- **Major Contribution**

The Meta-Analysis of Clinical Judgment Project:

Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction

Stefanía Ægisdóttir
Michael J. White
Paul M. Spengler
Alan S. Maugherman
Linda A. Anderson
Robert S. Cook
Cassandra N. Nichols
Georgios K. Lampropoulos
Blain S. Walker
Genna Cohen
Jeffrey D. Rush
Ball State University

Clinical predictions made by mental health practitioners are compared with those using statistical approaches. Sixty-seven studies were identified from a comprehensive search of 56 years of research; 92 effect sizes were derived from these studies. The overall effect of clinical versus statistical prediction showed a somewhat greater accuracy for statistical methods. The most stringent sample of studies, from which 48 effect sizes were extracted, indicated a 13% increase in accuracy using statistical versus clinical methods. Several variables influenced this overall effect. Clinical and statistical prediction accuracy varied by type of prediction, the setting in which predictor data were gathered, the type of statistical formula used, and the amount of information available to the clinicians and the formulas. Recommendations are provided about when and under what conditions counseling psychologists might use statistical formulas as well as when they can rely on clinical methods. Implications for clinical judgment research and training are discussed.

A large portion of a counseling psychologist's work involves deciding what information to collect about clients and, based on that information, predicting future client outcomes. This decision making can occur both at the microlevel, such as moment-to-moment decisions in a counseling session, and at the macrolevel, such as predictions about outcomes such as suicide risk, violence, and response to treatment (Spengler, Strohmer, Dixon, & Shivy, 1995). Because the quality of client care is often determined

THE COUNSELING PSYCHOLOGIST, Vol. 34 No. 3, May 2006 341-382
DOI: 10.1177/0011000005285875

by the accuracy of these decisions (Dawes, Faust, & Meehl, 1989; Meyer et al., 1998; Spengler, 1998), determining the best means for decision making is important.

Two major approaches to decision making have been identified: the clinical and the statistical, which is also called mechanical (Dawes et al., 1989). Clinical prediction refers to any judgment using informal or intuitive processes to combine or integrate client data. Psychologists use the clinical method when their experience, interpersonal sensitivity, or theoretical perspective determines how they recall, synthesize, and interpret a client's characteristics and circumstances.

Such intuitive or "gut-level" inferences are greatly reduced in the statistical approach. Predictions are based on empirically established relations between client data and the condition to be predicted (Dawes et al., 1989). A psychologist who declares that his or her clinical impression suggests a client may be suicidal has used the clinical method. By contrast, when using the statistical method, client data are entered into formulas, tables (e.g., actuarial tables), or charts that integrate client information with base rate and other empirical information to predict suicide risk. While the statistical method is potentially 100% reproducible and well specified, the clinical method is neither as easily reproduced nor as clearly specified (Grove, Zald, Lebow, Snitz, & Nelson, 2000).

Meehl (1954) contended that while the clinical method requires specific

Alan S. Maugherman is now in private practice at the Center for Psychological Development, Muncie, Indiana. Linda A. Anderson is now at the University Counseling and Psychological Services, Oregon State University. Robert S. Cook is now in private practice at Lifespan, North Logan, Utah. Cassandra N. Nichols is now at the Counseling and Testing Services, Washington State University. Blain S. Walker is now at the Tripler Army Medical Center, Honolulu, Hawaii. Genna R. Freels is now at the Louisiana Professional Academy, Lafayette, Louisiana. Funding for this project was provided to Paul M. Spengler by Grant MH56461 from the National Institute of Mental Health, Rockville, Maryland; by six grants to Paul M. Spengler from the Internal Grants Program for Faculty, Ball State University, Muncie, Indiana (two summer research grants, two summer graduate research assistant grants, an academic year research grant, and a new faculty research grant); and by three Lyell Bussell summer graduate research assistant grants, Ball State University, Muncie, Indiana. Preliminary findings from the clinical versus statistical prediction meta-analysis were presented at the annual meetings of the American Psychological Association in Washington, D.C., August 2000; Boston, August 1999; San Francisco, August 1998; Toronto, Ontario, Canada, August 1996; and New York City, August 1995, and by invited address at the annual meeting of the Society for Personality Assessment, New Orleans, March 1999. The authors extend special thanks to Kavita Ajmere, Corby Bubp, Jennifer Cleveland, Michelle Dorsey, Julie Eiser, Layla Hunton, Christine Look, Karsten Look, K. Christopher Rachal, Teresa Story, Marcus Whited, and Donald Winsted III for instrumental assistance in obtaining articles, coding studies, and managing data for the project. Correspondence concerning this article should be addressed to Stefanía Ægisdóttir, Department of Counseling Psychology, Teachers College 622, Ball State University, Muncie, IN 47306; e-mail: stefaegis@bsu.edu.

training, the statistical method does not. The statistical method requires only inserting data into a formula specifically designed for a particular judgment task. This may not be entirely true. Despite the use of formulas or tables to integrate information, the statistical method may require advanced training in the collection of relevant clinical and research-based information. Furthermore, advanced training may enhance a clinician's perceptions, which in turn may be quantified and used in a statistical model. For example, a clinician may believe a client has the potential for suicide, translate this impression into a number on a rating scale, and then statistically combine this number with other data to predict the client's risk for suicide (e.g., Westen & Weinberger, 2004).

To determine how counseling psychologists can be most effective in their decision making, knowing when and under what conditions each method is superior is important. The purpose of our meta-analysis is to articulate this knowledge.

THE CLINICAL VERSUS STATISTICAL PREDICTION CONTROVERSY

The search for the most accurate decision-making method is not new. In fact, this question has been debated for more than 60 years (Dawes et al., 1989; Meehl, 1954). The debate began with Meehl's (1954) book *Clinical Versus Statistical Prediction,* in which Meehl theoretically analyzed the relation between the clinical and statistical methods of prediction and summarized findings from existing literature. Meehl found that in all but 1 of 20 studies, statistical methods were more accurate than or equally accurate as the clinical method. He concluded that clinicians' time should be spent doing research and therapy, whereas work involving prognostic and classification judgments should be left to statistical methods.

Holt (1958), the most adamant defender of the clinical method, criticized Meehl's (1954) conclusions. Holt's critique involved essentially two issues: (a) the identification and assessment of predictive variables and (b) how they should be integrated. Holt believed that Meehl had given insufficient attention to the sophistication with which clinicians identify the criteria they are predicting, what variables to use in their prediction, and the strength of the relationship between predictors and criteria. In Holt's view, clinicians can identify these variables only through training and experience with comparable cases. After identifying the relevant variables, they are assessed. Assessment may be as much qualitative as quantitative. Holt's second criticism was that Meehl pitted "naïve clinical integration" of prediction against statistical decision making. A fairer comparison would compare statistical methods with "sophisticated clinical decision making and integration" (Holt,

1958) According to Holt, sophisticated clinical decision making is based on sophisticated data. These data are both qualitative and quantitative, have been gathered in a systematic manner, and have known relationships with what is being predicted. Unlike the statistical approach, the clinician remains the prime instrument, combining the data and making predictions that are tailored to each person. Holt presented data suggesting a superiority for sophisticated clinical rather than statistical procedures in predicting success in clinical training. On the basis of these findings, Holt argued for a combination of clinical and statistical methods (i.e., sophisticated clinical) that would be systematic and controlled and sensitive to individual cases.

Since this time, other narrative and box-score reviews of the literature on the differential accuracy of clinical and statistical methods have been published (e.g., Dawes et al., 1989; Garb, 1994; Grove & Meehl, 1996; Kleinmuntz, 1990; Russell, 1995; Sawyer, 1966; Wiggins, 1981). Narrative reviews are traditional literature reviews; box-score reviews count statistical significance and summarize studies in a table format. These reviews nearly always supported Meehl's (1954) conclusion that statistical methods were more accurate than or, at minimum, equally as accurate as clinical prediction methods (for a rare exception, see Russell, 1995). A recent meta-analysis of the clinical versus statistical literature (Grove et al., 2000) also supported earlier findings. Grove et al. (2000) found a consistent advantage ($d = .12$) for statistical prediction over clinical prediction across various types of nonmental health and mental health predictors and criteria.

Influence of the Statistical Versus Clinical Prediction Controversy

Despite the repeated conclusion that statistical prediction methods are more accurate than clinical procedures, the findings have had little influence on clinical practice (Dawes et al., 1989; Meehl, 1986). Dawes et al. (1989) and Meehl (1986) offered several reasons for this. They suggested that clinicians lack familiarity with the literature on clinical versus statistical prediction, are incredulous about the evidence, or believe that the comparisons were procedurally biased in favor of statistical prediction methods. They also proposed that certain aspects of education, training, theoretical orientation, and values might influence their reluctance to recognize advantages associated with statistical decision methods. Most clinicians highly value interpersonal sensitivity. Because of this, some may believe that the use of predictive formulas dehumanizes their clients. A corollary is that the use of group-based statistics or nomothetic rules is inappropriate for any particular individual. Practitioners are also subject to confirmatory biases such that they recall instances in which their predictions were correct but fail to recall those instances in which statistical prediction was more accurate. One might add another reason: Some accounts have simply been too broad to convince mental health practitioners. In some instances (e.g., Grove et al., 2000), the literature that has reviewed clinical versus statistical prediction includes research and criteria that range from mental health to medicine to finance.

Use of Statistical Prediction Models

Perhaps as a result of the limited influence of clinical versus statistical comparison studies, few statistical prediction models are available to counseling psychologists and psychotherapy practitioners (Meyer et al., 1998). Clinicians working in forensic settings, however, have developed such models. In fact, numerous funded research projects have been conducted to aid in classifying juvenile and adult prison inmates (e.g., Gottfredson & Snyder, 2005; Quinsey, Harris, Rice, & Cormier, 1998; Steadman et al., 2000; Sullivan, Cirincione, Nelson, & Wallis, 2001). One such effort is the Violence Risk Appraisal Guide (VRAG; Quinsey et al., 1998), which is a statistical system for predicting recidivism of imprisoned violent offenders.

The VRAG is based on more than 600 Canadian maximum security inmates who were released either back to the community, to a minimum security hospital, or to a halfway house. After a series of correlation analyses of predictor and outcome variables, a set of stepwise regression models was conducted. These analyses reduced the original 50 predictors to 12. These include psychopathy checklist scores (Hare, 1991), elementary school maladjustment scores, presence of a personality disorder, age at time of offense, separation from parents at an age younger than 16, failure on prior conditional release, nonviolent offense history score (using an instrument), marital status, schizophrenia diagnosis, most serious injury of offender's victim, alcohol abuse score, and gender of offender's victim. Each predictor was assigned a specified weight based on the empirical relationship with the outcome variable. Summing the resultant scores yields a probability estimate for an offender's future violence within the next 7 and 10 years. For instance, scores between +21 and +27 indicate a 76% likelihood for future violence, whereas scores between –21 and –15 suggest a probability of only 8%. The authors have validated this model for different groups of inmates (e.g., arsonists or sex offenders), with promising results (see Quinsey et al., 1998, for more detailed use of this statistical model).

