Prediction in Forensic and Neuropsychology

Sound Statistical Practices

Edited by

Ronald D. Franklin
St. Mary's Hospital
and Florida Atlantic University

LEA
2003
LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
Mahwah, New Jersey London

Neuropsychological Evaluations as Statistical Evidence[1]

Ronald D. Franklin
St. Mary's Hospital and Florida Atlantic University[2]

Joachim I. Krueger
Brown University

EVIDENCE DEFINED

According to Merriam Webster (1996) evidence has different vernacular and legal meanings. In the vernacular it is associated with proof and truth, as well as the observation of events. In law, the term is more precise, referring to "proof of fact(s)" presented at a trial. Evidence is essential in convincing the judge or jury of the facts in a case, thereby enabling the discovery of truth. Legal evidence can include "hard" findings such as photographs, audio recordings, plaster castings, fingerprints, and medical records. More often, however, evidence is provided by a fallible witness who can be questioned and cross-

[1]This chapter is prepared with individuals who have received at least one undergraduate and one graduate course in statistics and psychometric theory in mind. Readers lacking this preparation, or those whose exposure to the topics is weak or dated, should review current works such as those prepared by Glenberg (1996) and Thorndike (1997).

[2]Please address correspondence to PO Box 246, Candor, NC 27229 or rdfphd@yahoo.com

examined. In this chapter we consider the evidentiary basis of psychological test data. Psychometric theory, the foundation of psychological test interpretation, evolved from statistical hypothesis testing (see chapter 3, this volume). Recent trends in test interpretation (see Chan, R. C. K, 2001; Martens, Donders, & Millis, 2001; Mitrushina, Boone & D'Elia, 1999; Putzke, Williams, Blutting, Konold, & Boll, 2001; and Rosenfeld, Sands & Van Gorp, 2000) champion interpretation anchored in base rate. Although base rate must be considered as an import component of diagnostic formulation, rarity is not synonymous with disability. For example, it is rare when children are born with extra digits or red eyes, but both are expressions of normal variation and rarely impair growth or development even though they may be associated with other disorders that affect development. In considering psychological findings as evidence it is wise to remember Fisher's (1959) *frequency admonition* "...the infrequency with which, in particular circumstances, decisive evidence is obtained, should not be confused with the force, or cogency, of such evidence" (p. 93).

THE WITNESS AS EVIDENCE

Courts generally recognize two types of witness, the fact witness and the expert witness.

The Fact Witness

A witness of fact reports firsthand observations to the court. There is no expectation that the fact witness provide an opinion or personal view. Circumstances may permit statements of views from a fact witness when "(a) [opinions are] rationally based on the perception of the witness, and (b) helpful to a clear understanding of the witness' testimony or the determination of a fact in issue" (Stromberg et al., 1988, p. 646). A fact witness may be compelled to testify with no guarantee of payment. The uncertainty of payment may influence the thoroughness of a fact witness' literature review or examination. In medicine and psychology, a treating clinician is often called upon as a fact witness. When this occurs, the witness can be asked to comment on a variety of topics such as the degree of impairment, or the likely cause of a medical or psychological condition, why specific treatments were recommended (or not recommended), and what prognosis can be made. A psychologist who has evaluated a patient following traumatic brain injury may be asked to witness this fact. Usually fact testimony is limited to when and why a patient is seen. However, "facts" sometimes represent opinions

(ie., clinical judgments such as is the patient brain damaged) instead of factual information. What is more, the fact witness may not be allowed to present research findings supporting clinical judgments on the grounds that they are tentative working hypotheses rather than actual facts. Hence there may be rules circumscribing the kinds of information allowable as a fact. What is more, items presented as facts are subject to challenge in cross-examination. When findings provide greater support for one side than the other, an expert may be hired for the purpose of challenging opinions expressed by the fact witness. Challenges of facts typically occur in two forms; cross-examination and testimony by an expert witness. The attorney who cross-examines a fact witness may have a non-testifying expert review reports or notes of the fact witness as well as any depositions taken for the current or prior trials. Reviews of prior trials can include any similar cases for which the fact expert has made public statements, including depositions or trial testimony. When large financial settlements are possible, a consulting expert will likely provide the attorney who cross-examines with information designed to undermine those opinions expressed by the fact witness. There may be no requirement that either attorney inform the fact witness regarding the involvement of either a consulting expert or a testifying expert, although the attorney who calls the fact witness typically provides the fact witness with reports and depositions provided by a testifying expert. Most likely, if an attorney employs a non-testifying expert, neither the fact witness nor the other attorney(s) will have knowledge of the employment. So, a prudent fact witness assumes that every test and test protocol will be scrutinized by a hostile expert who has conducted a thorough literature review and has likely conducted independent research on some aspect of neuropsychology that is relevant to the case.

The Expert Witness

The designation of a witness as "expert" by the courts has specific meaning as defined by Federal Rules of Evidence 702 and 703 (http://expertpages.com/federal/federal.htm). These rules allow courts to qualify a witness as expert on the basis of knowledge, skill, experience, training, or education, and to allow admission of scientific data or other information used in the expert testimony. Experts usually enter a case voluntarily and may or may not actually testify. On some occasions, psychologists or other professionals are appointed by the courts to serve as expert witness. Compensation is usually provided for case preparation and testimony. Because the designation as "expert" does not guarantee

payment in all jurisdictions, many experts require payment of a retainer before they will meet with either the attorney or the patient. Obtaining a retainer is important because it insures that the neuropsychologist can perform both a thorough literature review and an appropriate examination.

Psychologists as Expert Witnesses. The entry of psychology as experts in the courts dates back to 1962 (Ofloff, Beavers, & DeLeon, 1999). Courts have consistently upheld the right of psychologists to qualify as experts for testimony concerning the presence of brain damage. Courts have historically been less supportive in qualifying psychologists as experts regarding the issue of causality (McCaffrey, Williams, Fisher, & Laing, 1997) but acceptance of psychological testimony for this purpose is growing.

Neuropsychologists as Expert Witnesses. Since its introduction in the legal arena, the use of neuropsychological testimony has been vigorously challenged. In their first edition of *Coping With Psychiatric and Psychological Testimony*, Ziskin and Faust (1998) argued that psychological data were based upon inadequate science. Faust, Ziskin, and Hiers (1991) further described neuropsychological data as inadequate legal evidence. Replies to this claim have been vigorous (see Barth, Ryan, & Hawk, 1992; McCaffrey et al., 1997), many writers noted that Ziskin and Faust's critique targeted the scientific method and its application to psychometric theory. Much of the debate focuses either directly or indirectly on the value of null hypothesis testing (NHST) to diagnosis and outcome prediction (see chapter 3, this volume).

Currently, the role of psychology is being resolved in the courts regarding testimony for both the presence and cause of brain damage. For example, Florida case law recently reversed the disallowance of neuropsychologists testifying with regard to causation:

> Because the practice of psychology has expanded to the point where psychologists who are not [medical] doctors are increasingly becoming involved in areas which were traditionally considered to be purely medical, a blanket prohibition of testimony by psychologists concerning causation of brain injury no longer seems practical. Instead, the more prudent approach is to allow trial judges, in their discretion, to qualify psychologists and neuropsychologists to testify on causation as any otherexpert would be qualified to testify in his or her area of expertise. (School Board vs. Cruz, 2000)[3]

The Georgia legislature has also defined neuropsychologists as professionals who can diagnose and treat organic brain disorders (T. G. Burns, personal communication 10/15/01).

