-west. Two of the protocols were obtained from a hospital in western Pennsylvania. The following groupings comprised the 40 neuropsychological protocols from people who sustained a brain injury.

Ten of the protocols are from individuals who sustained a brain injury apparently confined to the right hemisphere. All ten of these individuals are right-handed, do not have a self-reported history of learning disability, drug and alcohol abuse or a major psychiatric disorder. The criterion of right hemisphere injury was based exclusively on reports from brain imaging scans and, in some cases, neurological examinations which revealed some type of brain insult ostensibly localized to the right hemisphere in the absence of significant herniation, raised intracranial pressure or other mass effect. The right hemisphere group was composed of the following etiologies: tumors, gun shot wound, strokes, brain abscess, infarcts, and a contusion. Individuals who sustained right hemisphere injury from a motor vehicle accident were not included in this group, because such injuries usually result in diffuse damage which may go undetected by brain scans. The neuropsychological data was not used as a determinant in the establishment of the criterion of right hemisphere injury.

Ten of the protocols are from individuals who sustained a brain injury apparently confined to the left hemisphere. All ten of these individuals are right-handed, do not have a self-reported history of learning disability, drug and alcohol abuse or a major psychiatric disorder. The criterion of left hemisphere injury was based exclusively on reports from brain imaging scans and, in some cases, neurological examinations which revealed some type of brain insult ostensibly localized to the left hemisphere in the absence of significant herniation, raised intracranial pressure or other mass effect. The left hemisphere group was composed of the following etiologies: tumors, AVMs, strokes and brain abscesses. Individuals who sustained left hemisphere injury from a motor vehicle accident were not included in this group, because such injuries usually result in diffuse damage which may go undetected by brain scans. The neuropsychological data was not used as a determinant in the establishment of the criterion of left hemisphere injury.

Twenty of the protocols are from individuals who sustained diffuse brain injury (i.e., brain impairment involving both the right and left hemispheres). All twenty of these individuals are right-handed, do not have a self-reported history of learning disability, drug and alcohol abuse or a major psychiatric disorder. The criterion of diffuse injury was based on reports from a patient's medical record that indicated that the patient experienced significant neurological sequelae, ostensibly resulting in bilateral brain injuries, following a motor vehicle accident or closed head injury. The diffuse brain injury group was composed of the following etiologies: traumatic head injuries from motor vehicle accidents, motorcycle accidents and falls. The neuropsychological data was not used as a determinant in the establishment of the criterion of diffuse brain injury.

Instructions

1. Please provide a judgment on each and every protocol as to the presence vs absence of brain injury based on the neuropsychological data provided.

2. If a protocol is judged as indicating the presence of brain injury, please make a judgment as to the localization of the brain damage, i.e., right, left or diffuse.

3. For your convenience, normative data (taken from the neuropsychological literature) for test interpretation are provided at the back section of the binder. If you wish, you may use these normative data in the data interpretive process. If you prefer to use other normative data, you may do so.

4. Some protocols contain missing information/data. The fact that some protocols have missing data is not a controlled design feature of this study. The missing data is a consequence of factors that operate in real life neuropsychological testing situations.

5. You may consult books or journal articles in the judgment process.

6. Please do not consult, in any way, other professionals or neuropsychologists regarding the your data analysis, decision making process or final judgments.

7. Please do not copy or duplicate, in any way, the neuropsychological protocols and materials in the binder.

8. I would like to kindly request that you complete this task within 4 weeks and mail back the binder in the addressed and paid envelope provided.

Thank you so much for you voluntary participation! I hope you enjoy this task.

Good Luck!

Appendix D

Norms for the subtests of the WAIS-R.

From - Wechsler, D. (1981). WAIS-R manual. New York: Psychological Corporation-
Harcourt Brace Jovanovich.

(Wechsler, 1981, p. 151)

Scaled score on any single test	Number of SD's from the mean	Percentile rank
19	+3	99.9
18	+2 2/3	99.6
17	+2 1/3	99.0
16	+2	98.0
15	+1 2/3	95.0
14	+ 1 1/3	91.0
13	+ 1	84.0
12	+ 2/3	75.0
11	+ 1/3	63.0
10	0 (Mean)	50.0
9	-1/3	37.0
8	-2/3	25.0
7	-1	16.0
6	-1 1/3	9.0
5	-1 2/3	5.0
4	-2	2.0
3	-2 1/3	1.0
2	-2 2/3	0.4
1	-3	0.1

Norms for the Logical Memory and Visual Reproduction subtests of the Wechsler Memory Scale-Revised.

From - Wechsler, D. (1987). Wechsler Memory Scale-Revised Manual. Psychological Corporation-Harcourt Brace Jovanovich.