In addition to forensics, statistical prediction formulas have been developed to aid with student selection for undergraduate, graduate, and professional schools. As an example, Swets et al. (2000) described a statistical prediction formula used in selecting candidates at the University of Virginia

School of Law. This formula consists of four predictor variables: undergraduate grade point average (GPA), mean GPA achieved by students from the applicants' college, scores from the Law School Admissions Test (LSAT), and the mean LSAT score achieved by all students from the applicants' college. Scores from these predictors are combined into a decision index of which a specific score indicates a threshold for admission. This statistical prediction formula predicts grades for 1st-year students and is used in combination with variables that are harder to quantify to select students (cf. Swets et al., 2000). Harvey-Cook and Taffler (2000) developed a statistical model using biographical data, frequently found on application forms and resumes, to predict success in accounting training in the United Kingdom. This six-variable model was developed on 419 accounting trainees. Retesting it on an independent sample of 243 trainees, Harvey-Cook and Taffler showed that their model could classify 88% of those failing and 33% of those successful in accounting training. The authors concluded that their model delivered better and more cost-effective results than clinical judgment methods currently used for this purpose in the United Kingdom (Harvey-Cook & Taffler, 2000).

Test cutoff scores offer another instance of a statistical procedure that may aid clinical decision making. Indeed, cutoff scores may be more readily available and easily constructed than statistical formulas. As an example, three Minnesota Multiphasic Personality Inventory–2 (MMPI-2) scales have been useful in classifying substance abuse: MacAndrew Alcoholism–Revised (MAC-R), Addiction Potential Scale (APS), and Addiction Acknowledgment Scale (AAS) (Rouse, Butcher, & Miller, 1999; Stein, Graham, Ben-Porath, & McNulty, 1999). Relying on data from 500 women and 333 men seeking outpatient mental health services, Stein et al. (1999) found that cutoff scores on the MAC-R correctly classified 86% of the women and 82% of the men as either substance abusers or nonabusers. In the case of the AAS, cutoff scores could predict 92% of women and 81% of men as either substance abusers or nonabusers. Likewise, cutoff scores with the APS enabled accurate prediction of 84% of women and 79% of men as either abusing or not abusing substances. This method of classification greatly exceeds the base rates for chance classification. For women, the positive predictive power (ability to detect substance abusers) for MAC-R, AAS, and APS was 100%, 79%, and 53%, respectively. These values compare with a base rate of 16%. For men, the respective positive predictive power for MAC-R, AAS, and APS was 100%, 68%, and 77%, respectively, which compare with a base rate of 27%.

Purpose of This Meta-Analysis

The current meta-analysis seeks to address several omissions in the literature on clinical versus statistical prediction. Although Grove et al.'s (2000) important study confirmed prior conclusions about the relative merits of clinical and statistical prediction methods, questions still remain regarding the application of the findings to judgment tasks commonly encountered by mental health practitioners. First, their review combined literature from psychology, medicine, forensics, and finance. Consequently, conclusive results are not provided about prediction accuracy for mental health clinical and counseling practitioners relative to statistical methods. Second, even though Grove et al. examined the influence of various study design characteristics (e.g., type of criterion, professional background of clinical judges, judge's level of experience, and amount of data available to the judges versus the statistical formulas), the influence of these design characteristics on the accuracy of prediction was not investigated when the criteria were psychologically related. Instead, Grove et al. investigated the influence of these study design variables on the overall effect, including studies from the diverse professional fields listed earlier. Similarly, despite Grove et al.'s examination of the influence of criterion type on the overall effect of the difference between clinical and statistical prediction accuracy, their criteria breakdown was broad (i.e., educational, financial, forensic, medical, clinical-personality, and other). The breakdown offers little specific information on which counseling psychologists can rely to decide when and under what conditions they should use clinical or statistical methods.

The first aim of this meta-analysis was to synthesize studies that had examined the differential accuracy of clinical and statistical judgments in which the prediction outcome was relevant to counseling psychology. Second, we examined studies in which predictions by mental health professionals were compared with statistical methods. In a typical study comparing these two methods, clinicians first synthesized client data (e.g., interview data, psychological tests, or a combination of interview information and one or more psychological tests) and then made a classification judgment (e.g., diagnosis) or predicted some future outcome (e.g., prognosis). The accuracy of these judgments was compared with a statistical prediction scheme in which the same (sometimes less or more) information was entered into a statistical formula that had been previously designed on the basis of empirical relations between the predictors (specific client data) and the criterion (the prediction task of interest). Third, we examined questions generated from the years of debate about the relative merits of clinical and statistical prediction. More specifically, we examined how the differential accuracy between clinical and statistical methods was affected by (a) type of prediction, (b) setting from which the data were gathered, (c) type of statistical formula, (d) amount of information provided to the clinician and formula, (e) information provided to the clinician about base rates, (f) clinician access to the statistical formula, (g) clinician expertness, (h) our evaluation of the validity of the criteria for accurate judgment, (i) publication source, (j) number of clinicians performing predictions in a study, (k) number of criterion behaviors predicted in a study, and (l) publication year.

Meta-analyses provide detailed and comprehensive syntheses of the professional literature. As such, they are especially relevant for bridging the gap between the science of counseling psychology and how it is practiced by counseling psychologists (e.g., Chawalisz, 2003; Stricker, 2003; Wampold, 2003). The current meta-analysis addresses how counseling psychologists should best make decisions: when they should use clinical methods, when they would do well to use statistical methods, and when either is acceptable. In addition to relying on empirically supported treatment strategies, the counseling psychologist scientist-practitioner may be informed by the current meta-analysis about situations when statistical decision methods lead to more accurate clinical predictions than the clinical method.

Spengler et al. (1995), for instance, proposed an elaborated model of the scientist-practitioner, basing their clinical judgment model on Pepinsky and Pepinsky (1954). In this model, strategies were proposed to increase judgment accuracy relying on scientific reasoning. They suggested that to improve judgment accuracy, counseling psychologists (a) should be aware of their values, preferences, and expectations; (b) should use multiple methods of hypothesis testing (both confirming and disconfirming); and (c) should use judgment debiasing techniques (cf. Spengler et al., 1995). We argue that the current meta-analysis will further inform counseling psychologists as scientists not by providing information about the absolute accuracy of clinical judgment (i.e., when it may be most vulnerable to error) but instead by assessing the relative accuracy of clinical versus statistical prediction. Under conditions in which statistical prediction is superior, a successful debiasing method would use prediction methods based on empirical relations between variables (i.e., statistical methods). On the basis of this meta-analysis, we hope to also suggest options for future research and training relevant to decisions typically made by counseling psychologists.

METHOD

Study Selection

This study is part of a large-scale meta-analysis of the clinical judgment (MACJ) literature (Spengler et al., 2005). By using 207 search terms, the MACJ project identified 1,135 published and unpublished studies between

1970 and 1996 that met our criteria for inclusion in meta-analyses of mental health clinical judgment.[1] However, because of the extensive historical debate about the relative benefits of statistical versus clinical prediction, we extended our search strategy for the present study back to 1940, thus defining the current study's search period from 1940 to 1996. After an iterative process, we identified 156 studies that investigated some form of statistical prediction or model of clinical prediction for a mental health criterion compared with the accuracy of clinical judgment.

To be included in the meta-analysis, studies had to meet the following criteria: (a) a direct comparison was reported between predictions made by mental health practitioners (i.e., professionals or graduate students) and some statistical formula, (b) a psychological or a mental health prediction was made (e.g., diagnosis, prognosis, or psychological adjustment), (c) the clinicians and the statistical formula had access to the same predictor variables or cues (even though the amount of information might vary), (d) the clinicians and the formula had to make the same predictions, and (e) the studies had to contain data sufficient to calculate effect sizes. By using these selection criteria, 67 studies qualified for inclusion, yielding 92 effect sizes. When Goldberg (1965) and Oskamp (1962) were included, 69 studies produced 173 effect sizes (see below).

Specialized Coding Procedures

The MACJ project used a coding form with 122 categories or characteristics (see Spengler et al., 2005) that were grouped under the following conceptual categories: judgment task, judgment outcomes, stimulus material, clinician individual differences, standard for accuracy, method of study, and type of design. An additional coding form was constructed including study design characteristics identified in historical literature and more contemporary research as potentially affecting the differential accuracy of clinical and statistical prediction. These design characteristics became the independent variables. We also noted whether the statistical formulas were cross-validated. In this instance, cross-validated formulas refer to any statistical formulas that have been independently validated on a different sample from which the formula was originally derived. For example, if a score of 10 on an instrument developed to diagnose major depressive disorder correctly identifies 95% of persons with that disorder, to be considered a cross-validated formula (i.e., a score of 10 indicates major depression), that same score (10) had to be able to identify major depressive disorder with comparable accuracy using another sample of persons with the disorder. Coding disagreements were resolved by discussion among coders until agreement was reached.

Dependent Measure: Judgment Accuracy

The dependent variable for all analyses was judgment accuracy. For a study to be included, a criterion had to be established as the accurate judgment (e.g., prior diagnosis or arrest records). For instance, Goldberg (1970) compared clinical and statistical judgments of psychotic versus neurotic MMPI profiles to actual psychiatric diagnosis. MMPI profiles from psychiatric patients diagnosed as clearly psychotic or neurotic were presented to clinical psychologists. Their judgment about whether the MMPI profiles belonged to either a psychotic or a neurotic patient was compared with a statistical formula constructed to categorize patients as psychotic if five MMPI scales (the lie, 6 [Pa], 8 [Sc], 3 [Hy], 7 [Pt]) were elevated. These two types of judgments were compared with the prior diagnoses, which were considered the accurate judgment. In another example, Gardner, Lidz, Mulvay, and Shaw (1996) examined clinical and statistical prediction of future violence. Gardner et al. developed three statistical formulas to predict future violence on the basis of clinical (e.g., diagnosis and drug use) and demographic information as well as information about prior violence. Violence prediction based on these three models was compared with predictions made by clinicians who had access to the same information as the formulas. The accuracy of these judgments was then compared with records of violent behavior (psychiatric, arrest, or commitment records) or from patients' reports about their violent behavior. In this study, available records and patient self-reports about violent behavior served as the criteria for accurate judgment. Thus, specific criteria for accurate judgments had to be reported for a study to be included in this meta-analysis.

Effect Size Measure

As Cohen (1988) noted in his widely read book, effect sizes may be likened to the size of real differences between two groups. Estimates of effect size are thus estimates of population differences—they estimate what is really happening and are not distorted by sample size. The purpose of a meta-analysis is to estimate the effect size in a population of studies. In our case, a mean weighted effect size (d^+) was used to represent the difference between clinical and statistical prediction accuracy.[2] Effect size measured by d^+ represents the mean difference between two samples of studies expressed in standard deviation units (g) and corrected for sample size (Johnson, 1993). More specifically, the mean judgment accuracy of statistical prediction was subtracted from the mean judgment accuracy of clinical prediction divided by the pooled standard deviation and then corrected for sample size.

In this study, the effect size (d^+) represents the magnitude, not the statistical significance, of the relative difference between clinical and statistical prediction accuracy. A negative d^+ value indicates superiority of the statistical prediction method, whereas a positive d^+ indicates superiority of the clinical method. An effect of zero indicates exactly no difference between the two methods. In addition to d^+, we reported the 95% confidence interval for the effect size. Confidence interval provides the same information as that extracted from significant tests. It permits one to say with 95% confidence (i.e., $\alpha = .05$) that the true effect size falls within its boundaries. If the confidence interval includes zero, the population effect may be zero; one cannot say with confidence that a meaningful difference exists between the two groups. However, if the confidence interval does not include zero, one can conclude that a reliable difference exists between clinical and statistical prediction (e.g., Johnson, 1993).