The view taken in this chapter is that psychology is well suited to litigation because hypothesis testing using psychological test data is consistent with the spirit of the judicial process and because test findings are open to empirical review. Controversy about psychological evidence historically involves two points; variations in the role of hypothesis testing as a basis of neuropsychological evidence and the use of statistical data as evidence of neuropsychological deficit. This chapter reviews salient aspects of these issues, and offers alternatives to currently disputed statistical procedures.

Statistics in the evidentiary process. In contrast to facts, statistics are tools used by experts for the formulation of opinions. While statistical evidence does not directly signify truth, it can be submitted to inferential methods that help estimate the truth of relevant hypotheses. Royall (1997) stated that statistical evidence refers to "which [hypothesis] is better supported. We might reasonably expect that strong evidence cannot be misleading very often" (p. 6). The degree to which statistical evidence constitutes legal evidence is determined by established "rules of evidence." These rules provide judges discretion in allowing or disallowing statistical information as evidence depending upon the circumstances of the case. Rules of evidence permit a judge to limit information provided to juries because jurors "are not totally

---

[3]Jurisdictions are beginning to certify neuropsychologists as psychology sub-specialists. The state of Georgia, for example, (Official Code of Georgia Annotated Section 43-39-1) defines neuropsychology as "concerned with the relationship between the brain and behavior, including the diagnosis of brain pathology through the use of psychological tests and assessment techniques."

rational, they must be shielded from exposure to information which is more likely to be deceptive than illuminating" (Stromberg, et al., 1988, p. 594).

Neuropsychological test scores reflect statistical reasoning at various levels. Statistical modeling is ubiquitous in psychological training, test development, and test interpretation. In best practice, the psychologist interprets statistical information gained from testing in such a way that designated parties in the court (i.e., judges and jurors) can infer "truth." The degree to which neuropsychological testimony aids in the inference of "truth" determines the value of that testimony to the court.

HYPOTHESIS TESTING AS A BASIS OF PSYCHOLOGICAL EVIDENCE

Two statistical models of hypothesis testing are used in psychology, frequentist and Bayesian. The frequentist model is concerned with how frequently certain observations can be expected to occur given a certain hypothetical distribution (such as the number of snake-eye rolls out of 10 tosses of two fair dice). There are two frequentists approaches, often referred to as the Fisherian and Neyman-Pearson schools (see chap. 3, this volume.) In contrast to the frequentist approach, the Bayesian approach (chap. 4, this volume) considers prior probability distributions as well as frequency distributions present at the time of the observation. What is more, the Bayesian approach permits estimates regarding the outcome of single, yet unobserved, events.

Frequentist Models of Hypothesis Testing

Fisher's Theory. According to Fisher, observed data need to be compared with a preselected critical region within a theoretical distribution using Null Hypothesis Significance Testing (NHST). Specifically, NHST yields the probability of finding the observed data—or data more extreme—if the theoretical distribution is assumed to be true. Because the *p* value can vary considerably depending on the extremity and the number of observations, its interpretation is confounded. A quarter of a century ago, ten of the world's leading applied statisticians, co-authored a paper explaining that trials contain large sample sizes provide stronger evidence than trials containing small sample sizes (Peto et al., 1976, p. 593). Rejection trials (also known as

Rejection Support Null Hypothesis Testing or RS-NHST, see chapter 3, this volume), is a second expression of Fisher's model, that includes an alternative hypothesis with the null. However, the alternative hypothesis associated with the null is simply a statement of what has not occurred.

Neyman-Pearson Theory. In contrast to Fisher's method, the Neyman-Pearson approach to NHST requires setting up a substantive alternative to the null hypothesis. Then, the method permits a choice between two alternative hypotheses given the evidence by evaluating the probability of the data under each of the contesting hypotheses. As in Fisher's method, the data are compared in terms of their abilities to predict long-run averages. Hence, both approaches are less concerned with the meaning or the value of the data than they are with the mathematical relationships between the sets of data.

The Statistical Relationship Between Test Findings and Diagnosis. Frequentist theories evaluate relationships between a distribution of sample data with hypothetical distributions. The typical task of the psychologist, however, is to reach a judgment on individual cases based on actual test findings. Table 5.1 presents the nominal descriptions that are used to indicate relationships between test findings and the presence or absence of a disorder.

TABLE 5.1
Nominal References for 2x2 Hypothesis Decision Matrix.

test findings +/-	*disorder presence+/-*	*nominal reference*
+	+	sensitivity
-	-	specificity
-	+	1-sensitivity
+	-	1-specificity

First defined by Yerushalami (1947), the terms 'sensitivity' and 'specificity' address the same issues of hypothesis testing described in Table 3.1 (chap. 3, this volume). Sensitivity refers to the proportion of the population with the disorder who test positive. Specificity refers to the proportion of the population without the disorder who test negative.

Ideally, psychological tests have both high sensitivity and high specificity. What these indexes have to say about the psychologist's judgment in an individual case also depends on the overall 'prevalence' (i.e., the base rate) of the disorder in the tested population (Meehl & Rosen, 1955).

The Bayesian Approach to Hypothesis Evaluation

As noted in chap. 4, this volume, Bayesian methods lead to the estimation of likelihood ratios and posterior probabilities of certain hypotheses. This section presents a brief description of both, considering the efficacy of them as evidence.

Estimates of Posterior Probabilities.

Three posterior probability estimates are described in the psychology literature (e.g., Elwood, 1993; Glaros & Kline, 1988).

Positive Predictive Value. (PPV) refers to the probability the patient has the disorder given a positive test result $p(D_o|P_o)$. According to Bayes' Theorem, PPV (Equations 5.1 and 5.2) is a posterior probability that depends on the prevalence of the disorder, the sensitivity of the test, and the overall probability of obtaining a positive test result. The latter is the sum of two probabilities: The test score could be positive given the presence (sensitivity) or the absence of the disorder (1-specificity).

$$PPV = \frac{\text{prevalence} * \text{sensitivity}}{\text{prevalence} * \text{sensitivity} + (1 - \text{prevalence}) * (1\text{-specificity})} \quad (5.1)$$

or

$$p(D_o|P_o) = \frac{p(D_o) * p(P_o|D_o)}{p(D_o) * (p(P_o|D_o) + (p(D_n) * p(P_o|D_n))} \quad (5.2)$$

Negative Predictive Value. (NPV) is the probability the patient does not have the disorder given a negative test result ($p(P_n|D_n)$) as demonstrated in Equation 5.3.

$$NPV = \frac{(1\text{-prevalence}) \times \text{specificity}}{(1\text{-prevalence}) \times \text{specificity} + \text{prevalence} \times (1\text{-sensitivity})} \quad (5.3)$$

Overall Predictive Value. (OPV) calculates the probability that a test taker's classification is correct (Equation 5.4) . The probability that a positive diagnosis is correct is the product of PPV and the prevalence of the disorder. The probability that a negative diagnosis is the product of NPV and the complement of the prevalence. OPV is the sum of these two products.