Percentile Equivalents of Raw Scores for Logical Memory I, by Age

Age Group							
Raw Score	18-19	20-24	25-34	35-44	45-54	55-64	65-69
43	99	99	99	99	99	99	99
42	98	98	99	99	99	99	99
41	97	98	98	99	99	99	99
40	95	97	97	98	98	99	99
39	95	97	97	97	97	98	99
38	92	96	96	96	97	98	99
37	91	96	95	95	97	98	98
36	89	94	94	94	96	98	98
35	85	90	91	93	94	96	98
34	83	88	90	92	93	94	97
33	80	84	85	86	89	92	94
32	78	81	81	82	86	90	91
31	72	76	77	78	83	88	89
30	66	73	74	75	81	86	87
29	53	60	62	64	74	83	86
28	51	57	58	59	70	80	84
27	45	54	55	56	66	76	80
26	41	50	51	52	59	67	72
25	37	46	47	49	57	65	70
24	28	39	41	43	53	63	65
23	25	34	35	37	47	57	60
22	22	29	31	33	41	51	54
21	19	24	25	26	34	44	52
20	18	22	24	26	32	40	45
19	14	17	21	25	29	34	36
18	12	16	20	24	25	27	30
17	9	14	16	18	21	24	27
16	7	12	13	14	16	18	23
15	4	7	8	10	10	11	19
14	3	6	7	9	9	10	16
13	2	4	4	5	6	7	11
12	2	3	3	4	5	6	9
11	2	3	3	4	4	5	8
10		1	2	2	3	3	4

Percentile Equivalents of Raw Scores for Logical Memory II, by Age

Age Group							
Raw Score	18-19	20-24	25-34	35-44	45-54	55-64	65-69
43	99	99	99	99	99	99	99
42	99	99	99	99	99	99	99
41	98	99	99	99	99	99	99
40	97	98	98	99	99	99	99
39	96	97	97	98	98	99	99
38	96	97	97	98	98	99	99
37	95	97	97	97	98	99	99
36	93	95	95	96	97	98	99
35	92	94	94	95	96	98	98
34	88	92	93	94	96	98	98
33	86	90	91	92	95	98	98
32	84	88	89	90	94	97	98
31	82	85	86	88	93	97	97
30	79	83	84	86	92	96	97
29	76	79	80	81	89	95	96
28	70	75	75	75	85	92	94
27	63	72	73	74	81	87	90
26	57	66	67	69	78	85	87
25	51	61	61	62	72	82	83
24	47	57	58	59	69	79	81
23	41	53	54	56	66	76	78
22	37	49	50	51	62	73	75
21	33	44	45	46	57	68	70
20	28	40	41	42	53	64	65
19	25	35	36	37	45	53	60
18	23	30	31	33	42	51	58
17	19	24	27	30	38	48	52
16	16	22	24	27	31	37	47
15	15	20	22	25	29	33	45
14	13	18	21	24	26	29	40
13	12	16	19	22	23	24	36
12	8	10	14	19	19	20	32
11	7	9	13	17	17	18	30
10	6	8	10	13	14	15	21
9	5	6	8	11	12	13	16
8	4	5	7	9	10	12	15
7	3	4	5	7	8	9	14
6	1	2	3	5	6	8	12

Percentile Equivalents of Raw Scores for Visual Reproduction I, by Age

Age Group							
Raw Score	18-19	20-24	25-34	35-44	45-54	55-64	65-69
41	99	99	99	99	99	99	99
40	96	96	97	98	98	99	99
39	93	94	96	98	98	99	99
38	85	86	90	94	96	98	98
37	75	76	83	90	94	98	98
36	67	70	74	79	88	95	96
35	61	62	65	68	79	88	94
34	48	54	56	59	72	83	89
33	40	42	45	48	63	77	83
32	35	37	38	40	54	68	76
31	26	29	31	33	50	66	72
30	18	22	24	27	41	57	67
29	12	18	19	20	32	48	63
28	11	16	17	18	29	42	60
27	9	12	14	16	24	35	52
26	7	10	12	15	23	33	45
25	6	8	11	14	19	25	40
24	5	6	9	13	17	22	36
23	3	4	7	11	15	19	34
22	2	3	6	9	11	13	30
21	2	3	4	6	7	9	20
20	1	2	3	4	6	8	14
19	1	2	2	3	5	7	12
18	1	2	2	3	4	5	10
17	1	2	2	3	4	5	8