The data were reduced to one representative effect size per study in most cases. This prevented bias that would result if a single study was overrepresented in the sample (Cooper, 1998; Rosenthal, 1991). For instance, if a study reported more than one statistical or clinical prediction (e.g., brain impairment and lateralization of the impairment; Adams, 1974), an average of the reported judgment accuracy statistic was calculated and transformed into one effect size. Also, if a study reported results from both non–cross-validated and cross-validated statistical prediction schemes, only results from the cross-validated statistical formula were used. This was done to prevent bias in favor of the statistical method, given the possibility of inflated correlations (based on spurious relations) between predictor and criterion variables in non–cross-validated statistical formulas (for more discussion of these issues, see Efron & Gong, 1983). Table 1 notes whether the studies used cross- or non–cross-validated statistical formulas.

Even though one average effect size per study was usually calculated, 18 studies produced more than one effect size (see Table 1). These studies included more than one design characteristic (independent variables) that we hypothesized might influence clinical versus statistical prediction accuracy and reported accuracy statistics for various levels of the independent variable. An example would be a study investigating clinical versus statistical prediction under two conditions. In one condition, the clinicians have access to the statistical prediction scheme, whereas in another condition they do not. In our studies, we extracted two effect sizes. That is, the study's two conditions (with and without access to the statistical formula) were treated as two independent projects. Furthermore, a study was allowed to produce

TABLE 1: Studies Included in Meta-Analysis

		Accuracy Statistic	*Accuracy*		
Citation	*Prediction*	*Reported*	*Clinical*	*Statistical*	d$^+$
Adams (1974)[1]	Brain impairment	Hit rate	53	52	.02
Adams (1974)[2a]	Brain impairment	Hit rate	53	56	–.06
Alexakos (1966)[1]	Academic performance	Hit rate	39	51	–.24
Alexakos (1966)[2]	Academic performance	Hit rate	39	52	–.27
Astrup (1975)[b]	Psychiatric diagnosis	Hit rate	78	74	.09
Barron (1953)	Psychotherapy outcome	Hit rate	62	73	–.23
Blumetti (1972)	Length of psychotherapy	Hit rate	61	54	.15
Bolton et al. (1968)	Prognosis	Correlation	.35	.48	–.16
Carlin and Hewitt (1990)	Real vs. random MMPI profile	Hit rate	63	95	–.73
Conrad and Satter (1954)	Academic performance	Correlation	.36	.46	–.12
Cooke (1967a)	Psychiatric diagnosis	Hit rate	77	76	.02
Cooke (1967b)[a]	Psychiatric diagnosis	Correlation	.42	.51	–.11
Danet (1965)	Prognosis	Hit rate	64	70	–.13
Devries and Shneidman (1967)[a]	Matching MMPI profiles to persons	Hit rate	75	100	–.81
Dickerson (1958)	Compliance with counseling plan	Hit rate	57	52	.10
Dunham and Meltzer (1946)[b]	Length of hospital stay	Hit rate	38	58	–.40
Evenson, Altman, Sletten, and Cho (1975)	Length of hospital stay	Hit rate	64	71	–.14
Fero (1975)[1]	Prognosis	Correlation	.35	.57	–.29
Fero (1975)[2b]	Prognosis	Correlation	.35	.73	–.57
Gardner et al. (1996)[1b]	Offense / violence	Hit rate	62	74	–.25
Gardner et al. (1996)[2]	Offense or violence	Hit rate	62	71	–.20
Gardner et al. (1996)[3]	Offense or violence	Hit rate	62	70	–.17
Gardner et al. (1996)[3]	Offense or violence	Hit rate	62	70	–.17

Gaudette (1992)[a]	Brain impairment	Correlation	.39	.45	–.07
Goldberg (1965) × 65[d]	Psychiatric diagnosis	Hit rate	62	64	–.04
Goldberg (1970)[1a]	Psychiatric diagnosis	Correlation	.28	.31	–.03
Goldberg (1970)[2]	Psychiatric diagnosis	Correlation	.28	.44	–.18
Goldberg (1970)[3a]	Psychiatric diagnosis	Correlation	.28	.46	–.21
Goldstein, Deysach, and Kleinnecht (1973)	Brain impairment	Hit rate	95	75	.57
Grebstein (1963)[a]	IQ	Correlation	.62	.56	.08
Gustafson, Greist, Stauss, Erdman, and Laughren (1977)[a]	Suicide attempt	Hit rate	65	80	–.33
Halbower (1955)	Prognosis	Correlation	.60	.79	–.21
Hall (1988)[a]	Offense or violence	Hit rate	55	81	–.59
Heaton et al. (1981)[b]	Brain damage	Hit rate	79	72	.14
Holland, Holt, Levi, and Beckett (1983)	Offense or violence	Correlation	.21	.28	–.08
Holt (1958)	Academic or training performance	Correlation	.25	.13	.12
Hovey and Stauffacher (1953)	Personality characteristics	Hit rate	74	63	.23
Johnston and McNeal (1967)	Length of hospital stay	Hit rate	72	72	.00
Kaplan (1962)[1]	Prognosis	Hit rate	58	70	–.25
Kaplan (1962)[2]	Prognosis	Hit rate	66	70	–.09
Kelly and Fiske (1950)	Academic or training performance	Correlation	.28	.29	–.01
Klehr (1949)	Psychiatric diagnosis	Hit rate	57	64	–.14
Kleinmuntz (1967)	Adjustment	Hit rate	66	71	–.10
Klinger and Roth (1965)[b]	Psychiatric diagnosis	Hit rate	94	71	.63
Lefkowitz (1973)	Marital adjustment	Hit rate	54	55	–.03
Leli and Filskov (1981)[1]	Brain impairment	Hit rate	41	62	–.44
Leli and Filskov (1981)[2]	Brain impairment	Hit rate	50	62	–.24
Leli and Filskov (1984)[1a]	Brain impairment	Hit rate	60	83	–.52
Leli and Filskov (1984)[2a]	Brain impairment	Hit rate	71	83	–.30
Lemerond (1977)	Suicide attempt	Hit rate	53	50	.06

(continued)

TABLE 1 (continued)

		Accuracy Statistic	Accuracy		
Citation	*Prediction*	*Reported*	*Clinical*	*Statistical*	d$^+$
Lewis and MacKinney (1961)	Career satisfaction	Correlation	.12	.21	–.09
Lindsey (1965)	Homosexuality	Hit rate	70	57	.03
Lyle and Quast (1976)	Brain impairment	Hit rate	67	68	–.02
McHugh and Apostolakos (1959)[b]	Academic field	Hit rate	70	46	.50
Meehl (1959)	Psychiatric diagnosis	Hit rate	69	74	–.11
Melton (1952)[1b]	Academic performance	Error rate	.47	.40	–.26
Melton (1952)[2]	Academic performance	Error rate	.45	.35	–.35
Meyer (1973)[1]	Psychiatric diagnosis	Hit rate	65	63	.05
Meyer (1973)[2]	Psychiatric diagnosis	Hit rate	67	63	.08
Miller, Kunce, and Getsinger (1972)[a]	Adjustment / employability	Correlation	.65	.28	.49
Moxley (1970)[1a]	Length of psychotherapy stay	Hit rate	55	67	–.25
Moxley (1970)[2a]	Length of psychotherapy stay	Hit rate	61	67	–.13
Moxley (1970)[3a]	Length of psychotherapy stay	Hit rate	67	67	.00
Moxley (1970)[4a]	Length of psychotherapy stay	Hit rate	69	67	.04
Oskamp (1962) × 16[d]	Prior hospitalization	Hit rate	72	65	.13
Oxman, Rosenberg, Schnurr, and Tucker (1988)[1a]	Psychiatric diagnosis	Hit rate	66	80	–.32
Oxman et al. (1988)[2a]	Psychiatric diagnosis	Hit rate	66	62	.08
Perez (1976)[1a]	Homicidality	Hit rate	50	83	–.77
Perez (1976)[2a]	Homicidality	Hit rate	51	83	–.77
Perez (1976)[3a]	Homicidality	Hit rate	51	83	–.77
Perez (1976)[4a]	Homicidality	Hit rate	58	83	–.61
Popovics (1983)	IQ	Correlation	.20	.19	.01

Sarbin (1942)	Academic performance	Correlation	.56	.60	–.06
Shaffer, Perlin, Schmidt, and Stephens (1974)[1a]	Suicide attempt	Correlation	.40	.40	.00
Shaffer et al. (1974)[2]	Suicide attempt	Correlation	.40	.18	.24
Shaffer et al. (1974)[3b]	Suicide attempt	Correlation	.40	.12	.30
Shagoury and Satz (1969)[a]	Brain impairment	Hit rate	83	87	–.11
Stricker (1967)	Psychiatric diagnosis	Hit rate	70	79	–.21
Szuko and Kleinmuntz (1981)	Lie detection	Correlation	.23	.52	–.32
Taulbee and Sisson (1957)[1]	Psychiatric diagnosis	Hit rate	64	78	–.31
Taulbee and Sisson (1957)[2]	Psychiatric diagnosis	Hit rate	64	63	.02
Thompson (1952)[b]	Juvenile delinquency	Hit rate	65	90	–.60
Walters et al. (1991)[a]	Malingering	Hit rate	59	73	–.30
Watley (1966)	Academic performance	Hit rate	68	75	–.14
Watley and Vance (1964)[a]	Academic performance	Hit rate	70	80	–.25
Webb et al. (1977)[b]	Occupational choice	Hit rate	35	55	–.41
Wedding (1983)[1]	Brain impairment	Hit rate	55	63	–.17
Wedding (1983)[2]	Brain impairment	Hit rate	55	60	–.11
Weinberg (1957)[1a]	Personality characteristics	Correlation	.49	.54	–.15
Weinberg (1957)[2a]	Personality characteristics	Correlation	.40	.54	–.06
Werner, Rose, Yesavage and Seeman (1984)[a]	Offense / violence	Correlation	.14	.40	–.27
Wiggins and Kohen (1971)[1a]	Academic performance	Correlation	.33	.69	–.51
Wiggins and Kohen (1971)[2a]	Academic performance	Correlation	.33	.50	–.20
Wirt (1956)[a]	Prognosis	Hit rate	54	79	–.54
Wittman and Steinberg (1944)[b]	Prognosis	Hit rate	43	78	–.77

NOTE: Hit rate refers to percentage correct; correlation refers to correlation with the criteria; error rate difference was calculated by a t test that was then transformed into d^+. All accuracy statistics are reported raw (untransformed). d^+: Effect size of difference between clinical and statistical prediction accuracy; negative d^+ indicates superiority of statistical prediction. Studies producing more than one effect size have each effect identified by superscripts 1 to 4. MMPI = Minnesota Multiphasic Personality Inventory.

a. Studies that did not use cross-validated statistical formulas.

b. Studies that were diagnosed as outliers. These studies were excluded in some analyses.

c. Studies that were only included in the overall effect. The average effect is reported.

d. Average hit rate for Goldberg (1965) and Oskamp (1962) studies.

more than one effect size if it used more than one statistical prediction scheme (e.g., regression formula and test cutoff score).