$$OPV = PPV * \text{prevalence} + NPV * (1\text{-prevalence}) \quad (5.4)$$

Clearly, OPV increases as PPV or NPV increase. The role of the prevalence of the disorder is less intuitive, because regardless of PPV and NPV, OPV increases as the prevalence becomes more extreme. If prevalence is .5, the a priori uncertainty regarding the presence of the disorder in the tested individual is at its maximum, which keeps OPV fairly low even if PPV and NPV are high. The question then is whether OPV is high enough to permit the claim that testing has improved the accuracy of the diagnosis beyond what it would be without testing. One way to evaluate such improvement is to ask whether OPV is superior to making a diagnosis randomly, on the basis of prevalence alone (Wiggins, 1973). If, for example, the prevalence of a disorder were .1, one might randomly make a positive diagnosis for every tenth client. The OPV of such a procedure would be .82 (i.e., $.9^2+.1^2$). It is difficult to justify this procedure because it amounts to non-optimal probability matching. If the assessor were to make a negative diagnosis in each case, OPV would be .9. The drawback of this method, of course, is that no positive diagnosis would ever be made, thus precluding any correct identification of the disorder. A good test with high sensitivity and specificity is a necessary tool if psychological assessors are to improve their diagnosis beyond these unsatisfactory alternatives. Such improvement becomes increasingly difficult as prevalence base rates become more extreme (which typically means as disorders become rarer).

To see if testing actually improves diagnosis, the ratio of OPV over the complement of the disorder's prevalence (1-prevalence) may be used (assuming that the disorder is rare; if it is frequent, the ratio is OPV/prevalence). Note that it is possible that both PPV and NPV are greater than their respective base rates (prevalence and 1-prevalence), while OPV is smaller than the accuracy one would achieve making uniformly negative diagnoses (i.e., 1-prevalence if $p(D_o|P_o) < .5$). As an example, consider the case in which the prevalence of the disorder is .2, sensitivity is .6 and specificity is .7. In this case, PPV/prevalence = 1.67, NPV/(1-prevalence) = 1.09, while OPV/(1-prevalence) = .96.

In psychological testing, the term Base Rate has two additional meanings. First, it refers to the frequency distribution of scores within populations – BR_f. For example, a standard score of 100 has a frequency distribution of 50% in the normal standardization sample. However the same score of 100 would have a frequency distribution of < 1% in the Huntington's disease sample (The Psychological Corporation, 1997, p. 147). BR_f may help the psychologist understand if a finding is rare, but it conveys no diagnostic information per se. Second, Base Rate is also used to describe the cumulative frequency of the score differences between tests in a given population – BR_d. Here, for example, differences of one standard deviation between Verbal and Performance IQ occur at a cumulative frequency of 15.5 % (The Psychological Corporation, 1997, p. 305). Again, BR_d provides no information about diagnosis, unless the rarity of these combined scores exclusively defines a disorder. Confounds occur with each of the three expressions of Base Rate. Prevalence Base Rates (BR_p) are problematic because even though estimates of disease prevalence exist they are sometimes disputed, and can vary considerably across cultures and geographic regions. What is more, if BR_p is known psychologists rarely know either the sensitivity or the specificity for most psychological tests associated with a specific diagnosis. BR_f and BR_d are more problematic because they convey no information that is unique, either to the patient or to specific disorders. Score frequencies are largely undefined for specific populations, and where they exist inconsistency is the rule rather than the exception. Remember Fisher's *Infrequency Admonition*.

Likelihood Ratios. As noted in chap. 3, this volume, likelihood ratios (λ) represent the ratio of the probabilities of the data under two hypotheses. When we compare the OPV (see Equation 5.4) of tests measuring two different diagnoses, we can consider (λ) as a measure of evidence for the first test (OPV_1) vis-à-vis the second test (OPV_2).

To the degree that two findings measure different diagnoses, the ratio of (λ) OPV constitutes a likelihood ratio such that:

$$(\lambda) = \frac{OPV_1}{OPV_2} \qquad (5.5)$$

For example, consider the comorbid diagnoses of oppositional defiant disorder (ODD) and attention deficit disorder (ADD). If I have a test for ODD with an OPV_1 of .92 given the patient's score on the ODD measure, and a second test for ADD having an OPV_2 of .72 given the patient's score on the ADD measure, then by substituting the values for ODD/ADD in Equation 5.5, (λ) = .92/.72. We could conclude that evidence supporting the diagnosis of ODD is stronger than the evidence supporting a diagnosis of ADD. Later in this chapter we will consider the strength of this evidence as well as the degree to which this evidence can be weak or misleading. This ratio provides an efficacious measure of statistical evidence for determining which psychological test best characterizes a diagnosis "beyond a reasonable degree of medical certainty" as defined by Brigham, Babitsky & Mangraviti, (1996). See Chapter 3, this volume for further discussion of medical certainty.

Likelihood ratios are problematic when conditional probabilities are equal for both the numerator and the denominator, resulting in the same λ when both OPV_1 and $OPV_2 = .9$ or when OPV_1 and $OPV_2 = .001$. In the first instance, a high degree of confidence is warranted. In the second instance, little confidence is warranted. Also, on those occasions where the denominator is zero, λ is undefined.

THE INADEQUACIES OF STATISTICAL HYPOTHESIS TESTING AS EVIDENCE

Both frequentist models reflect similar views regarding hypothesis testing. Because they are ubiquitously associated with null hypothesis testing in psychology, they both suffer from similar inadequacies as evidence. Royall (1997) argued that both the Neyman-Pearson model (p. 56) and Fisher's model (p. 79) produce invalid outcomes that can "lead to different results in two situations where the evidence is the same." Proofs supporting Royall's statements are beyond the scope of this work and interested persons should review his original text. It is important for readers to understand that his formative arguments represent a Bayesian perspective (see Chapter 3, this volume).

Null Hypothesis Significance Testing

Null hypothesis statistical testing (NHST) has become the standard for many scientific publications. Even so, there have been critics of the method since its inception. Tryon (1998) attributed the success of NHST to the ease at which it can be correctly calculated and interpreted. Yet, prominent psychometricians often misinterpret NHST results (Cohen, 1994) and documented misuse of the procedure has occurred for three decades (Dar, Serlin, & Omer, 1994). Most of the introductory psychology textbooks printed between 1965 and 1994 presented NHST inaccurately (McMan, 1995). Critics present three classes of problems associated with NHST: (a) the logical foundations (b) interpretation difficulties and (c) failure to also use supplementary or alternative inference methods (Krueger, 2001).

Logical Foundation. The logical foundation of null hypothesis testing was eloquently challenged by Howson and Urbach (1989), who viewed the inductive step of either accepting or rejecting a null hypothesis as capricious. Arbitrary decisions must be taken (i.e., proper statistic derived from "experience" or personal judgment) in order to render a conclusion. A leading advocate of the null hypothesis method proposed by Fisher once stated, "There is no answer to [the question] 'Which significance test should one use except the subjective one? Personal views intrude always" (Kempthorne, 1966, p. 12).

Interpretation Difficulties. Problems interpreting the Neyman-Pearson model lead to the introduction of a distribution having a critical region against which observations could be compared with a test-statistic selected a priori. Rejection criteria were recommended for this critical region and a hypothesis was "rejected" or "failed rejection" dependent on how sample data compared to this critical region. Unfortunately, the meaning of "accept and reject a hypothesis or reject and fail to reject" remain obscure. What is more, it is commonly assumed that acceptance or rejection of a hypothesis may be a function of the size of the sample rather than anything associated with the theory.