Percentile Equivalents of Raw Scores for Visual Reproduction II, by Age

Age Group							
Raw Score	18-19	20-24	25-34	35-44	45-54	55-64	65-69
41	99	99	99	99	99	99	99
40	99	99	99	99	99	99	99
39	95	95	95	96	98	99	99
38	92	93	93	94	97	99	99
37	87	88	89	90	95	98	98
36	83	86	87	88	94	98	98
35	79	80	80	81	91	97	98
34	70	72	74	77	87	94	95
33	62	64	69	74	80	85	94
32	55	58	61	64	73	82	93
31	46	50	51	53	65	77	92
30	39	41	44	48	61	74	85
29	33	34	37	40	53	66	83
28	26	27	30	33	45	59	81
27	21	22	25	29	38	48	78
26	16	18	21	24	33	46	72
25	15	16	18	20	30	42	67
24	13	14	16	18	27	37	61
23	10	12	14	17	24	33	54
22	8	10	13	16	23	31	49
21	6	8	11	14	20	29	45
20	5	8	10	13	19	27	43
19	4	8	10	12	16	22	38
18	4	7	9	11	15	20	34
17	3	6	7	9	12	17	27
16	3	5	6	7	10	14	25
15	3	5	5	6	9	13	23
14	2	4	4	5	8	11	21
13	2	3	3	4	7	10	18
12	2	3	3	3	6	9	14
11	1	2	2	3	4	7	12
10	1	1	1	2	3	4	11

Norms for the Trail Making Test.

From - Lezak, M. D. (1983). Neuropsychological assessment. New York: Oxford University Press.

Adapted from - Davies, A. D. M. (1968). The influence of age on the trail making test performance. Journal of Clinical Psychology, 24, 96-98.

age	20-39 (n=180)		40-49 (n=90)		50-59 (n=90)		60-69 (n=90)		70-79 (n=90)	
Part	A	B	A	B	A	B	A	B	A	B
Percentile										
90	21	45	22	49	25	55	29	64	38	79
75	26	55	28	57	29	75	35	89	54	132
50	32	69	34	78	38	98	48	119	80	196
25	42	94	45	100	49	135	67	172	105	292
10	50	129	59	151	67	177	104	282	168	450

Note: Time in seconds.

OR

From- Auch-Fromm, D., & Yeudall, L. T. (1983). Normative data for the Halstead- Reitan neuropsychological tests. Journal of Clinical Neuropsychology, 5, 221-238.

Normative Data for the Trail Making Test in Seconds Stratified by Age

	PART A				PART B		
Age	n	M	SD	Range	M	SD	Range
15-17	32	23.4	5.9	15.2-39.0	47.7	10.4	25.4-81.0
18-23	76	26.7	9.4	12.0-60.1	51.3	14.6	23.3-101
24-32	57	24.3	7.6	11.8-46.0	53.2	15.6	29.1-98.0
33-40	18	27.5	8.3	16.0-52.7	62.1	17.5	39.0-111
41-65	10	29.7	8.4	16.5-42.0	73.6	19.4	41.9-102

OR

From - Russell, E. W., Neuringer, C., & Goldstein, G. (1970). Assessment of brain damage: A neuropsychological key approach. New York: Wiley.

Revised Norms for Rating Equivalents of Raw Scores

Rating Equivalents of Raw Scores

	0	1	2	3	4	5
Trails A	<19	20-33	34-48	49-62	63-86	87+
Trails B	<57	58-87	88-123	124-186	187-275	276+

Note: Time in seconds

Note: 0 and 1 = normal range; 2= mild impairment; 3= moderate impairment; and, 4 and 5 = severe impairment.

Norms for the Finger Tapping Test, Halstead -Reitan.

From - Auch-Fromm, D., & Yeudall, L. 1. (1983). Normative data for the Halstead- Reitan neuropsychological tests. *Journal of Clinical Neuropsychology*, 5, 221-238.

Males

		Preferred hand			Nonpreferred hand		
Age	n	M	SD	Range	M	SD	Range
15-17	17	47.6	5.8	38.0-55.6	43.6	4.9	33.4-51.8
18-23	44	49.6	6.9	26.6-64.6	45.4	6.9	26.8-58.6
24-32	31	50.6	6.6	38.2-66.2	46.0	6.1	28.8-55.0
33-40	12	53.4	5.9	39.0-61.0	49.8	4.7	41.0-57.8
41-64	4	44.4	5.8	35.8-48.2	41.4	3.5	36.6-44.4

Females

Age	n	M	SD	Range	M	SD	Range
15-17	15	42.7	7.9	30.2-54.0	41.1	6.2	31.6-51.0
18-23	30	43.6	7.5	30.6-65.6	41.2	6.5	32.8-61.8
24-32	25	45.2	6.7	31.0-60.0	40.9	5.7	28.6-53.6
33-40	6	45.8	5.5	40.6-55.6	44.3	4.6	40.6-53.2
41-64	6	40.4	4.8	34.2-48.4	38.6	4.8	32.0-46.6

Note: Average number of taps over five trials.

OR

From - Reitan, R.M. Manual for administration of neuropsychological test batteries for adults and children.

Finger Tapping Test	Mean	S.D.
	50.74	7.29

Note. Mean for control subjects.
N=50.

OR

From - Russell, E. W., Neuringer, C., & Goldstein, G. (1970). Assessment of brain damage: A neuropsychological key approach. New York: Wiley.