Despite the potential for bias, the questions addressed by these analyses were essential to understanding what factors influence clinical versus statistical judgment accuracy. Their value thus outweighed the risk of adding bias to the overall results. Both Cooper (1998) and Johnson and Eagly (2000) suggest that this is acceptable under conditions similar to ours.

Nonetheless, two studies (Goldberg, 1965; Oskamp, 1962) produced so many effect sizes that they were treated as separate cases. Oskamp (1962) reported hit rates of 16 different cross-validated statistical formulas compared with the judgment accuracy of the average clinician. Goldberg (1965) reported the prediction accuracy of 65 different statistical methods compared with the average clinician. To prevent an overrepresentation of data from Goldberg and Oskamp, the overall effect size was reported with and without these two studies. They were not included in our analysis of study design characteristics.

Strategy for Data Analyses

The overall effect size was first calculated for 69 studies, including Goldberg (1965) and Oskamp (1962), producing 173 effect sizes. Second, as noted under "Study Selection," the overall effect size was also calculated without the Goldberg and Oskamp studies, producing 92 effect sizes. Third, the overall effect size was calculated using only the 49 studies that compared cross-validated statistical formulas with the clinical method. This last procedure resulted in 60 effect sizes and eliminated artificially inflated accuracy of statistical methods caused by chance associations among variables (e.g., Dawes et al., 1989; Meehl, 1954). Next, the influence of study design characteristics (independent variables) on the overall effect size of these 49 studies was calculated[3] (see Results for description and hypothesized effects). Because heterogeneity was not eliminated when study design characteristics were considered, an outlier analysis was performed to reduce heterogeneity[4] (Johnson & Eagly, 2000). The 41 studies (48 effect sizes) remaining after this analysis represented the most conservative sample of studies and were used to calculate the influence of study design characteristics on the overall effect.

It was necessary to remove 12 outliers (20%) from the 60 effects produced by the use of cross-validated statistical formulas to reach a homogenous set of effects, Q (47) = 65.83, $p > .05$. The resulting set of 41 studies produced 48 effects and shared a common effect size. They are sufficiently homogenous to be characterized as coming from the same population. Outliers, by contrast, have extreme effects. Their extremity may be because of error or may be because the studies have characteristics that are deviant relative to most studies in the sample (Hunter & Smith, 1990). According to Hedges (1987) and others, removing up to 20% of studies to reach a homogeneous set of data is not uncommon for meta-analyses of psychological topics. This conservative sample of studies, those that used cross-validated statistical formulas and did not have unrepresentatively extreme effect sizes, was used to interpret differences between clinical and statistical prediction accuracy.

RESULTS

Studies investigated the clinical and statistical judgment accuracy of the following 11 prediction tasks: brain impairment, personality, length of hospital stay or treatment, diagnosis, adjustment or prognosis, violence or offense, IQ, academic performance, if an MMPI profile was real or fictional, suicide attempt, and homosexuality. In all studies, the prediction accuracy of mental health professionals and graduate students in the mental health field were compared with a statistical prediction method. The studies, prediction tasks, and accuracy statistics are listed in Table 1.

Overall Effect Size

Figure 1 presents a distribution of the difference between clinical and statistical prediction accuracy transformed into a weighted effect size (d^+) in stem-and-leaf diagram format for the most conservatively selected set of effect sizes ($n = 48$). As may be seen, effect sizes ranged from .57 in favor of the clinical method to –.73 in favor of the statistical method. Visual inspection reveals a slight skew in favor of the statistical method. Similar results were obtained when outlier effects were included, without Goldberg (1965) and Oskamp (1962). For these 92 effect sizes, effects ranged from .63 to –.81 (see Note 3).

On the basis of suggestions from the literature (Dawes et al., 1989; Garb, 1994; Grove et al., 2000; Grove & Meehl, 1996; Kleinmuntz, 1990; Meehl, 1954; Russell, 1995; Sawyer, 1966; Wiggins, 1981), we expected that statistical prediction would be, in general, more accurate than clinical prediction. Our hypothesis was confirmed. By using the Grove et al. (2000) approach to interpret the relative difference between clinical and statistical prediction accuracy, Figure 1 reveals that of the 48 effect sizes, 25 effects (52%) favored statistical prediction methods, 18 (38%) reported no difference between the two methods, and 5 (10%) favored the clinical method.

Overall effect sizes (both in the form of d^+ and r^+) for the four sets of analyses performed are presented in Table 2. Effect sizes (d^+) ranged from

0	−.8	
0	−.7	
1	−.7	3
0	−.6	
0	−.6	
0	−.5	
0	−.5	
0	−.4	
1	−.4	4
1	−.3	
2	−.3	21
5	−.2	97655
5	−.2	43110
3	−.1	775
8	−.1	44432110
3	−.0	997
4	−.0	3321
6	.0	012223
4	.0	5688
1	.1	2
1	.1	5
2	.2	34
0	.2	
0	.3	
0	.3	
0	.4	
0	.4	
0	.5	
1	.5	7
0	.6	

FIGURE 1. Stem-and-Leaf Diagram for Clinical and Statistical Prediction Differences of Transformed Effect Sizes (d^+) for the Most Conservative Sample of Studies ($N = 48$ Effects)

NOTE: A negative d^+ indicates superiority of statistical method.

–.16, without Goldberg (1965) and Oskamp (1962), to –.12 for studies using cross-validated formulas with outlier studies excluded. As Table 2 also shows, the test for homogeneity of effects, Q, was significant ($p < .001$) in all instances except when outlier effects were removed from the analysis. This indicates that when outlier effects were included, the variability among effect sizes was too great for a reliable interpretation. Therefore, significant heterogeneous effect sizes (reflected by Q values) must be interpreted with caution. By contrast, confidence can be placed in the overall effect size of $d^+ = –.12$ ($r^+ = –.06$) produced by nonoutlying studies with cross-validated formulas. In this case, the effect size is reliable as indicated by a nonsignificant Q. Additionally, the 95% confidence interval did not cross zero, further supporting the reliability of the effect.

TABLE 2: **Overall Effect Sizes for Clinical Versus Statistical Prediction**

	Mean Weighted Effect Size (d^+)	*95% Confidence Interval*	r^+	*Homogeneity Index* Q
With Goldberg (1965) and Oskamp (1962) (*N* = 69 studies and 173 effects)	–.16	–.18 to –.14	–.08	460.54***
Without Goldberg (1965) and Oskamp (1962) (*N* = 67 studies and 92 effects)	–.15	–.17 to –.13	–.08	470.71***
Cross-validated studies (*N* = 49 studies and 60 effects)	–.14	–.17 to –.12	–.07	371.82***
Cross-validated studies without outliers (*N* = 41 studies and 48 effects)	–.12	–.14 to –.09	–.06	65.84

NOTE: Negative d^+ indicates superiority of statistical method.
***$p < .001$.

Because of these two conditions, we can conclude that statistical prediction methods are, in general, more accurate than clinical prediction methods. This is a small effect by conventional standards (Cohen, 1988), but it is consistent and reliable. When consistency and reliability are established in a meta-analysis, confidence can be placed in the effect as a true effect (i.e., statistical is better than clinical prediction). Another perspective on the meaning of this effect is possible through the use of a binomial effect size display analysis (Rosenthal & Rubin, 1982). This analysis permits the level of improvement achieved with one decision-making method versus another to be determined. In this case, clinical decisions are accurate 47% of the time, $(.50 - r/2) \times 100$, whereas statistical decisions are accurate 53% of the time, $(.50 + r/2) \times 100$. As a result, the likelihood of a successful decision can go up 13% when statistical rather than clinical methods are used.

Study Design Characteristics

The overall effect size of the difference between clinical and statistical methods was analyzed as a function of study design characteristics (i.e., independent variables hypothesized to influence the overall effect size). We used only the most conservative sample of studies. The study design characteristics were fitted to the effect sizes using a procedure analogous to the analysis of variance (Hedges & Olkin, 1985). The between-class effect (Q_B) in this procedure is similar to a main effect in an analysis of variance. The 95% confidence interval and a test of the homogeneity of the effect size within each class (Q_W) were also calculated. The homogeneity of studies within a class (Q_W) determines the confidence that can be placed in interpreting the effect size (d^+). If studies are heterogeneous within a class (e.g., prediction task), nonrandom variance in d^+ remains in that class (Johnson, 1993).

Results of these analyses are reported in Table 3. Each study design characteristic or independent variable had several levels. For example, studies varied in how much information (predictor cues) they provided to clinicians and the formula on which clinicians rely in making predictions. Some gave the same amount to the clinician and the statistical formula, others gave the clinicians more, others gave the clinician less, and some did not report how much information was provided to either. Before our analysis, we determined that to have sufficient reliability and representativeness, three or more studies would have to examine a design characteristic level to be included in the analysis. Because only two studies gave the clinician less information than the formula, that condition is not included in the overall test of the design characteristic. Within-class effects (Q_W) values are included in Table 3 for information only. Seven studies did not report how much information was given to clinicians and the formula. These seven were also excluded from our analyses.

Type of prediction. Since Meehl's (1954) original publication, the methodology and the nature of the studies comparing clinical and statistical prediction accuracy have been questioned. One particular criticism concerned the unrepresentativeness of the judgment tasks evaluated. It was argued that unusual predictions (e.g., predicting future grade point average; Holt, 1970) did not provide a fair evaluation of the accuracy of clinicians' judgment. In response to this criticism, subsequent studies broadened the scope of the prediction tasks (cf. Rock, Bransford, & Maisto, 1987). We identified 11 prediction tasks and examined six of them (Table 3).

Effect sizes were expected to vary across different prediction tasks. This expectation was confirmed, Q_B (5) = 12.42, $p < .05$. Three of the six prediction tasks that met our criteria for analysis had an effect size in which the confidence interval did not include zero. These were predictions of adjustment or prognosis ($d^+ = -.14$), offense or violence ($d^+ = -.17$), and academic performance ($d^+ = -.14$). For the other three prediction tasks, the confidence interval included zero, which indicates that any differences between the clinical and statistical methods were unreliable. Data were insufficient to analyze differences in determining personality type, estimating client IQ, determining whether an MMPI profile was fictitious, predicting suicide attempt, or determining homosexuality. In no instance was the clinical method consistently more accurate than the statistical method.