Howson and Urbach (1989) described many instances showing how the results of hypothesis testing using either the Fisher or Neyman-Person standard are in conflict with reality. Two well-known examples germane to psychology were reported by Cohen (1994). Meehl and Lykken (cited in Meehl, 1990) cross-tabulated 15 presumably unrelated items taken from 57,000 Minnesota high school students. All of the cross-tabulation were statistically significant, 96% of them at $p <$ .000001! Meehl (1990) thus noted that, "the notion that the correlation between arbitrarily paired trait variables will be, while not literally zero, of such minuscule size as to be of no importance, is surely wrong" (p. 212). In the second oft-cited example, Cohen reported a 2% incidence of schizophrenia in adults. One screening for schizophrenia has a sensitivity of 95% and specificity of 97%. Given a positive test for schizophrenia, and a test sensitivity of 95%, one might conclude that the patient with a positive test result has schizophrenia because there is less than a 5% chance the test is in error. However, given the low incidence of schizophrenia, the true probability that the case is normal is about .60 (calculated using the Bayesian model presented in Equation 1; see chap. 4, this volume, for a more detailed description of this problem). As noted by Howson and Urbach (1989): Well supported hypotheses are often rejected by a significance test. Inference by significance test also clashes with entrenched ideas about the nature of evidence, requiring the rejection of hypotheses that seem highly confirmed, allowing (in randomized tests) quite extraneous experiments such as the selection of cards from a pack to influence one's attitude toward hypotheses which have nothing to do with cards. (p. 175)

Psychologists who react to negative beliefs derived from rejection of a null hypothesis as though they are valid, "accept" null hypotheses by behaving as though they were true (Malgady, 1998). Subjective, and possibly unconscious or obscure, value judgment may enter into this inference. Nickerson (2000) cited other criticisms of NHST, which are summarized in Table 5.2

TABLE 5.2
Other Criticisms of Null Hypothesis Statistical Testing

a priori unlikelihood that H_0 is true
sensitivity of NHST to sample size
noninformativeness of test outcomes
inappropriateness of all-or-none decisions regarding significance
arbitrariness of the decision criterion test bias
possible inflation of Type I errors in the literature
presumed frequency of Type II errors
ease of violating assumptions of statistical tests
influence on experimental designs

The Directional Hypothesis. If a neuropsychologist has reason to suspect that scores from the patient's test protocol will be greater than or less than those obtained from the comparison group, a directional hypothesis is sometimes used. The directional hypothesis is evaluated using a one-tailed statistical test that compares scores with only one half of the theoretical distribution. The one-tailed test effectively doubles the likelihood of finding a "significant" effect (Dietrich & Kearns, 1983). Because the directional hypothesis does not allow inferences in cases where findings are the opposite of those predicted, it should be avoided in clinical practice.

Despite the ongoing criticisms of NHST, the methods have also found its apologists. Hagen (1998), for example, attributes shortcomings to improper use by evaluators. He observed that "the null hypothesis is not a statement about the sample (i.e., the patient), [it] is a statement about the population [e.g., standardization sample] from which the sample is drawn" (p. 801). When investigators and clinicians render conclusions about patients from null hypothesis test results, they make inappropriate attributions. Table 5.3 presents Nickerson's (2000) synopsis of reasons that he believes NHST has withstood the many criticisms (see also Krueger, 2001).

TABLE 5.3
Reasons Null Hypothesis Statistical Testing Remains Impervious to Criticism

lack of understanding of NHST or confusion regarding conditional probabilities
the appeal of formalism and the appearance of objectivity
the need to cope with the threat of chance
deep entrenchment of the approach within the field [of psychology] as evidenced in the behavior of advisers, editors, and researchers
it appears to provide the user with a relatively simple and straightforward method for separating meaningful and irrelevant data

Standardization Samples in Psychology

Perhaps because of the ubiquity of NHST in psychology research and test standardization, psychologists often make assumptions about data reported as "norms" for psychological tests. First, they assume that data described by means and standard deviations are measured using equal intervals (i.e., 30 - 35 is the same score difference as 100 - 105). Second, they assume that the distances between scores are equal across tests (i.e., 85 - 95 on Test A: 35 - 45 on Test B). Three, data are assumed to follow a "bell curve" distribution. Spreen and Strauss (1998) published reviews of more than 91 tests and measures that are used by neuropsychologists. Table 5.4 presents an analysis of the standardization samples for tests described therein, clustered into three principle groups. Education refers to tests that were developed for, or whose development has been significantly influenced by, school classification requirements (viz., PL 94-142, PL 99-457; see Sattler, 1988). Specialty tests were developed for special populations, primarily psychiatric inpatients and outpatients. Npsych. addresses those measures that were designed in neuropsychology laboratories (See the web page www.geocities.com/rdfphd/index.htm).

Of the tests described by Spreen and Strauss (1998), three groups were co-normed (viz.,Woodcock-Johnson Psychoeducational Test Battery; the Wechsler Intelligence Scale for Children - III [WISC-III] and the Wechsler Individual Achievement Test [WIAT]; and, The Wechsler Adult Intelligence Scale III [WAIS-III], The Children's Memory Scale [CMS], and the Wechsler Memory Scale III [WMS-III]). Co-norming refers to administration of multiple tests to the same individuals within the same block of time. Co-norming is described as "linking samples" by the Psychological Corporation. A rather large group of children (N = 1,284) participated in the linking sample for the WIAT with intelligence tests (WISC-III, WPPSI-R, or WAIS-R). No adults participated in the link. The Children's Memory Scale also has "linking samples" with WISC-III and WPPSI-R using 108 children in each of five age groups (Cohen, 1994). The WAIS-III/WMS-III Technical Manual (the Psychological Corporation, 1997, p. 16) notes that the standardization group for the WMS-III consists of half the WAIS-III standardization group.

TABLE 5.4
Summary of Sample Sizes Reported by Spreen and Strauss (1998)

Category	*Total*	*Specialty*	*Education*	*Npsych.*
Mean sample size	2001	2519	2951	548
Standard deviation	5603	2974	1061	965
Ratio /σ	2.79	1.80	0.54	1.76
Census match %	20.8	36.0	100	02.9

Several problems arise when parametric statistics are used to evaluate scores taken from tests that were not co-normed. First, age cohorts vary inconsistently and unpredictably across and within samples. Second, cell sizes (typically age cohorts) are often insufficient to generalize information to other populations (cell sizes as small as 2 were reported by Spreen & Strauss, 1998.). Third, standardization samples cannot be generalized due to the persistent use of poorly defined "available" subjects (less than 2% of the samples were large enough to permit meaningful stratification). Fourth, only the Education group provides samples having a standard deviation less than the mean. This suggests that only tests in the Education group can be meaningfully compared using the standard deviation as a measure of effect. These criticisms are particularly germane to the neuropsychology measures. Consequently, direct comparisons of standardized scores across neuropsychological measures is risky. This does not mean that scores from the neuropsychology group are uncomparable, only that direct comparison of average scores or difference scores from a reference mean introduces considerable unknown error. When conditions such as these exist, comparisons using nonparametric statistical tests are recommended. As noted by Cliff (1996):

> The calculations of the power of statistics and the relative power of different statistics must also be done on the basis of assumptions about the characteristics of the data. When the parent distributions have characteristics different from those assumed, the absolute and relative powers can consequently be different. Since we often have reason to believe that our data do not conform to classical (i.e., parametric) assumptions, working with statistics that have good power under a broad range of situations is preferable to using ones that have optimum power under a narrow range of special ones. (p. 17)

Bayesian Methods

In considering data as evidence, the classical Bayes model as presented in chap. 4, (this volume) is problematic as well. Different prior probabilities lead to different conclusions. When prevalence data are available, multiple estimates may exist. Bayesian findings are further limited in their evidentiary value because they overly rely on belief, especially when prior probabilities are subjective. A final criticism of the Bayesian model is based on the difficulty of representing complete ignorance by a probability distribution and ignorance represents a form of prior belief. Consider the value calculated for positive predictive power in Equation 1 when the prevalence equals zero (a form of complete ignorance or disbelief in the existence of a disorder).