Tapping (No.)

		Rating Equivalents of Raw Scores					
		0	1	2	3	4	5
Dom.	M	55	54-50	49-43	42-32	31-20	19-0
	F	51	50-46	45-39	38-28	27-16	15-0
Nondom.	M	49	48-44	43-37	36-26	25-14	13-0
	F	45	44-40	39-33	32-22	21-10	9-0

Note: 0 and 1 = normal range; 2=mild impairment; 3=moderate impairment; and, 4 and 5 = severe impairment.

Norms for the Category Test.

From - Russell, E. W., Neuringer, C., & Goldstein, G. (1970). Assessment of brain damage: A neuropsychological key approach. New York: Wiley.

	Ratings Equivalents of Raw Scores					
	0	1	2	3	4	5
Category Test Errors	25	26-52	53-75	76-105	106-131	132+

Note: 0 and 1 = normal range; 2=mild impairment; 3=moderate impairment; and, 4 and 5 = severe impairment.

OR

From - Reitan, R. M. Manual for administration of neuropsychological test batteries for adults and children.

Category Test	Mean	S.D.
	32.38	12.62

Note. Mean number of errors for control subjects.
N=50.

OR

From - Auch-Fromm, D., & Yeudall, L. T. (1983). Normative data for the Halstead- Reitan neuropsychological tests. Journal of Clinical Neuropsychology, 5, 221-238.

Age	n	M	SD	Range
15-17	32	35.8	16.2	16-68
18-23	71	35.9	21.2	9-106
24-32	55	30.5	13.6	10-68
33-40	18	36.3	14.3	11-67
41-64	10	53.0	21.0	29-96

Note: Number of errors.

Norms for the Controlled Word Association Test (i.e., FAS Test).

From - Lezak, M. D. (1983). Neuropsychological assessment. New York: Oxford University
Press.
Adaptive from - Benton, A. L., & Hamsher, K. deS. (1976). Multilingual aphasia examination.
. Iowa City: University of Iowa.

Controlled Oral Word Association Test: Adjusted Formula for Males and Females

Add points to raw scores of 10 and above as indicated:

	Age					
Education (years completed)	25-54		55-59		60-64	
	M	F	M	F	M	F
less than 9	9	8	11	10	14	12
9-11	6	5	7	7	9	9
12-15	4	3	5	4	7	6
16 or more	-	-	1	1	3	3

Adjusted score	Percentile Range	Classification
53+	96+	Superior
45-52	77-89	High Normal
31-44	25-75	Normal
25-30	11-22	Low Normal
23-24	5-8	Borderline
17-22	1-3	Defective
10-16	1	Severe defect
0-9	1	Nil -Trace

Norms for the Wisconsin Card Sorting Test.

From - Heaton, R. K. (1981). A manual for the Wisconsin Card Sorting Test. Odessa, Fl: Psychological Assessment Resources, Inc.

	Age (years)			
	40 (n=100)	40-49 (n=19)	50-59 (n=16)	59 (n=15)
Full Scale IQ	113.9(11.7)	112.4(13.4)	120.3(9.4)	109.7(9.9)
Categories Achieved	5.6 (1.0)	4.8(1.8)	5.6(1.1)	4.2(2.0)

Note: Means and standard deviations.

AND

	Education (years)		
	12 (n=20)	12-15 (n=77)	15 (n=53)
Full Scale IQ	105.2(9.8)	110.8(10.9)	121.8(8.8)
Categories Achieved	5.1(1.4)	5.2(1.5)	5.7(1.0)

Note: Means and standard deviations. .

Appendix E

May, 1991

Dear Dr.

I recently received your responses to the 50 neuropsychological protocols. Thank you for returning the material in a timely fashion.

Attached are a few follow-up questions that I kindly request you complete. I expect that it will take no longer than 15 minutes to respond. Please complete the questions and return them as soon as possible (a stamped envelope is provided).

I will be contacting you soon, if I haven't already, to complete a form to satisfy administrative requirements so that the $100.00 honorarium can be mailed out to you.

Thanks again for your cooperation, time and participation.

Sincerely,

Marc Gaudette, MA
Doctoral Candidate

Follow-up Questions

Name:________________________ Date:______________

1. Please estimate how much time you spent on the judgment and decision making task.

Approximately_______hours.

2. Using the scale below, please provide a Mean rating and a Range rating of how confident you were making the presence vs absence judgment?

Not at all confident Very confident

1 2 3 4 5 6 7

Mean rating of confidence making the presence vs absence judgment_____
My level of confidence, making the presence vs absence judgment, ranged from_____ to _____

3. Using the scale below, please provide a Mean rating and a Range rating on how confident you were making the localization judgments, i.e., right, left and diffuse?