Data collection setting. Holt (1958, 1970) stated that in many studies, clinicians were asked to make unusual predictions with limited data. These

TABLE 3: Effect Size by Study Design Characteristics, Cross-Validated Studies Without Outliers (*N* = 41 Studies, 48 Effects)

	Between-Class Effect (Q_B)	*n*	*Mean Weighted Effect Size* (d^+)	*95% CI Lower*	*95% CI Higher*	r^+	*Homogeneity Within Class* (Q_W)
Type of prediction	12.42*						
Brain impairment		7	–.05	–.22	.12	–.03	6.33
Hospital or treatment length		3	–.02	–.13	.09	–.01	2.33
Diagnosis		9	–.05	–.12	.01	–.03	7.39
Adjustment or prognosis		11	–.14	–.19	–.10	–.07	6.69
Future offense or violence		4	–.17	–.24	–.11	–.09	1.46
Academic performance		8	–.14	–.18	–.09	–.07	11.30
IQ		(1)	(.01)	(–.36)	(.00)	(.00)	(0.00)
Personality type		(1)	(.23)	(–.18)	(.64)	(.11)	(0.00)
Random MMPI profile		(1)	(–.73)	(–1.24)	(–.22)	(–.34)	(0.00)
Suicide attempt prediction		(2)	(.16)	(–.06)	(.38)	(.07)	(.63)
Homosexuality		(1)	(.27)	(–.24)	(.78)	(.13)	(0.00)
Data collection setting	6.96**						
Clinicians from same setting as data		21	–.14	–.17	–.11	–.07	28.03
Clinicians not from same setting as data		25	–.06	–.11	–.01	–.03	30.37
Not reported		(2)	(–.06)	(–.31)	(.18)	(–.03)	(.29)
Statistical formula type	11.94**						
Linear statistical formula		27	–.15	–.19	–.12	–.08	32.29
Test cutoff score		8	–.10	–.15	–.06	–.05	8.06
Logically constructed rule		11	–.03	–.09	.03	–.02	11.87
Model of clinical judgment		(2)	(–.14)	(–.23)	(–.05)	(–.07)	(2.47)
Amount of information	4.06*						
Same amount for clinicians and formula		24	–.06	–.11	–.01	–.03	28.26
Clinicians have more than formula		15	–.13	–.16	–.09	–.06	25.77*
Clinicians have less than formula		(2)	(–.12)	(–.42)	(.19)	(–.06)	(1.64)
Not reported		(7)	(–.15)	(–.20)	(–.10)	(–.08)	(3.81)

(continued)

TABLE 3 (continued)

	Between-Class Effect (Q$_B$)	*n*	*Mean Weighted Effect Size* (d$^+$)	*95% CI Lower*	*95% CI Higher*	r$^+$	*Homogeneity Within Class* (Q$_W$)
Information about base rates	4.25						
Base rate provided		8	–.02	–.12	.08	–.01	5.40
Base rate same as natural setting		8	–.11	–.15	–.06	–.05	6.37
Clinicians do not know base rate		27	–.13	–.16	–.09	–.06	45.56*
Not reported		(5)	(–.14)	(–.19)	(–.09)	(–.07)	(3.04)
Availability of statistical formula	2.70						
Available		5	–.14	–.18	–.09	–.07	7.33
Not available		40	–.09	–.12	–.06	–.05	53.34
Not reported		(3)	(–.15)	(–.21)	(–.10)	(–.08)	(.05)
Clinician expertness	2.47						
Experts in the prediction task		7	–.05	–.14	.03	–.03	14.96*
Nonexperts in the prediction task		41	–.12	–.15	–.10	–.06	48.41
Publication source	2.06						
American Psychological Association journal		19	–.11	–.15	–.07	–.05	22.53
Non–American Psychological Association journal		16	–.11	–.15	–.06	–.05	21.85
Dissertation		13	–.15	–.19	–.10	–.07	19.39
Confidence in criterion for accuracy	.12						
Low		12	–.12	–.17	–.08	–.06	12.03
High		36	–.12	–.14	–.08	–.06	53.69

NOTE: A negative *d*$^+$ indicates superiority of the statistical method. Values in parentheses were not included in the overall analysis. Either too few studies examined the design characteristic level, or the study did not report how the characteristic was treated. Values are included for information only. CI = confidence interval. MMPI = Minnesota Multiphasic Personality Inventory.

**p* < .05. ***p* < .01. ****p* < .001.

conditions, in Holt's opinion, did not adequately represent the clinical approach and invalidated the comparison between clinical and statistical methods. Yet if the data given to clinicians are from the same setting that the clinicians inhabit, then clinicians should be familiar with the predictors. They should know how to use them to form effective judgments. Based on this rationale, we expected that if the clinicians were drawn from the same setting as were the data (e.g., clinicians who worked in the clinic from which the client data were drawn), the difference between clinical and statistical prediction would be less than if the clinicians and the data were not from the same setting.

We found that data collection setting indeed influenced effect size, Q_B (1) = 6.96, $p < .01$, although not as we expected. Contrary to our expectations, when clinicians were not from the same setting as the data on which they based their predictions ($d^+ = -.06$), the effect size was smaller than if the data were from their work setting ($d^+ = -.14$). In neither instance did the effect-size confidence interval include zero. Therefore, even though statistical methods yield more accurate predictions independent of the familiarity of clinicians with the data, the difference between the clinical and the statistical method is smaller when clinicians do not come from the same setting as the data used to make their judgments. Thus—and most unexpected—existing studies indicate that clinicians seem to be more accurate when they are working with less familiar or novel information.

Statistical formula type. Studies have compared several types of statistical formula with the clinical method. We organized the studies on four types of formula: linear statistical models (e.g., regression or discriminant function), logically constructed rules (e.g., Goldberg rule for differential diagnosis of psychosis and neurosis from MMPI), test cutoff scores, and models of clinical judgment (mechanization of clinicians' judgment processes). Even though no specific type of statistical formula was expected to yield the most accurate results, on the basis of Dawes and Corrigan's (1974) hypotheses we assumed that simple linear models would be as accurate as more complex models. We also expected, on the basis of narrative and empirical evidence (e.g., Dawes et al., 1989; Grove et al., 2000), that any of the four types of statistical method would be more accurate than the clinical method.

Our hypotheses were partially supported. The type of statistical formula used affected overall effect size, Q_B (3) = 11.94, $p < .01$. Effect sizes (d^+) ranged from –.15 for linear statistical formulas (i.e., regression and discriminant function analysis) to –.03 for logically constructed rules or signs (e.g., patterned MMPI rules). All categories of statistical formulas, except logically constructed rules, yielded effect sizes in which the confidence interval did not include zero. Logically constructed rules did not differ from clinical prediction methods. Hence, statistical methods are more accurate than clinical methods but only when purely statistical models are used. Logical rules are unexpectedly no better and no worse than clinical methods. Data were insufficient to compare models mechanizing clinicians' judgment processes to the clinical method.

Amount of information. Yet another controversy surrounding clinical versus statistical prediction concerns the amount of information made available to clinicians and the formula. Holt (1970), for instance, noted that in many studies the only data available to clinicians were quantitative (e.g., an MMPI profile). This, Holt contended, placed the clinical method at a serious disadvantage because "given a chance to show what it can do, it has to have meaningful, qualitative data to work with" (p. 343). In contrast, Dawes et al. (1989) argued that even if clinicians were provided more information than the formulas, regardless of whether the information was qualitative, they will still fail to surpass the statistical method (see also Faust, 1986, 2003). This assumption was supported by Grove et al. (2000). To resolve this controversy, we studied the influence of the amount of information provided to the clinicians and the statistical formula. We hypothesized, in accordance with Dawes et al. (1989), Faust (1986, 2003), and Grove et al. (2000), that the differential accuracy between clinical and statistical prediction would not be affected by the amount of informational cues.

Our findings indicate that the amount of information provided to the clinicians influenced their prediction accuracy, Q_B (2) = 4.06, $p < .05$. Increasing the amount of information, however, decreased clinicians' judgment accuracy. More information may thus not be better.

Information about base rates. Humans tend to ignore base rate information when making judgments under ambiguous conditions (Kahneman & Tversky, 1973). Several studies reviewed in this meta-analysis investigated the influence of base rate information on clinicians' predictions. Because base rate information is usually underutilized, we anticipated that effect sizes would be the same when clinicians knew the base rate for the outcome they were predicting (i.e., base rate provided or base rate the same as the natural setting) and when base rate information was not provided.

This was the case. Whether clinicians had base rate information available when making judgments did not matter, Q_B (2) = 4.25, $p > .05$. Even so, the difference between the two methods tended to diminish when clinicians had base rate information ($d^+ = -.02$). The range of this small effect includes zero, which suggests no difference between the two methods. The statistical method surpassed the clinical method when the clinicians were not informed about base rate ($d^+ = -.13$) and when the base rate for the prediction was the same as in the clinician's work setting ($d^+ = -.11$).

Availability of statistical formula. An important question raised in the debate about clinical and statistical prediction is whether clinical prediction can be improved when clinicians are given the statistical formula and its outcome. Because some studies provided clinicians with the statistical formula before they made predictions, we could compare clinical and statistical methods with and without access to the formula. In their narrative review of the literature, both Sines (1970) and Dawes et al. (1989) stated that even though clinicians had access to and were free to use the statistical formulas, their predictions remained less accurate than statistical prediction methods. Therefore, we hypothesized that accessibility to the statistical formula would not influence the differential accuracy of clinical and statistical prediction.

As predicted, clinicians' access to the statistical formulas did not improve their accuracy. With and without access to statistical formulas, clinical prediction remained less accurate than statistical prediction, Q_B (1) = 2.70, $p > .05$. In the five studies in which clinicians were provided with the statistical formula, the effect size was –.14, and in the 40 studies in which the clinicians did not have access, the effect size was –.09. In both these instances, the confidence interval did not include zero, providing support for the stability of the effect.

Clinical expertness. Holt (1970) noted that studies often failed to compare comparable clinical and statistical prediction methods. For example, the average judge (not the best judge) was compared with the best statistical formulas (for exceptions, see Goldberg, 1965; Oskamp, 1962). Although few studies compared the best judge with the best formula, clinicians identified as experts in the predictions under study may provide a more competitive contrast with the statistical formula. Spengler et al. (2005) found in a recent meta-analysis a small but reliable increase in clinical judgment accuracy with increased experience. They did not, however, assess the impact of expertness as distinct from experience on judgment accuracy. Drawing on the notion that clinicians vary considerably in their ability to make accurate judgments and that the best clinicians can do well (Holt, 1970), we hypothesized that experts in the prediction task would be more accurate than nonexperts. Likewise, we anticipated that the performance of experts would be similar to or better than that of statistical methods.

Our predictions were not supported, Q_B (1) = 2.47, $p > .05$, even though a trend in the hypothesized direction was observed. In seven studies, clinicians were considered experts in the judgment task. These studies yielded an effect size of –.05, although the confidence interval included zero. In the 41 studies in which judges were not considered experts, d^+ was larger (–.12) and the confidence interval did not cross zero. Thus, when judgments are made by expert clinicians, the difference between clinical and statistical methods seems to disappear. However, when the clinicians are nonexperts, they are consistently outperformed by statistical formulas.

Publication source. The direction of publication bias is that studies reporting significant results are more often published than studies presenting nonsignificant results (e.g., Greenwald, 1975; Rosnow & Rosenthal, 1989). Because of competition for publication in major journals such as those published by the American Psychological Association, studies in these journals may have larger effects. From this assumption, we anticipated that the differential accuracy of clinical and statistical prediction would be larger in studies retrieved from American Psychological Association journals than from studies retrieved from other sources. This hypothesis was not supported, Q_B (2) = 2.06, $p > .05$.

Confidence in the criterion for accuracy. We assigned a dichotomous rating of high or low to the confidence we placed in the reliability of each study's criterion for accuracy (see Spengler et al., 2005). If we determined that the criteria had suspected unreliability, we rated our confidence in the accuracy criterion as low (e.g., peer or supervisor ratings of successful academic training; Kelly & Fiske, 1950). A high rating was given if use of a relatively reliable and valid criterion for accuracy was reported (e.g., neuroradiological or operative verification of brain damage; Heaton, Grant, Anthony, & Lehman, 1981). We hypothesized larger effect sizes among studies with a highly valid and reliable criterion for accuracy owing to the well-known relation between criterion reliability and a ceiling on predictive validity (for further discussion, see Schmidt & Hunter, 1996). This hypothesis was not confirmed, Q_B (1) = .12, $p > .05$. Furthermore, the effect-size confidence interval did not cross zero. Therefore one may conclude that the greater accuracy of statistical over clinical prediction is not affected by our rating of the quality of the criterion used to determine accuracy.