Impeachment

Inconsistent testimony can produce impeachment. Opposing attorneys may attempt to discredit a witness's credibility whenever the witness presents strong evidence for or against a client. Videotaped depositions, reviews of prior testimony in deposition for the current or prior trials, discrepancies between published statements and evidentiary statements, and "mousetrap" questioning can result in confusion and a claim of impeachment (Babitsky & Mangraviti, 1999).

The *p* value has two distinct and conflicting roles in NHST (see chap. 3, this volume). First, it measures the strength of evidence. Second, it represents the probability of obtaining misleading evidence. Because of these conflicting roles, impeachment accompanies all null-hypothesis-based tests of statistical significance. The Bayesian approach (see chap. 4, this volume) has been recommended as a remedy for circumventing this problem (Box & Tiao, 1992; Gouvier, Hayes & Smiroldo, 1998; Harlow, Mulaik, & Steiger, 1997; McCaffrey et al , 1997). One specific form of Bayesian analysis, likelihood ratios, measures evidence in a way that mitigates against impeachment.

STATISTICAL EVALUATION USING LIKELIHOOD RATIOS AS EVIDENCE OF NEUROPSYCHOLOGICAL DEFICIT

Theoretical Basis for Using Likelihood Ratios as Evidence

The likelihood ratio (λ) entails three interrelated notions: the likelihood principle, the likelihood function, and the law of likelihood (Royall, 1997). The *likelihood principle* supposes that when presented with two sets of data, the likelihood of selecting one observation of the same value from either group is equal. The *likelihood function* refers to the likelihood of selecting a single value from a group of values. When the likelihood of observing a specific value is greater in one group than the other, then the likelihood function provides evidence supporting selection from the group having greater likelihood of containing the value of interest; hence, the function provides support for one group vis-à-vis the other. The strength of evidence supporting one group over the other is supported by evidence from the likelihood ratio. The *law of likelihood* applies to two hypotheses and indicates when a given set of observations is evidence for one hypothesis versus the other. It explains how observations should be interpreted as evidence for A vis-à-vis B, but makes no mention of how those observations should be interpreted as evidence in relation to A alone. Neither of the two NHST models (viz., *p* values or rejection trials) provides adequate evidence for hypothesis testing because each is based on the law of improbability or the law of changing probability. The law of likelihood is based on the Bayes model. But, unlike Bayesian probability models, which are useful only when prior probabilities are known, likelihood ratios only evaluate current data. Likelihood ratios remain stable irrespective of prior probabilities. The law of likelihood effectively defines the concept of statistical evidence as it relates to comparison with another set of observations. The question "When do test observations support one conclusion or another" is best answered by the law of likelihood.

Calculation of the likelihood ratio compares two conditional probabilities as expressed in Equation 5.5. Phrased as a question, the ratio asks "Given a score of x, which diagnosis is more likely, A or B?" Royall (1997) equated statistical evidence with this likelihood function: "The evidence in the observations is represented by the likelihood function, and is measured by likelihood ratios. It is not represented and measured by probabilities, either frequentist sample-space probabilities or Bayesian posterior probabilities" (p. 176). Two applications of

likelihood are salient to psychology: evaluation of the evidentiary strength of specific neuropsychological measures and evidentiary support of testing specific dependent variables arising out of the individual assessment.

CALCULATING EVIDENCE

When the psychologist considers test findings as evidence, tests can be grouped into three classes based upon the amount of information available in the research literature. The first group we will call *Diagnostic Tests* because they have known sensitivity and specificity for a disorder with a known prevalence. The second group, *Abilities Tests*, are associated with some and possibly multiple disorders having known prevalence, but the sensitivity and specificity information are either unknown or unassociated. The third group, *Construct Tests*, have theoretical relationships with theoretical disorders that have unknown prevalence and no known sensitivity or specificity. In those situations where it is possible to select tests in advance, the psychologist will use *Diagnostic Tests* if they are available.

Evidentiary Strength of Diagnostic Tests. Diagnostic Tests vary in their ability to identify pathological processes. Occasionally inferences about their relative efficacy can be gained form their demonstrated sensitivity and specificity for a given diagnoses. This group of tests is most appropriate for Bayes analysis as described earlier. The psychologist calculates OPV for each test and associated diagnosis, then compares the OPV using ratios as described earlier. See Equation 5.5 and the associated discussion.

Evidentiary Strength of Abilities Tests. Perugini, Harvey, Lovejoy, Sandstrom, K., & Webb. (2000) present a method for calculating OPV for tests having known specificity and sensitivity, but unknown prevalence (see Equation 5.6).

$$OPV = \frac{\text{Sensitivity} + (1\text{-Specificity})}{N} \qquad (5.6)$$

According to E. A. Harvey (personal communication, May 21, 2002) this formula uses sample size to compensate for unknown prevalence. The OPV allows the psychologist to compare these findings with those of other tests for which sensitivity, specificity, and sample sizes are known. When comparing *Abilities Tests* with *Clinical Tests*, this formulation should serve as a lowest common denominator.

Evidentiary Strength of Construct Tests. When the psychologist must use a test having unknown sensitivity and specificity, then the standardized percentile rank can be used to compare the patient's relative standings on various tests[4]. When it is necessary to compare Construct Tests with other types of test, the comparison should be made at the percentile rank level. Diagnosis can be made using this process by comparing historical information and psychological test scores to diagnostic criteria listed in the most current version of Diagnostic and Statistical Manual of Mental Disorders (DSM) or the International Classification of Disease (ICD). Neither nosology incorporates psychological test findings in their inclusion or exclusion criteria for most diagnosis. Table 5.5 provides an example for calculating evidentiary strength of construct tests, λd (as defined in Equation 5.6) for the differential diagnosis of Dementia of the Alzheimer's Type (DAT) and Traumatic Brain Injury (TBI) using data presented in the WAIS-III/WMS-III Technical Manual.

The ratio shows a greater likelihood that the patient's observed standard scores (SS) on one measure (Auditory Immediate Memory) are due to DAT rather than TBI (i.e., $\lambda d = 27/2 = 9.04$)[5]. Calculations such

[4]The standardized percentile score corrects for direction of deficit, so that all scores compared have either high or low scores reflective of poor performance. For example, in comparing a WAIS-III performance percentile ranking of 75 with theTrails B percentile score of 60, the Trails B score must be subtracted from 100 (i.e., 100 - 60 = 40) in order to standardize the scores because high scores indicate poor performance on Trails B whereas low scores reflect poor performance on the WAIS-III. The resulting comparison of 75/40 produces an Index score (?) of 1.87, a difference that is too small to establish significance "beyond a reasonable degree of medical certainty."