Not at all confident Very confident

1 2 3 4 5 6 7

Mean rating of confidence making the localization judgment of right hemisphere injury_____
My level of confidence, making the localization judgment of right hemisphere injury, ranged from____ to _____

Mean rating of confidence making the localization judgment of left hemisphere injury_____
My level of confidence, making the localization judgment of left hemisphere injury, ranged from____ to _____

Mean rating of confidence making the localization judgment of diffuse injury_____
My level of confidence, making the localization judgment of diffuse injury, ranged from____ to _____

4. Please estimate the percentage of protocols you correctly judged in making the presence vs absence judgment.

____

5. Please estimate the percentage of protocols you correctly judged in making the localization judgment, i.e., right, left and diffuse.

I correctly judged______________% of the 10 right hemisphere injury protocols.

I correctly judged______________% of the 10 left hemisphere injury protocols.

I correctly judged______________% of the 20 diffuse injury protocols.

6. Using the scale provided, please rate each test's importance in your decision making process for the judgment of presence vs absence of brain injury.

Not at all important Very important

1 2 3 4 5 6 7

Information subtest_____

Digit Span subtest_____

Vocabulary subtest_____

Arithmetic subtest_____

Comprehension subtest_____

Similarities subtest_____

Trails A_____

Trails B_____

FAS Test_____

Category Test_____

Finger Tapping Test_____

Wisconsin Card Sorting Test_____

Logical Memory subtest, immediate recall trial_____

Logical Memory subtest, delayed recall trial_____

Visual Reproduction subtest, immediate recall trial_____

Visual Reproduction subtest, delayed recall trial_____

Picture Completion subtest_____

Picture Arrangementsubtest_____

Block Design subtest_____

Object Assembly subtest_____

Digit Symbol subtest_____

Verbal IQ_____

Performance IQ_____

Full Scale IQ_____

Age_____

Education_____

Gender_____

Occupation_____

7. Using the scale provided, please rate each test's importance in your decision making process for the judgment of localization of brain injury.

Not at all important Very important

1 2 3 4 5 6 7

Information subtest____

Digit Span subtest____

Vocabulary subtest____

Arithmetic subtest____

Comprehension subtest____

Similarities subtest____

Picture Completion subtest____

Picture Arrangementsubtest____

Block Design subtest____

Object Assembly subtest____

Digit Symbol subtest____

Trails A____

Trails B____

FAS Test____

Category Test____

Finger Tapping Test____

Wisconsin Card Sorting Test____

Logical Memory subtest, immediate recall trial____

Logical Memory subtest, delayed recall trial____

Visual Reproduction subtest, immediate recall trial____

Visual Reproduction subtest, delayed recall trial____

Verbal IQ____

Performance____

Full Scale ____

Age____

Education____

Gender____

Occupation____

8. If you would like to make any comments about the judgment task and materials provided, please do so below.

Appendix F

Judges' Subjective Weighting of the 27 Cues for the Presence vs Absence Judgment

	Experts			Novices		
Cues	1	2	3	1	2	3
Age	5[a]	6	6	4	3	6
Education	6	6	6	6	6	7
Gender	2	1	2	2	2	1
Occupation	6	5	4	6	4	5
Verbal IQ	5	4	5	6	6	4
Performance IQ	5	6	5	6	6	4
Full Scale IQ	5	5	2	6	6	4
Information	4	1	3	4	6	6
Digit Span	5	2	4	4	4	6
Vocabulary	4	2	4	4	5	6
Arithmetic	4	3	4	4	4	6
Comprehension	4	2	5	4	5	6
SIMILARITIES[b]	5	2	5	4	5	6
Picture Completion	4	3	3	4	4	6
Picture Arrangement	4	5	4	4	6	6
BLOCK DESIGN	5	6	4	4	6	6
Object Assembly	5	5	3	4	4	5
DIGIT SYMBOL	5	4	5	5	6	6

Trail A	6	3	3	2	4	7
TRAIL B	6	6	3	2	6	7
Wisconsin Card Sort	6	5	4	6	4	6
Category Test	6	6	3	4	6	5
FINGER TAPPING RIGHT	6	4	3	3	4	4
FINGER TAPPING LEFT	6	4	3	3	4	4
irverb,WMS-R	6	4	4	4	4	5
DRVERB, WMS-R	6	5	3	1	4	6
IRVIS, WMS-R	5	4	4	4	4	5
DRVIS, WMS-R	5	5	3	1	4	6
FAS	6	5	4	4	6	5

[a]Scale=1 (test score was not at all important to the judgment) to 7 (test score was very important to the judgment).

[b]Test scores in CAPITAL letters refer to the nine predictor cues.