Other independent variables. Three additional independent variables were identified to clarify the differential accuracy of clinical and statistical prediction. These were the number of prediction tasks performed within a study, the number of clinicians making predictions, and year of publication or completion. While more reliable values were expected for higher numbers of predictions and greater numbers of clinicians, no specific hypotheses were associated with these variables. Similar to assumptions made by Spengler et al. (2005), we anticipated that as study age becomes more recent, the clinical method would fair better in relation to the statistical. This trend would be because of the increasing attention given through the years to improving the accuracy of clinical judgment (e.g., Arkes, 1981, 1991; Dawes et al., 1989; Faust, 1986; Garb, 1989, 1998; Spengler et al., 1995). Regression analyses (i.e., focused comparison; Rosenthal & Rubin, 1982) were performed to examine the effects of these three independent variables on the overall effect size. None of these continuous variables significantly predicted the effect size ($p > .05$).

DISCUSSION

Our examination of the differential prediction accuracy of clinical and statistical methods from studies completed over a 56-year span shows that in general, statistical prediction methods are somewhat more accurate than the clinical method. This confirms, in most instances, the independent and parallel meta-analytic findings of Grove et al. (2000); it is also in accord with earlier narrative reviews of clinical and statistical prediction (e.g., Dawes et al., 1989; Grove & Meehl, 1996; Meehl, 1954; Sawyer, 1966). Using the most conservative sample of studies, we found an effect size of –.12 favoring the accuracy of statistical over clinical methods. This overall effect is virtually identical to the effect size (.12) of Grove et al. across a wide range of decision contexts. Note that although the signs of effects from the two studies are different, this difference is only because of the opposite coding method used. These effect sizes reflect a 13% increase in accuracy using statistical rather than clinical prediction techniques.

Should a relatively small effect, such as that which we observed, be dismissed as unimportant? We think not. First, partisan arguments have appeared in the literature that strongly favor either clinical or statistical predictions. Our analysis and that of Grove et al. (2000) argue for more temperance on both sides. Although the statistical method is almost always the equal of the clinical method and is often better, the improvement is not overwhelming. Much more research is needed—in particular, programmatic lines of research on statistical prediction—that translates into practical applications for practicing psychologists (e.g., Quinsey et al., 1998). Likewise, supporters of clinical decision making must show how their approach can be improved.

A second reason for not ignoring such a modest effect involves how it is to be used. Consider heart attack prevention. Would not any improvement here be important? The Steering Committee of the Physicians Health Study

Research Group (1988) thought so. They discontinued a randomized double-blind experiment on the use of aspirin to reduce heart attacks after finding a preliminary effect of r = .034. This small effect was nonetheless large enough that continuing to give control subjects a placebo would have been unethical. After converting this study's d^+ values (–.12) to r^+ (= –.06), the effect of aspirin in reducing heart attack is half the size of the effect found in the current meta-analysis.

One area in which the statistical method is most clearly superior to the clinical approach is the prediction of violence, r = –.09. Out of 1,000 predictions of violence, the statistical method should correctly identify 90 more violent clients than will the clinical method (Rosenthal, 1991). The victims of violence would not consider this effect small. Some predictions are more important; therefore, we recommend that statistical prediction techniques be developed, considered, and used for the most important types of decisions made by counseling psychologists and other mental health professionals (e.g., danger, suicide, and/or parole).

Our study went beyond the work of Grove et al. (2000) by examining how specific aspects of the prediction studies influenced the nature of the difference between clinical and statistical prediction. These study design variables were chosen because they had been hypothesized by other reviewers to explain why clinical and statistical predictions differ in accuracy. We found that in some instances, the difference between clinical and statistical prediction was influenced by study design characteristics. Knowing the type of prediction is important: Predictions of violence or academic performance were much more accurate with statistical techniques, whereas treatment length was predicted equally well by both methods.

We found that being familiar with the prediction setting did not help clinicians do better than statistical methods; in fact, clinicians fared worse. The type of statistical formula was also important. All statistical types, except logically constructed rules, did better than clinicians. Even with the exception, the result was a draw; the clinician did not do better. Some have argued that certain studies have "stacked the deck" against the clinician by providing less information than the formula (e.g., Holt, 1970). We found that when clinicians were given the same or more information than the statistical formula, the formula did better. Information was insufficient to assess those conditions in which clinicians had less information. In contrast, the overall effect size was not influenced by clinicians' access to the statistical formula. Furthermore, the overall effect size was not influenced by the outlet in which the study was published. Finally, trends were observed that future studies should address. Clinicians considered experts in a prediction task did better than nonexperts and did as well as statistical methods.

Also, having base rate information available resulted in clinicians' approaching the prediction accuracy of statistical methods.

Implications for Practice

On the basis of our analyses, we tentatively suggest when and under what conditions counseling psychologists and other mental health practitioners might best use clinical or statistical methods to make accurate predictions about their clients (Westen & Weinberger, 2004). Certain limitations apply, however. Akin to the psychotherapy literature on empirically validated treatments (e.g., Waehler, Kalodner, Wampold, & Lichtenberg, 2000; Wampold, 1997), recommendations about what works for clinical predictions can logically be made based only on those applications that have so far been tested. Likewise, for many hypotheses that we tested, the sample sizes were so small that more research is needed before firm conclusions can be made about those judgments.

Given the convergence between our meta-analysis and the work of Grove et al. (2000), statistical rules ought to be employed when feasible. This is especially true if judgment accuracy is important and errors are costly. This recommendation, however, is made with qualification. First, as was recognized by an American Psychological Association task force on the use of psychological assessment (Meyer et al., 1998), prediction rules for many judgment tasks are scarce. That is, researchers have yet to study a large array of possible applications of statistical prediction techniques.

Second, not all statistical formulas are effective. Examples are Goldberg's rules for predicting neurotic versus psychotic diagnoses from the MMPI. These logically constructed rules were developed from large samples and were extensively cross-validated. A review of 406 samples of patients and nonpatients found that Goldberg's index did not generalize well across samples (Zalewski & Gottesman, 1991). This finding corresponds to our results showing that logically constructed rules, despite being cross-validated, were not more accurate than clinician judgments. Other statistical prediction methods, in contrast, have been found to be successful (e.g., regression formulas for predicting recidivism; Hilton et al., 2004; Quinsey et al., 1998).

Third, counseling psychologists should educate and familiarize themselves with available statistical prediction methods, such as regression formulas, test cutoff scores, and hit rates (for further information, see Anastasi & Urbina, 1996; Crocker & Algina, 1986; Greene, 2000). This is especially true for critical decisions in which false-negative judgments can be costly. Even a small increase in accuracy is important if one is predicting suicide, domestic violence, or postparole adjustment. For example, statistical formulas exist to predict violent behavior (Hilton et al., 2004; Quinsey et al., 1998), as do cutoff scores and hit rates for improving classification accuracy using several psychological tests (e.g., MMPI/MMPI-2 cutoffs for malingering; Graham, Watts, & Timbrook, 1991; posttraumatic stress disorder; Bury & Bagby, 2002; substance abuse; Rouse et al., 1999; Stein et al., 1999). Ignoring available statistical prediction schemes may do a disservice to clients and could even be unethical when false-negative outcomes carry severe consequences (e.g., Dawes, 2002). Counseling psychologists already rely on nomothetic assessment techniques when making predictions about clients (e.g., Strong Interest Inventory, MMPI-2/MMPI-A, or Millon Multiaxial Clinical Inventory-III). Use of tests and inventories as statistical strategies requires knowledge of models and test and measurement principles to effectively bridge nomothetic findings to the individual situation (for further discussion, see Faust, 1997). Furthermore, counseling psychologists already rely on test cutoff scores, hit rates, and decision trees to aid accurate classification (e.g., the Substance Abuse Subtle Screening Inventory; Miller, 1985). Our findings suggest that counseling psychologists should highly weigh scores from valid and reliable psychological tests in their clinical decision making.

Fourth, when counseling psychologists have familiarized themselves with available statistical formulas and prediction techniques, they should use those formulas and techniques to improve their prediction accuracy. While this is especially true when judgment accuracy is critical, such as predicting future violence and offense, it is also true for predictions of prognosis, psychological adjustment, and academic performance. Several suggestions can be drawn from the literature reviewed here. For instance, in studies examining prognostic predictions (e.g., improvement in psychotherapy; Barron, 1953; Bolton, Butler, & Wright, 1968; Kaplan, 1962; Wirt, 1956; Wittman & Steinberg, 1944), the statistical prediction involved cutoff scores from psychological tests in almost all instances (e.g., MMPI profile within normal limits and high scores on Barron's ego strength scale). Thus on the basis of these findings, counseling psychologists should place substantial weight on valid and reliable instruments when making prognostic predictions about their clients. Furthermore, counseling psychologists should rely on past academic accomplishment and scores from aptitude tests when predicting academic success, especially as it relates to completion of undergraduate training (Alexakos, 1966; Conrad & Satter, 1954; Melton, 1952; Sarbin, 1942; Watley, 1966). We call on research to develop statistical formulas especially relevant to counseling psychology practice to perform these and other tasks. For instance, empirical tests of the relevance of GPA and Graduate Record Examination scores in predicting success in counseling psychology programs are needed, either alone or in combination with other hypothesized predictors of successful training.

Fifth, counseling psychologists should use available base rate information when making decisions to aid their accuracy. Prevalence data routinely reported in epidemiological studies (e.g., National Comorbidity Survey; Kessler et al., 2003) and reported in *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.; *DSM-IV-TR*; American Psychiatric Association, 2000) are useful, as are normative data for specific settings or client type. Prevalence data may aid diagnostic accuracy such that more prevalent diagnoses (higher base rate) are more likely to be accurate than less prevalent diagnoses, given certain symptoms. Spengler et al. (1995) hypothesized that teaching counseling psychologists how to use base rates in their decision making would increase judgment accuracy, and they found support for this in a pre- and posttest control group comparison of a 6-week educational course on clinical decision making (see Spengler & Strohmer, 2001). Further research is needed on the use of base rates for increasing the accuracy of clinical judgments.

Sixth, counseling psychologists should be skeptical about the relative accuracy of their clinical judgments, even when they are working with familiar cases in familiar contexts. Studies show that clinical predictions were worse than statistical predictions in these conditions. These counterintuitive results may be because of clinicians' overconfidence or habituation to the setting (e.g., Arkes, 1981, 1991; Faust, 1986). When working with familiar cases, clinicians may become overly confident in their decision making and use a positive hypothesis-testing strategy. Such strategies serve to confirm hypotheses by recalling supporting information and discounting or forgetting contradictory information. Incorporating statistical methods in decision making involving familiar cases could reduce this bias.