[5]Royall (1997) presented proofs that ratios of 8 and 1/8 are consistent with 5% level of risk. On occasions requiring "quite strong"

as these are easily performed on spreadsheets (the spreadsheet for this table is available from the web page at http: / / www . geocities . com / rdfphd / index . html).

TABLE 5.5.
Likelihood Ratios from WMS-III for Differential Diagnosis[6]

test	*SS*	*DAT*			*TBI*			*DAT/TBI*
			σ	%		σ	%	λ_d
AI	62	68.7	11.0	27	98.3	19.3	02	**09.04**
VI	65	70.6	10.9	30	74.9	13.9	24	01.27
IM	55	62.9	11.4	24	78.9	17.7	09	02.76
AD	64	66.1	09.6	41	89.6	21.8	12	03.44
VD	72	67.5	08.1	71	74.3	13.9	43	01.63
ARD	65	65.6	08.6	47	93.6	16.6	04	**11.10**
GM	62	60.4	08.6	37	81.9	16.5	11	05.03
WM	74	80.4	16.6	35	91.9	11.9	07	05.28

Evidentiary Support for MMI (λm). Establishing that a patient has attained MMI implies that the patient's condition has become static and no significant additional change is likely. Table 5.6 provides comparisons of the data from Table 5.5 with a second evaluation occurring one year later. Evidence provided by λm indicates the patient has reached MMI ($\lambda m < 8$ in all categories), conditional upon the

evidence, he recommended ratios of 32 and 1/32 where the level of risk is .001. Data are interpreted using the boundries of $\lambda = 8$ or $\lambda = .125$.

[6]For each score above the 50th percentile, you must add λ to the λ for the cumulative posterior percentilews as demonstrated on the corresponding spreadsheet on the internet web page at HTTP//www. geocities.com/rdfphd/index.htm).

assumption that after one year no further change is expected, where λm = Percentile Rank 1/Percentile Rank 2 (i.e., Auditory Immediate λm = .271/.302 = 1.114).

TABLE 5.6
Determination of MMI

	Adm. Standard Scores		*Std. Sample*		*Percentile Ranks*		
	1	2	Mean	SD	1	2	(λ_m)
Aud. Immediate	62	63	68.7	11.0	0.271	0.302	1.114
Visual Immediate	65	60	70.6	10.9	0.304	0.165	0.545
Immediate Memory	55	45	62.9	11.4	0.244	0.058	0.238
Aud. Delayed	64	74	66.1	9.6	0.413	0.795	1.922
Visual Delayed	72	82	67.5	8.1	0.711	0.963	1.355
Aud. Rec. Delayed	65	73	65.6	8.6	0.472	0.805	1.705
General Memory	62	71	60.4	8.6	0.574	0.891	1.553
Working Memory	74	81	80.4	16.6	0.350	0.514	1.470

Evidentiary Support for Posterior Function (λe). The comparison of posterior with current function has been the subject of many studies[7], yet controversy remains regarding the best way to evaluate this type of change. Table 5.7 provides an example of the equivalent evaluation ratio (λe) for this purpose. Here, all tests have the same mean and standard deviation. Rankings are again compared with λe reflecting WISC percentiles/WAIS percentiles. Rather than predict prior abilities, this function evaluates the differences between current and prior scores. The prior scores can originate in academic and other records (preferred) or derivations from predictive equation. In this example, both Digit Span and Block Design reflect boundary differences of sufficient magnitude to opine that abilities were superior during the earlier evaluation.

[7]PsycSCAN lists 223 documents between 1990 and 2001 containing the keyword *premorbid*.

TABLE 5.7
Comparison to Posterior Function

Test	*Battery SS*		*Percentile Rank*		λ_e
	WISC	*WAIS*	*WISC*	*WAIS*	
Vocabulary	112	98	0.79	0.45	1.76
Information	92	87	0.30	0.19	1.54
Similarities	101	83	0.53	0.13	4.10
Comprehension	98	92	0.45	0.30	1.51
Digit Span	**105**	**78**	**0.63**	**0.07**	**8.85**
Arithmetic	99	82	0.47	0.12	4.11
Picture Completion	113	95	0.81	0.37	2.18
Picture Arrangement	110	101	0.75	0.53	1.42
Object Assembly	107	94	0.68	0.34	1.97
Block Design	**114**	**74**	**0.82**	**0.04**	**19.86**

DETERMINING IF EVIDENCE IS MISLEADING

The interpretation of likelihood ratios (λ) requires consideration of the magnitude of the ratio as well as its likelihood of producing weak or misleading evidence. Weak (*w*) evidence refers to low likelihood ratios. Misleading (*m*) evidence refers to data supporting hypothesis A when Hypothesis B is in fact true. Royall (1997) recommended using the bounds of < λ 1/8 (.125) or > 8 as evidence of preference for one hypothesis vis-à-vis the other. Table 5.8 presents the probability of obtaining misleading or weak evidence with a λ of 8:1/8 and 32:1/32 as adapted from Royall (p 97).

Returning to Table 5.5, we see evidence for every test that the patient' standard scores are more like both DAT and TBI patients than the "normal" standardization sample (based on the "normal" standardization sample mean score of 100 and standard deviation of 15).

From Table 5.7 we can conclude (using Table 5.8) that with eight paired observations (col. 1) the evidence is not likely to be misleading (m = .01|λ = 8:1/8 or m = .001| λ = 32:1/32), but it may be weak (w = .046|λ = 8:1/8 or w = .202 |λ = 32:1/32). We can also conclude that the test scores alone poorly discriminate between DAT and TBI but they provide good evidence that the patient differs from "normal."

TABLE 5.8
Probabilities of Misleading or Weak Information

# obs	λ = 8 or 1/8		λ = 32 or 1/32	
	m	*w*	*m*	*w*
7	.005	.143	.005	.143
8	.010	.046	.001	.202
9	.003	.083	.003	.083
10	.006	.026	.002	.120
11	.002	.048	.001	.048
12	.004	.016	.001	.072

Evaluating Multiple Measures

Sequential Bayesian Analysis. The preceding discussion considered relationships between two data sets or a single score with a reference group. Although many comparisons of test scores employ these models, psychologists also evaluate single abilities or constructs using multiple measures. It is not uncommon to include several measures of attention or memory in a single test battery. Spreen and Strauss (1998) point out that "The process of clinical interpretation takes into account not only the probabilities of individual test results, but also the combination of many test results, the observations during the process of testing, the question posed for the examiner, and the characteristics of a specific disorder"(p. 28). Bayesian models imply that a sequential analysis of data allow each observation or score to serve as an estimation of posterior probabilities for subsequent observations or scores (see Equations 4.5 – 4.7 and accompanying text, this volume). Therein, each new piece of data would increase the positive predictive power of the test battery. However, an empirical study of ADHD appears to contradict this view.