Judges' Subjective Weighting of the 27 Cues for the Localization Judgment

	Experts			Novices		
Cues	1	2	3	1	2	3
Age	6	2	6	2	1	3
Education	6	3	6	5	1	3
Gender	2	1	2	2	1	1
Occupation	6	1	4	5	1	3
Verbal IQ	5	6	6	6	6	6
Performance IQ	5	6	6	6	6	6
Full Scale IQ	5	2	2	6	3	4
Information	4	6	3	4	3	6
Digit Span	5	2	5	4	3	6
Vocabulary	4	6	4	4	3	6
Arithmetic	4	6	5	4	3	6
Comprehension	4	5	6	4	3	6
SIMILARITIES[a]	5	6	6	5	3	6
Picture Completion	4	4	3	4	3	6
Picture Arrangement	4	4	5	5	6	6
BLOCK DESIGN	5	6	5	5	6	6
Object Assembly	5	4	5	4	3	6
DIGIT SYMBOL	5	2	5	4	3	6

Trail A	6	3	3	2	1	1
TRAIL B	6	3	3	2	1	1
Wisconsin Card Sort	6	3	5	5	1	2
Category Test	6	3	4	6	1	2
FINGER TAPPING RIGHT	6	6	4	5	6	7
FINGER TAPPING LEFT	6	6	4	5	6	7
irverb,WMS-R	6	6	5	4	3	7
DRVERB, WMS-R	6	6	4	1	3	7
irvis, WMS-R	5	6	5	4	3	7
DRVIS, WMS-R	5	6	4	1	3	7
FAS	6	6	5	6	6	6

[a]Scale=1 (test score was not at all important to the judgment) to 7 (test score was very important to the judgment).

[b]Test scores in CAPITAL letters refer to the nine predictor cues.

Appendix G

Means and Standard Deviations for the Test Scores for the Four Sets of Protocols

	Protocols[a]				
	Normals	RHBD	LHBD	DBD	
Variables[b]	X (SD)	X (SD)	X (SD)	X (SD)	
Age	44.2 (16.0)	39.0 (15.0)	43.6 (14.7)	32.4 (10.7)	NS[c]
Education	13.4 (1.8)	13.6 (2.7)	12.5 (3.7)	14.2 (3.4)	NS
VIQ	97.8 (7.9)	98.0 (13.0)	86.7 (15.7)	97.8 (14.1)	NS
PIQ	102.6 (12.4)	84.7 (11.2)	89.1 (15.8)	94.6 (13.4)	S[c], d
FSIQ	99.2 (9.5)	92.0 (12.7)	87.3 (15.8)	95.8 (13.1)	NS
Information	9.9 (2.0)	10.3 (2.4)	7.3 (3.7)	9.4 (3.3)	NS
Digit Span	7.9 (2.0)	10.2 (1.5)	7.3 (3.8)	9.6 (2.7)	NS
Vocabulary	10.1 (2.1)	9.8 (3.3)	7.5 (2.5)	9.5 (2.6)	NS
Arithmetic	9.1 (2.9)	9.4 (2.8)	7.6 (2.9)	10.7 (3.2)	S. e
Comprehension	9.7 (1.5)	9.5 (3.4)	7.1 (3.2)	9.4 (2.6)	NS
Similarities	10.7 (1.9)	9.2 (2.2)	8.8 (2.6)	9.3 (3.1)	NS
Picture Comp.	8.8 (2.4)	7.6 (1.6)	7.9 (1.5)	9.1 (2.6)	NS
Picture Arr.	10.0 (3.2)	8.1 (2.3)	8.7 (2.4)	8.4 (2.9)	NS
Block Design	10.5 (2.8)	6.9 (2.7)	9.4 (3.6)	10.2 (2.8)	S. d,f
Object Assembly	9.7 (2.4)	7.9 (3.6)	8.9 (5.0)	9.4 (3.1)	NS
Digit Symbol	13.2 (2.2)	7.8 (2.8)	7.2 (2.4)	8.8 (2.8)	S. d,g,h
Trail A	27.2 (9.0)	40.5 (26.0)	58.4 (59.0)	29.8 (10.0)	NS
Trail B	64.1 (24.5)	126.5 (80.6)	138.4 (94.6)	82.4 (38.2)	S. g
WCS	4.5 (1.5)	4.1 (2.3)	4.6 (1.3)	4.9 (1.4)	NS

	Normals X (SD)	RHBD[a] X (SD)	LHBD[b] X (SD)	DBD[c] X (SD)	
Category	59.7 (22.7)	60.3 (33.8)	64.0 (39.9)	52.7 (26.4)	NS
Tapping Right	50.2 (4.0)	45.1 (5.2)	40.4 (8.5)	45.0 (6.3)	S. g
Tapping Left	42.2 (4.5)	39.9 (5.6)	42.0 (3.4)	41.2 (6.7)	NS
IRVERB[d]	27.7 (7.0)	22.1 (7.0)	13.6 (9.9)	22.2 (7.2)	S. e,g
DRVERB[e]	24.1 (8.0)	19.4 (6.6)	9.1 (9.9)	15.8 (7.4)	S. g,h,i
IRVIS[f]	35.4 (7.6)	26.1 (6.7)	28.6 (9.1)	30.9 (6.3)	S. d
DRVIS[g]	30.2 (9.1)	18.8 (8.0)	22.3 (12.5)	25.4 (7.6)	S. d
FAS	34.1 (8.1)	37.0 (9.6)	24.4 (16.6)	31.5 (10.9)	NS

[a] Protocols: RHBD=Right hemisphere brain damage. LHBD= Left hemisphere brain damage. DBD=Diffuse brain damage.