Finally, in several instances, clinical judgment was the equal of statistical methods. Our findings show that clinicians can rely on either their own inferential processes (the clinical method) or on statistical methods when detecting brain impairment from psychological tests, determining personality characteristics, predicting length of hospitalization and treatment, and diagnosing. In these predictions, clinicians were as accurate as the statistical approaches. Two reasons might account for these findings. First, experience with these tasks may have provided clinicians with information regarding validity of relations between client data and outcome (Dawes, 1994; Ericsson & Lehman, 1996; Faust, 1991; Lichtenberg, 1997; Spengler, 1998). In fact, these prediction tasks, in comparison with the tasks in which mechanical methods were more accurate, tend to offer a greater opportunity for immediate feedback to the practicing clinician. Second, these prediction tasks may have more objective tools available to aid judgment accuracy, such as psychological tests and systematic guidelines (e.g., use of *DSM-IV-TR* and test cutoff score for classification), than other prediction tasks in which statistical methods surpassed the clinical method.

Implications for Training

Counseling psychology has given much attention to the importance of the scientist-practitioner model of training. Epistemological issues in research and practice have been discussed, and the use of empirically validated treatment strategies for specific clients and problems has been encouraged. Our analysis supports the argument that training should also encourage the use of statistical methods to decrease judgment biases and errors. A portion of pre- and postdoctoral training could be used to show trainees how frequently they have the opportunity to use statistical methods of prediction and how these methods could improve their predictions. Test cutoff scores, hit rates, and decision trees are all statistical strategies and can improve clinical predictions. Adequate training in statistics and probability theory has been found to increase judgment accuracy (Nisbett & Ross, 1980), as has education about inferential errors and common heuristics (e.g., Arkes, 1981, 1991; Spengler & Strohmer, 2001). Methods for demystifying statistical prediction should be incorporated into counseling psychology curricula. Furthermore, trainees should be familiarized with the construction of simple regression models to aid judgment accuracy. Some research on the benefits of training counselors in effective decision-making strategies already exists (Berven, 1985; Berven & Scofield, 1980; Falvey & Hebert, 1992; Kurpius, Benhamin, & Morran, 1985; Spengler & Strohmer, 2001). More research is needed on how best to implement statistical decision making and comparing its effectiveness in different contexts before educators and trainers can most effectively introduce statistical models of decision making.

In their proposed scientist-practitioner model for assessment, Spengler et al. (1995) suggested that counseling psychologists collect local clinical data for decisions that might be considered routine; these data would be used to evaluate and improve the effectiveness of their practice. Psychologists could then use these data to mechanize routine decisions. Counseling psychologists could also adjust existing statistical formulas to fit their client type and setting, basing their adjustment on local data. Some of the many decisions that might be assisted by statistical formulas include amount of treatment, school-to-work transition, career choice, and predictions of dangerousness to self or others. Blocher (1987) provided a simple and accessible model for counseling psychologists to learn how to empirically evaluate their practice. Likewise, Stricker (e.g., Stricker & Trierweiler, 1995) has written extensively about training psychologists as local clinical scientists who collect local practice data to improve their decision making.

Counseling psychologists should be trained in statistical methods for evaluating their effectiveness with individual clients. Examples of promising statistical prediction techniques for clinical practice come from a newer area of psychotherapy research called patient-focused research (Howard, Moras, Bril, Martinovich, & Lutz, 1996). By using statistical techniques, such as probit analysis, survival analysis, and hierarchical linear modeling, an individual client's progress is compared with expected recovery curves to improve clinical decision making. Lambert, Hansen, and Finch (2001) developed a methodology based on a data set of 10,000 cases to identify clients likely to fail in treatment (i.e., signal cases). They used the Outcome Questionnaire-45 (Lambert et al., 1996) to monitor client weekly progress. Providing clinicians with the simple feedback to change the course of their treatment, compared with a no-feedback group, led to statistically and clinically significant improvements in recovery rates for the signal cases. Another area of statistical prediction research relevant to counseling training applications is the use of statistical techniques to measure clinically significant and reliable change (Ogles, Lambert, & Masters, 1996).

Implications for Research

This meta-analysis represents only the second meta-analysis conducted in this area of the literature (cf. Grove et al., 2000). The present findings are not without limitations. The arguments in favor of the small, but reliable, edge of statistical prediction techniques are strong, but we are struck by the limits of these studies. Few programmatic lines of research have accumulated bodies of evidence for specific applications (e.g., clinical versus statistical methods to aid suicide assessment risk). More systematic studies, such as those performed by Goldberg (1965) and Quinsey et al. (1998), are clearly needed. Likewise, more models should be developed for clinical practice. Otherwise, these findings will remain only an academic issue of little practical interest to mental health professionals.

A recent proposal by Katsikopoulos, Machery, Pachur, and Wallin (2004) suggested development of user-friendly models, arguing that statistical models must be context based, simple, and friendly for the clinician. Katsikopolous et al. proposed constructing statistical methods they termed "friendly heuristics." These friendly heuristics are simple guidelines that rely on a limited amount of client data and do not need complex integration. Often, these friendly heuristics mirror the cognitive processes underlying clinical judgment.

Examples include relying on few but important client data to render a prediction (e.g., if X communicates a desire to die and has access to a gun, X is likely to commit suicide). Similarly, test cutoff scores and hit rates reported for many psychological tests are simple and easy to use and therefore friendly. To further aid judgment accuracy and save time, more friendly heuristics for use in clinical settings must be developed and tested.

A common criticism of the body of studies on clinical versus statistical prediction accuracy is lack of ecological validity (e.g., Holt, 1970; Rock et al., 1987; Westen & Weinberger, 2004). This criticism has some merit. In various studies, the clinicians made global dichotomous judgments about events that they seldom encounter and about which they may have limited knowledge or training. More research is clearly needed on predictions common to counseling psychologists that incorporate statistical and clinical prediction methods.

In conclusion, we do not argue against clinical prediction as a decision-making strategy (cf. Hammond, 1996). Counseling psychologists make far too many decisions in which the absolute right answer is not the issue and in which determining statistical decision rules may never be practical (e.g., moment-to-moment decisions; Spengler et al., 1985). For these decisions, the clinical strategy is necessary. However, after reviewing 56 years of research, we conclude that clinical prediction should not be the only method. After being shown by two meta-analyses and several independent analyses to be at least equal and often superior to clinical decision making in counseling psychology and mental health contexts, statistical methods must be one of the strategies of the careful clinician. Quoting Meehl (1954), “We have no right to assume that entering the clinic has resulted in some miraculous mutations and made us singularly free from the ordinary human errors which characterized our psychological ancestors” (p. 28).

NOTES

1. A complete description of the search process can be found in Spengler et al. (2005). To limit the retrieval of studies to a manageable, yet representative, sample, studies that appeared between 1970 and 1996 were included in the search. Electronic databases included PsychInfo, ERIC, Dissertation Abstracts, BRS, MEDLINE, and Social Science Index. Unavailable dissertations and journal articles were purchased, and authors were contacted to obtain material that was difficult to retrieve. After we identified likely clinical judgment studies, forward and backward cross-referencing was conducted until no new studies were obtained. By using this strategy, more than 35,000 articles were identified; 4,617 were coded and 1,135 met our inclusion criteria for the project. We chose this open-ended strategy to maximize the number of studies that would be reviewed. Because our search did not include studies published after 1996, we performed a file drawer analysis. In this analysis, we can project how many studies reporting significantly different effect sizes will be needed to change our overall results. This analysis indicated that to reduce our effect size to zero (within the 95% confidence interval

boundaries), indicating no difference between clinical and statistical prediction, 99 additional studies are needed using cross-validated formulas and producing nonoutlying effect sizes.

2. Hit rates reported as percentage correct were directly transformed into *d* using the DSTAT program (Johnson, 1993). When hit rates were reported as a correlation of the prediction with the criterion (*r*) for clinical and statistical prediction, the *r* values were first converted to *Z* scores. Differences in *Z* scores were then transformed into a chi-square value. The chi-square value of *Z* score differences was transformed into *d* (Rosenthal, 1991).

3. These results are not reported but are available on request by contacting the first author.

4. Johnson (1993) and Hedges and Olkin (1985) recommend outlier analyses in meta-analyses when moderating variables fail to explain observed heterogeneity of studies. In this procedure, outliers are sequentially removed until the hypothesis of homogeneity cannot be rejected (i.e., the probability of Hedges's *Q* exceeds .05). In the current meta-analysis, the study design characteristics that we hypothesized would influence the overall effect size (e.g., prediction task) did not reduce the heterogeneity among the studies. This made the interpretation of the effect size difficult because the many studies differed from each other in terms of magnitude and direction of the effects. Therefore to make our findings more interpretable, we performed an outlier analysis.

REFERENCES

*References marked with an asterisk indicate studies included in the meta-analysis.

Adams, K. M. (1974). *Automated clinical interpretation of the neuropsychological battery: An ability-based approach*. Unpublished doctoral dissertation, Wayne State University, Detroit, Michigan.

*Alexakos, C. E. (1966). Predictive efficiency of two multivariate statistical techniques in comparison with clinical predictions. *Journal of Educational Psychology*, *57*, 297-306.

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

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology*, *49*, 323-330.

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, *110*, 486-498.

Anastasi, A., & Urbina, S. (1996). *Psychological testing* (7th ed.). New York: Prentice Hall.

*Astrup, C. A. (1975). Predicted and observed outcome in followed-up functional psychosis. *Biological Psychiatry*, *10*, 323-328.

*Barron, F. (1953). Some test correlates of response to psychotherapy. *Journal of Consulting Psychology*, *17*, 235-241.

Berven, N. L. (1985). Reliability and validity of standardized case management simulations. *Journal of Counseling Psychology*, *32*, 397-409.

Berven, N. L., & Scofield, M. E. (1980). Evaluation of clinical problem-solving skills through standardized case-management simulations. *Journal of Counseling Psychology*, *27*, 199-208.

Blocher, D. H. (1987). Process models for professional counseling. In *The professional counselor*. New York: MacMillan.

Blumetti, A. E. (1972). *A test of clinical versus actuarial prediction: A consideration of accuracy and cognitive functioning*. Unpublished doctoral dissertation, University of Florida, Gainesville.

Bolton, B. F., Butler, A. J., & Wright, G. N. (1968). Clinical versus statistical prediction of client feasibility. *Wisconsin Studies in Vocational Rehabilitation* [Monograph VII], University of Wisconsin Regional Rehabilitation Research Institute, Madison, WI.

Bury, A. S., & Bagby, R. M. (2002). The detection of feigned uncoached and coached posttraumatic stress disorder with the MMPI-2 in a sample of workplace accident victims. *Psychological Assessment*, *14*, 472-484.

*Carlin, A. S., & Hewitt, P. L. (1990). The discrimination of patient-generated and randomly generated MMPIs. *Journal of Personality Assessment*, *54*, 24-29.

Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Conrad, H. S., & Satter, G. A. (1954). *The use of test scores and quality-classification ratings in predicting success in electrician's mates school. Project N-106: Research and development of the Navy's aptitude testing program*. Princeton, NJ: Research and Statistical Laboratory College Entrance Examination Board.

*Cooke, J. K. (1967a). Clinicians' decisions as a basis for deriving actuarial formulae. *Journal of Clinical Psychology*, *23*, 232-233.