Perugini, et al. (2000) compared permutations of seven tests in their abilities to classify 43 children as either ADHD or control. Table 5.9 presents analysis of the Perugini et al. data using Bayesian definitions of PPV, NPV, OPV, and the ADHD prevalence of .04 (American Psychiatric Association, 2000, p. 90).Using Perugini et al. data, Bayes predictive measures of empirical findings do not support the conventional view that using more tests improves predictive accuracy. This analysis shows best OPV when only 1 test, Trials, is used. Failure in predictive enhancement is likely a consequence of unknown measurement error that may exceed predictive power of the tests. In Bayesian analysis both error and predictive power are magnified. One method for controlling error in Bayesian decision making is to employ a Bayesian network.

Bayesian Network Analysis. Bayes networks evaluate believed relations between sets of variables relevant to some problem. They combine discrete, continuous and propositional (true/false) variables using algorhythms that are capable of incorporating multiple and varied information sources. Figure 5.1 presents a Bayes network for the diagnosis of ADD and related disorders that incorporates the Perugini et al. data with other measures. This analysis is available for download from the web site at http://www.geocities.com/rdfphd/index.html. The network analysis makes no assumptions regarding data distributions. Prior knowledge is included in the decision process. Findings allow inferences about patients rather than data. Findings are simply presented in diagrams. Findings are highly accurate when networks are properly constructed. Networks produce no undefined results. For psychology, the networks are superior to algorithms available for DSM evaluation because they include objective and test data. Also, as new information emerges, networks can "learn" and improve accuracy. This is accomplished by introducing either new cases or research findings. Error is controlled by using multiple data or observation sources .

Currently software for developing and analyzing networks is available. Expert system software, such as *Netica* (www.norsys.com), provides an excellent visual presentation of data that can be understood by most jurors and other consumers of psychological information. At this time Bayes networks provide the most promising method of statistical analysis as evidence in forensic and neuropsychology. There is, however, a pressing need for the construction and validation of Bayes networks for psychological diagnosis.

TABLE 5.9
Bayesian predictive measures for Perugini et al. *

Tests	*Sensitivity*	*Specificity*	*PPV*	*NPV*	*OPV*
1/7	.76	0.45	.06	0.95	.94
2/7	.62	0.91	.20	0.99	.77
3/7	.28	1.00	.25	1.00	.00
4/7	.05	1.00	.05	1.00	.00
1/3	.76	0.59	.07	0.97	.92
2/3	.52	1.00	.52	1.00	.00
3/3	.10	1.00	.10	1.00	.00
Trails	.29	0.91	.09	0.99	.88
Digit Span	.38	0.95	.18	0.99	.76
CPT Index	.67	0.73	.09	0.98	.90
Trails	.70	0.51	.05	0.96	.94

*Base rate of ADHD = .04 (American Psychiatric Association, 2000,p. 90)

CONCLUSIONS

The expertise of a psychologist in court is dependent on the evaluator's ability to present statistical evidence as legal evidence that is clear to the trier of fact. Because presentation of complex statistical processes often confuses jurors, psychologist should avoid statistical explanations when possible. However, a thorough understanding of statistical constructs may be necessary during cross examination. Statistical theory is so important in the design, development, and evaluation of psychological tests that psychologists who are unprepared to explain them risk not only professional embarrassment, but could be held liable for malpractice should their client "loose" as a consequence of their poor preparation.

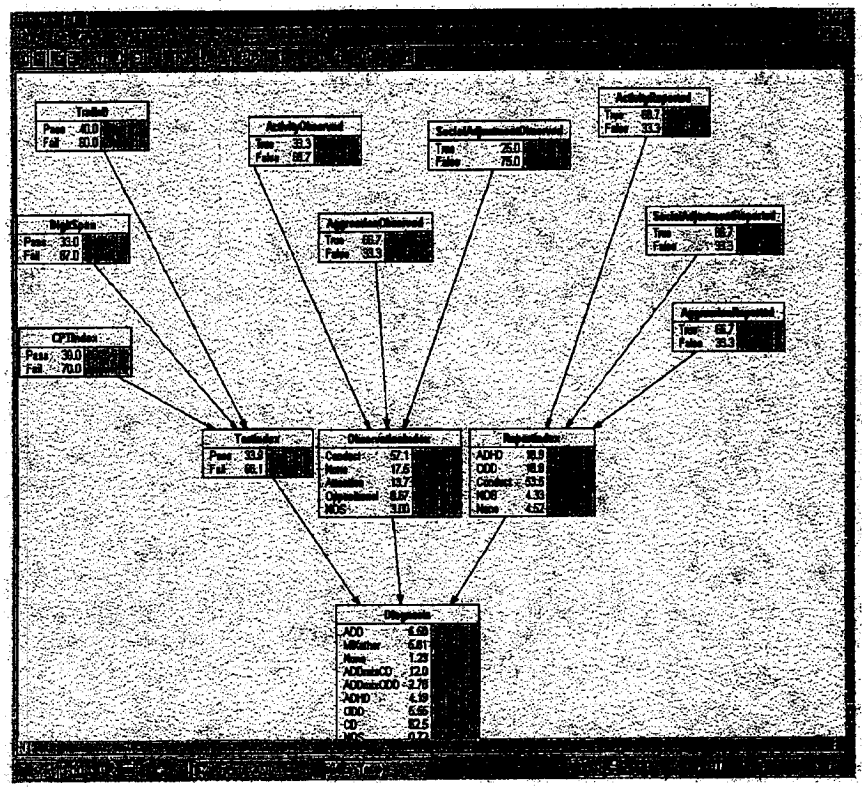

Figure 5.1. Bayes network

Statistical findings constitute evidence when they assist in demonstrating the superiority of one set of data vis-à-vis another. Reliance on NHST as the basis for establishing statistical evidence constitutes impeachment because the measure of significance employed probability model is of value in establishing the efficacy of a test when (p) has two different and contradictory meanings. Bayes' posterior prevalence for a neuropsychological disorder is well established, but has questionable value as evidence for the superiority of one set of test results vis-à-vis a second set or when prevalence is unknown. Likelihood ratios provide the best extant model for comparing two sets of data as evidence, but can be challenged because their product can be undefined. Bayes networks offer the most viable analytic alternative for psychological evidence inforensic and neuropsychology but networks need to be constructed and verified.

CLINICAL IMPLICATIONS

When it is necessary to demonstrate that the difference observed between two test scores is significant *beyond reasonable degree of medical certainty*, the simplest calculation is pair-wise division of standardized percentile scores[8]. The quotient of this division functions as a likelihood ratio (λ). Table 5.10 presents percentile scores when λ = 8 and 32 that are appropriate for pair-wise comparisons. As the table indicates, when the percentile score from Test A = 50 (SS = 100), it would be necessary for the patient to produce a percentile score of 6.25 (SS ~77) on the second test in order to demonstrate a difference that is "beyond a reasonable degree of medical certainty." A percentile score of 0.781 on Test B (SS ~ 62) would be necessary to establish a λ of ~32, a difference that is "well beyond a reasonable degree of medical certainty."

Calculation in this way may be less precise than regression-based analysis, but unlike regression models, the analysis allows comparison across data types when correlation coefficients between tests are unknown or when inadequate information exists for calculating Bayesian posterior probabilities. Even qualitative rankings from ordinal data types can be compared in this manner. For example, the classifications of Mild, Moderate, or Severe that are gleaned from hospital, rehabilitation, or other medical records can be assigned percentile values of 10%, 5%, and 1% respectively to indicate their relative relationships. Evaluating findings in this way tends to be more conservative than regression based methods, thereby allowing the clinician to place high confidence in the results.