[b] Variables: IRVERB=Immediate recall trail, Logical Memory subtest, WMS-R. DRVERB= Delayed recall trial, Logical Memory subtest, WMS-R. IRVIS= Immediate recall trial, Visual Reproduction subtest, WMS-R. DRVIS=Delayed recall trial, Visual Reproduction subtest, WMS-R.

[c] NS=One-way analysis of variance was not significant (i.e., $p > 0.05$). S= One-way analysis of variance was significant (i.e., $p < 0.05$).

d=Normals vs RHBD

e=DBD vs LHBD

f=DBD vs RHBD

g=Normals vs LHBD

h=Normals vs DBD

i=RHBD vs LHBD

Appendix H

Neuropsychological Data.

Appendix G presents the means and standard deviations of 27 of the 29 possible cues (occupation and gender were not quantified). Examining the 25 test scores for the normal group it is clear that this so called normal group generally achieved test scores in the average range of cognitive functioning on the majority of the tests. Exceptions included a relatively low scaled score on Digit Span (7.9) and Picture Completion (8.8), and a relatively high Category test score of 59.7 (indicative of impaired performance). Otherwise, the normal group's test scores were solidly in the average range. In addition, the normal group's mean scores outperformed (not always statistically) the other three groups on 17 of the 25 tests. Therefore, there appears to be sufficient evidence to suggest that the so called normal group was essentially composed of people who obtained scores in the average range based on normative data in the published literature.

In terms of the right hemisphere group, conventional neuropsychological principles and empirical data concerning hemispheric specialization (see Lezak, 1983; Kolb and Whishaw, 1990) suggest that this group should have greater relative difficulty on tests purportedly mediated by the right hemisphere (e.g., perceptual and spatial functioning, left hand motor functioning and nonverbal memory functioning, and PIQ lower than VIQ). The data shows that VIQ was greater than PIQ by a notable 13 points, and this group had the lowest PIQ; the two most perceptual-spatial tasks on the WAIS-R, i.e., Block Design and Object Assembly were lower for this group than any other group; although not statistically significant, the left hand finger tapping score was lower in this group than in the other groups; and the immediate and recall trails of the nonverbal task of the WMS-R were lower for this group than the other groups. Therefore, the right hemisphere protocols used in this study as a group appeared to conform to conventional neuropsychological principles and previous published data concerning hemispheric specialization (see Lezak, 1983).

As for the left hemisphere group, conventional neuropsychological principles and empirical data on hemispheric specialization suggest that this group should have the greatest difficulty on tasks associated with the functioning of the left hemisphere (e.g., verbal tasks, right hand finger tapping, verbal memory, and VIQ lower than PIQ). The data for the left hemisphere group showed that they had a slightly higher (although not statistically) PIQ relative to VIQ, and this group had the lowest VIQ compared to the other three groups. Also, heavily mediated verbal tasks from the WAIS-R (i.e., Information, Vocabulary, Comprehension and Similarities) were lower (although not statistically) in this group than in any other group. The immediate and delayed trails of the Logical memory subtest (i.e., verbal memory) of the WMS-R were lower in this group compared to the other groups. Finally, verbal fluency measured by the FAS test was lower (although not statistically) in this group than in any other group. Therefore, test scores from the left hemisphere brain damaged protocols appeared to be generally characteristic of cognitive dysfunction associated with left hemisphere brain injury based on conventional neuropsychological principles and previous data (Lezak, 1983; Kolb & Whishaw, 1990).

The diffuse brain damaged group typically have sustained some form of traumatic head injury (e.g., motor vehicle accident) that has affected cognitive functioning bilaterally (e.g., involving both the right and left hemispheres). Typically, individuals who sustain a diffuse brain injury will show a greater number, but not necessarily more severe cognitive dysfunction, on neuropsychological tests compared to a brain injury lateralized to just one hemisphere. The test scores from the diffuse brain damaged group appeared to be more similar to the normal group than to the right or left hemisphere groups. That is, the diffuse group did not seem as impaired as the other two brain damaged groups. Specifically, the normal group outperformed the right and left hemisphere brain damaged groups in 19 of the 25 comparisons, while the diffuse brain damaged group outperformed the right and left hemisphere brain damaged groups in 15 of the 25 comparisons. Intuitively, this finding indicates that the protocols used to represent the diffuse group were not as severely brain damaged as the protocols used to represent the other two brain damage groups. Nonetheless, the normal group outperformed the diffuse brain damaged group on

17 of the 25 comparisons.