[8]The standardized percentile score corrects for direction of deficit, so that all scores compared have either high or low scores reflective of poor performance. For example, in comparing a WAIS-III performance percentile ranking of 75 with theTrails B percentile score of 60, the Trails B score must be subtracted from 100 (i.e., 100 - 60 = 40) in order to standardize the scores because high scores indicate poor performance on Trails B whereas low scores reflect poor performance on the WAIS-III. The resulting comparison of 75/40 produces an Index score (λ) of 1.87, a difference that is too small to establish significance "beyond a reasonable degree of medical certainty."

At the time of this chapter's publication, psychologists in clinical settings rely almost exclusively on statistical models that originated in the 19th century. Isn't it time to move on and develop Bayes networks as well as other forms of modern analysis as aids to clinical diagnosis?

TABLE 5.10
Percentile Score Differences Between Value A and Value B Required for Significance When λ = 8 and λ = 32.

A%	B%@ λ= 8	B%@ λ= 32
1	00.125	00.016
5	00.625	00.078
10	01.250	00.156
15	01.875	00.234
20	02.500	00.313
25	03.125	00.391
30	03.750	00.469
35	04.375	00.547
40	05.000	00.625
45	05.625	00.703
50	06.250	00.781
55	13.125	03.128
60	20.625	05.156
65	28.750	07.118
70	37.500	09.375
75	46.875	11.719
80	56.857	14.219
85	67.500	16.857
90	78.750	19.688
95	90.625	22.656

REFERENCES

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders*. (4th ed.). Washington DC: Author.

Babitsky, S., & Mangraviti, J. J. (1999). *How to excel during depositions*. Falmouth, MA: SEAK, Inc.

Barth, J. T., Ryan, T. V., & Hawk, G. L. (1992). Forensic neuropsychology: A reply to the method skeptics. *Neuropsychology Review, 2*, 251-266.

Box, G. E. P. & Tiao, G. C. (1992). *Bayesian inference in statistical analysis*. New York: Wiley.

Brigham, C. R., Babitsky, S., & Mangraviti, J. J. (1996). *The independent medical evaluation report*. Falmouth, MA: SEAK, Inc.

Chan, R. C. K. (2001). Base rate of post-concussion symptoms among normal people and its neuropsychological correlates. *Clinical Rehabilitation, 15*(3), 266-273.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62*, 75-82.

Dietrich, F. H., & Kearns, T. J. (1983). *Basic statistics*. San Francisco: Dellen.

Elwood, R. W. (1993). Clinical discriminations and neuropsychological tests: An appeal to Bayes' Theorem. *The Clinical Psychologist, 7*, 224-233.

Faust, D., Ziskin, J. & Hiers, J. B. (1991). *Brain damage claims: Coping with neuropsychological evidence*. Los Angeles: Law and Psychology Press.

Fisher, R. A. (1959). *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of Clinical Psychology, 44*, 1013-1023.

Glenberg, A. M. (1996). *Learning from data: An introduction to statistical reasoning*. (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Gouvier, W. D., Hayes, J. S., & Smiroldo, B. B. (1998). The significance of base rates, test sensitivity, test specificity, and subjects' knowledge of symptoms in assessing TBI sequelae and malingering. In C. R. Reynolds (Ed.), *Detection of malingering during head injury litigation* (pp. 55-80). New York: Plenum.

Hagen, R. L. (1998). A further look at wrong reasons to abandon statistical testing. *American Psychologist, 53*, 797-798.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. La Salle, Il: Open Court.

Kempthorne, O. (1966). Some aspects of experimental inference. *Journal of the American Statistical Association, 61*, 11-34.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56*, 16-26.

Malgady, R. G. (1998). In praise of value judgments in null hypothesis testing and of "accepting" the null hypothesis. *American Psychologist, 53*, 797-798.

Martens, M., Donders, J., & Millis, S. R. (2001). Evaluation of invalid response sets after traumatic head injury. *Journal of Forensic Neuorpsychology, 2*(1), 1-8.

McCaffrey, R. J., Williams, A. D., Fisher, J. M., & Laing, W. C. (1997). *The practice of forensic neuropsychology: Meeting challenges in the courtroom*. New York: Plenum.

McMan, J. C. (1995, August). *Statistical significance testing fantasies in introductory psychology textbooks*. Paper presented at the 103rd Annual Convention of the American Psychological Association, New York.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108-141.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194-216.

Merriam Webster, Inc. (1996). Merriam Webster's collegiate dictionary (10th ed.). Springfield, MA: Author.

Mitrushina, M. N., Boone, K. B., & D'Elia, L. F. (1999). *Handbook of normative data for neuropsychological assessment*. New York. Oxford University Press.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

Ofloff, J. R. P., Beavers, D. J., & DeLeon, P. H. (1999). Psychology and the law: A shared vision for the 21st century. *Professional Psychology: Research and Practice, 30*, 331-332.

Perugini, E. M., Harvey, E. A., Lovejoy, D. W., Sandstrom, K., & Webb, A. H. (2000). The predictive power of combined neuropsychological measures for Attention-Deficit/Hyperactivity Disorder in children. *Clinical Neuropsychology, 6*, 101-114.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., & Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prononged observation of each patient, I: Introduction and design. *British Medical Journal, 34*, 585-612.

Putzke, J. D., Williams. M. A., Glutting, J. J., Konolid, T. R., & Boll, T. J. (2001). Developmental memory performance: Inter-task consistency and base-rate variability on the WRAMAL. *Journal of Clinical & Experimental Neuropsychology, 23*(3), 253-264.

Rosenfeld, B., Sands, S. A., & Van Gorp, W. G. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives of Clinical Neuropsychology, 15*(4), 349-359.

Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.

Sattler, J. M. (1988). *Assessment of children*. (3rd ed.) San Diego, CA: Author.

School Board vs. Cruz. 25 F.L.W. D1085. (Fla. 5th DCA, 2000).

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms, and commentary*. (2nd ed.) NY: Oxford University Press.

Stromberg, C. D., Haggarty, D. J., Mishkin, B., Leibenluft, R. F., Rubin, B. L., McMillian, M. H., & Trilling, H. R. (1988). *The Psychologist's Legal Handbook*. Washington, DC: The Council for the National Register of Health Service Providers in Psychology.

The Psychological Corporation. (1997). *WAIS-III/WMS-III technical manual*. San Antonio, TX: Author.

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education*. (6th ed.). Upper Saddle River, NJ: Merrill.

Tryon, W. W. (1998). The inscrutable null hypothesis. *American Psychologist, 53*, 796-807.

Wiggins, J. S. (1973). *Personality and prediction*. Reading, MA: Addison-Wesley.

Williams, J. M. (1997). The forensic evaluation of adult traumatic brain injury. In R. J. McCaffrey, A. D. Williams, J. M. Fisher & W. C. Laing, (Eds.). *The practice of forensic neuropsychology: Meeting challenges in the courtroom*. (pp. 37- 70). New York: Plenum.

Yerushalami, J. (1947). Statistical problems in assessing methods of medical diagnosis. *Public Health Reports, 62*, 1432-1449.

Ziskin, J., & Faust, D. (1998). *Coping with psychiatric and psychological testimony*. Los Angeles: Law and Psychology Press.