Appendix I

Correlations Between Predictor Cues and the Ecological Criteria

Predictor Cues and the Presence/Absence Criterion: The data and analyses in this study found that Digit Symbol, right hand finger tapping, delayed trial of the Logical Memory subtest of the WMS-R (DRVERB) and the delayed trial of the Visual Reproduction subtest of the WMS-R (DRVIS) significantly correlated with the presence/absence ecological criterion. Therefore, four of the nine predictor cues significantly correlated with the criterion.

Digit Symbol's strong association with the presence/absence judgment (r = 0.62) is consistent with Lezak's (1983) review. That is, she reported that Digit Symbol is the most sensitive of the WAIS's subtest to cortical dysfunction. The normals scored about 1 SD above the mean while the brain damaged groups scored about 2/3 SDs below the mean.

Finger Tapping right hand was found to be significantly associated with the presence/absence judgment (r = 0.38). Apparently, brain insult lowered right hand finger tapping below that which the normals scored, therefore allowing for a significant relationship to emerge. In contrast, the level of left hand finger tapping was comparable for the four groups of protocols. For right hand finger tapping, the normals scored solidly in the average range, while the brain damaged groups scored about 1 SD below the mean (see normative data in Appendix D). For left hand finger tapping, all groups scored about 1 SD below the mean (see normative data in Appendix D).

The delayed trials of the Logical Memory subtest and the Visual Reproduction subtest of the WMS-R were significantly associated with the presence/absence judgment (r = 0.40). This is consistent with Squire's (1987) position that the delayed recall trials of memory tests tend to be more sensitive to brain damage than immediate recall trials.

It was surprising that Trail B did not produce a significant correlation for the presence/absence criterion (r = 0.27). Trail B is often proclaimed to be an especially sensitive measure of brain dysfunction. The data presented in Appendix G clearly showed that the normals generally scored in the average range according to the normative data provided in Appendix D, while the brain damaged groups generally scored in the brain impaired range.

PIQ's significant correlation with the presence/absence criterion was not surprising. Mandleberg and Brooks (1975) found that Performance IQ was lower and took much longer to recover than Verbal IQ following severe head injury. Botwinick (1984) reported that Performance IQ declined with increasing age more notably than Verbal IQ. Overall, these studies indicate that the subtests comprising the Performance IQ tend to be more sensitive and less robust to brain dysfunction than the Verbal subtests. The immediate recall trials of the Logical Memory and the Visual Reproduction subtests also significantly correlated with the criterion. These subtests' sensitivity to the presence of brain injury is consistent with the literature on memory and various types of neurological insults (see Squire, 1986). That is, memory problems tend to be the most frequent complaints in people with a brain injury, and memory dysfunction tends to be associated with most neurologic disorders.

Predictor Cues and the Localization Criterion: The delayed recall trial of the Logical memory subtest of the WMS-R and the FAS test were the only two of the nine predictor cues that significantly correlated with the localization criterion.

The significant relationship between the delayed recall trial of the Logical Memory subtest and the localization criterion ($r = 0.43$) was consistent with Delaney et al.'s (1980) finding that this subtest is especially sensitive to left hemisphere impairment. In this study, the right hemisphere brain damaged group scored significantly higher on this subtest compared to the left hemisphere brain damaged group.

The FAS test's reported sensitivity to left hemisphere brain insult was supported in the data from this study ($r = 0.35$). There was about a 13 point difference in the mean values between the right and left hemisphere groups (the diffuse group scored about in the middle).

The correlation of Block Design with the localization criterion ($r = 0.28$) nearly reached significance (recall that a value of about $r = 0.28$ or greater produced a significant correlation). A somewhat higher correlation was expected though, given that Block Design is considered especially sensitive to right hemisphere insult.

The Information and Digit Span subtests, and the immediate recall trial of the Logical Memory subtest also significantly correlated with the localization criterion. The Information subtest has not been found to be especially sensitive to brain injury (unless the person is aphasic), while the Digit Span subtest and immediate recall trial of the Logical Memory subtest are moderately sensitive measures.

An important variable to consider in the interpretation of the correlation matrix of the predictor cues and the two criteria is the neuropsychological data on which the correlations were based. As was stated in the section above, because of the nature of the design elements in this study, neuropsychological data associated with "severe" brain injuries were probably not consistent with the protocols used. Therefore, in theory, tests scores associated with progressivley more severe brain insults were not indicative of the neuropsycholgical data in this study. Thus, because the brain damaged groups were not representative of a full range of severity (i.e., mild, moderate, severe), the neuropsychological test scores were restriced and the resulting correlations were probably attenuated (Nunnally, 1978).