*Cooke, J. K. (1967b). MMPI in actuarial diagnosis of psychological disturbance among college males. *Journal of Counseling Psychology*, *14*, 474-477.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

*Danet, B. N. (1965). Prediction of mental illness in college students on the basis of "nonpsychiatric" MMPI profiles. *Journal of Counseling Psychology*, *29*, 577-580.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York: Free Press.

Dawes, R. M. (2002). The ethics of using or not using statistical prediction rules in psychological practice and related consulting activities. *Philosophy of Science*, *69*, S178-S184.

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

Dawes, R. M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668-1674.

*Devries, A. G., & Shneidman, E. S. (1967). Multiple MMPI profiles of suicidal persons. *Psychological Reports*, *21*, 401-405.

*Dickerson, J. H. (1958). *The Biographical Inventory compared with clinical prediction of post counseling behavior of V.A. hospital counselors*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

*Dunham, H. W., & Meltzer, B. N. (1946). Predicting length of hospitalization of mental patients. *American Journal of Sociology*, *52*, 123-131.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*, 36-48.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, *47*, 273-305.

*Evenson, R. C., Altman, H., Sletten, I. W., & Cho, D. W. (1975). Accuracy of actuarial and clinical predictions for length of stay and unauthorized absence. *Diseases of the Nervous System*, *36*, 250-252.

Falvey, J. E., & Hebert, D. J. (1992). Psychometric study of clinical treatment planning simulation (CTPS) for assessing clinical judgment. *Journal of Mental Health Counseling*, *14*, 490-507.

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

Faust, D. (1991). What if we had really listened? Present reflections on altered pasts. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Vol. I. Matters of public interest* (pp. 185-217). Minneapolis: University of Minnesota Press.

Faust, D. (1997). Of science, meta-science, and clinical practice: The generalization of a generalization to a particular. *Journal of Personality Assessment*, *68*, 331-354.

Faust, D. (2003). Holistic thinking is not the whole story: Alternative or adjunct approaches for increasing the accuracy of legal evaluations. *Assessment*, *10*, 428-411.

Fero, D. D. (1975). *A lens model analysis of the effects of amount of information and mechanical decision making aid on clinical judgment and confidence*. Unpublished doctoral dissertation, Bowling Green State University, Bowling Green, OH.

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

Garb, H. N. (1994). Toward a second generation of statistical prediction rules in psychodiagnosis and personality assessment. *Computers in Human Behavior*, *10*, 377-394.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association.

*Gardner, W., Lidz, C. W., Mulvay, E. P., & Shaw, E. C. (1996). Clinical versus actuarial predictions of violence in patients with mental illnesses. *Journal of Consulting and Clinical Psychology*, *64*, 602-609.

*Goldberg, L. R. (1965). Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs: General and Applied*, *79*(9), 1-27.

*Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving clinical inferences. *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.

Gottfredson, D. M., & Snyder, H. N. (2005). *The mathematics of risk classification: Changing data into valid instruments for juvenile courts* (NCJ 209158). Washington, DC: National Center for Juvenile Justice, Office of Juvenile Justice and Delinquency Prevention.

Graham, J. R., Watts, D., & Timbrook, R. E. (1991). Detecting fake-good and fake-bad MMPI-2 profiles. *Journal of Personality Assessment*, *57*, 264-277.

*Grebstein, L. C. (1963). Relative accuracy of actuarial prediction, experienced clinicians, and graduate students in a clinical judgment task. *Journal of Consulting Psychology*, *27*, 127-132.

Greene, R. L. (2000). *The MMPI-2: An interpretive manual* (2nd ed.). Boston: Allyn & Bacon.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1-20.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*, 293-323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical vs. mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19-30.

*Gustafson, D. H., Greist, J. H., Stauss, F. F., Erdman, H., & Laughren, T. (1977). A probabilistic system for identifying suicide attempters. *Computers and Biomedical Research*, *10*, 83-89.

*Halbower, C. C. (1955). *A comparison of actuarial versus clinical prediction to classes discriminated by the Minnesota Multiphasic Personality Inventory*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

*Hall, G. C. N. (1988). Criminal behavior as a function of clinical and actuarial variables in a sexual offender population. *Journal of Consulting and Clinical Psychology*, *56*, 773-775.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable justice*. New York: Oxford University Press.

Hare, R. D. (1991). *The revised psychopathy checklist*. Toronto, Canada: Multi-Health Systems.

Harvey-Cook, J. E., & Taffler, R. J. (2000). Biodata in professional entry-level selection: Statistical scoring of common format applications. *Journal of Occupational and Organizational Psychology*, *73*, 103-118.

*Heaton, R. K., Grant, I., Anthony, W. Z., & Lehman, R. A. (1981). A comparison of clinical and automated interpretation of the Halstead-Reitan battery. *Journal of Clinical Neuropsychology*, *3*, 121-141.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science: The empirical cumulativeness of research. *American Psychologist*, *42*, 443-455.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hilton, N. Z., Harris, G. T., Rice, M. E., Lang, C., Cormier, C. A., & Lines, K. J. (2004). A brief actuarial assessment for the prediction of wife assault recidivism: The Ontario Domestic Assault Risk Assessment. *Psychological Assessment*, *16*, 267-275.

*Holland, T. R., Holt, N., Levi, M., & Beckett, G. E. (1983). Comparison and combination of clinical and statistical predictions of recidivism among adult offenders. *Journal of Applied Psychology*, *68*, 203-211.

Holt, R. R. (1958). Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology*, *56*, 1-12.

Holt, R. R. (1970). Yet another look at clinical and statistical prediction: Or, is clinical psychology worthwhile? *American Psychologist*, *25*, 337-349.

*Hovey, H. B., & Stauffacher, J. C. (1953). Intuitive versus objective prediction from a test. *Journal of Clinical Psychology*, *9*, 341-351.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z. M., & Lutz, W. (1996). Efficacy, effectiveness, and patient progress. *American Psychologist*, *51*, 1059-1064.

Hunter, J. E., & Smith, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Johnson, B. T. (1993). *DSTAT 1.10: Software for the meta-analytic review of research literature.* Hillsdale, NJ: Lawrence Erlbaum.

Johnson, B. T., & Eagly, A. H. (2000). Quantitative synthesis of social psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social psychology.* London: Cambridge University Press.

*Johnston, R., & McNeal, B. F. (1967). Statistical versus clinical prediction: Length of neuropsychiatric hospital stay. *Journal of Abnormal Psychology*, *72*, 335-340.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *801*, 237-251.

*Kaplan, R. L. (1962). *A comparison of actuarial and clinical predictions of improvement in psychotherapy*. Unpublished doctoral dissertation, University of California, Los Angeles.

Katsikopoulos, K., Machery, E., Pachur, T., & Wallin, A. (2004). *The search for models of clinical judgment: Fast, frugal, and friendly in Paul Meehl's spirit.* Unpublished manuscript. Retrieved October 15, 2004, from http://jeannicod.ccsd.cnrs.fr/documents/disk0/00/00/05/33/ijn_00000533_00/ijn_00000533_00.pdf

Kelly, E. L., & Fiske, D. W. (1950). The prediction of success in the VA training program in clinical psychology. *American Psychologist*, *5*, 395-406.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., et al. (2003). The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association*, *289*, 3095-3106.

*Klehr, R. (1949). Clinical intuition and test scores as a basis for diagnosis. *Journal of Consulting Psychology*, *13*, 34-38.

*Kleinmuntz, B. (1967). Sign and seer: Another example. *Journal of Abnormal Psychology*, *72*, 163-165.

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

*Klinger, E., & Roth, I. (1965). Diagnosis of schizophrenia by Rorschach patterns. *Journal of Projective Techniques and Personality Assessment*, *29*, 323-335.

Kurpius, D. J., Benjamin, D., & Morran, D. K. (1985). Effect of teaching a cognitive strategy on counselor trainee internal dialogue and clinical hypothesis formulation. *Journal of Counseling Psychology*, *32*, 262-271.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology*, *69*, 159-172.

Lambert, M. J., Hansen, N. B., Umphress, V., Lunnen, K., Okiishi, J., Burlingame, G., et al. (1996). *Administration and scoring manual for the Outcome Questionnaire (OQ 45.2)*. Wilmington, DE: American Professional Credentialing Services.

*Leli, D. A., & Filskov, S. B. (1981). Clinical-actuarial detection and description of brain impairment with the W-B Form 1. *Journal of Clinical Psychology*, *37*, 623-629.

*Leli, D. A., & Filskov, S. B. (1984). Clinical detection of intellectual deterioration associated with brain damage. *Journal of Clinical Psychology*, *40*, 1435-1441.

Lemerond, J. N. (1977). *Suicide prediction for psychiatric patients: A comparison of the MMPI and clinical judgments*. Unpublished doctoral dissertation, Marquette University, Madison, WI.

*Lewis, E. C., & MacKinney, A. C. (1961). Counselor vs. statistical prediction of job satisfaction in engineering. *Journal of Counseling Psychology*, *8*, 224-230.

Lefkowitz, M. B. (1973). *Statistical and clinical approaches to the identification of couples at risk in marriage*. Unpublished doctoral dissertation, University of Florida, Gainesville.

Lichtenberg, J. W. (1997). Expertise in counseling psychology: A concept in search of support. *Educational Psychology Review*, *9*, 221-238.

*Lindsey, G. R. (1965). Seer versus sign. *Journal of Experimental Research in Personality*, *1*, 17-26.

Lyle, O., & Quast, W. (1976). The Bender Gestalt: Use of clinical judgment versus recall scores in prediction of Huntington's disease. *Journal of Consulting and Clinical Psychology*, *44*, 229-232.

McHugh, R. B., & Apostolakos, P. C. (1959). Methodology for the comparison of clinical with actuarial predictions. *Psychological Bulletin*, *56*, 301-309.

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. (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.

*Melton, R. S. (1952). *A comparison of clinical and actuarial methods of prediction with an assessment of the relative accuracy of different clinicians*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Meyer, K. (1973). *The effect of training in the accuracy and appropriateness of clinical judgment*. Unpublished doctoral dissertation, Adelphi University, Garden City, NY.

Meyer, J. J., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. L., et al. (1998). *Benefits and costs of psychological assessment in healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group* (Part I). Washington, DC: American Psychological Association.

*Miller, D. E., Kunce, J. T., & Getsinger, S. H. (1972). Prediction of job success for clients with hearing loss. *Rehabilitation Counseling Bulletin*, *16*, 21-29.

Miller, G. A. (1985). *The Substance Abuse Subtle Screening Inventory (SASSI): Manual*. Bloomington, IN: Spencer Evening World.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

Ogles, B., Lambert, M. J., & Masters, K. S. (1996). *Assessing outcome in clinical practice*. Needham Heights, MA: Allyn & Bacon.

*Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs: General and Applied*, *76*, 1-27.

Oxman, T. E., Rosenberg, S. D., Schnurr, P. P., & Tucker, G. J. (1988). Diagnostic classification through content analysis of patients' speech. *American Journal of Psychiatry*, *145*, 464-468.

Pepinsky, H. B., & Pepinsky, N. (1954). *Counseling theory and practice*. New York: Ronald Press.

*Perez, F. I. (1976). Behavioral analysis of clinical judgment. *Perceptual and Motor Skills*, *43*, 711-718.

*Popovics, A. J. (1983). Predictive validities of clinical and actuarial scores of the Gesell Incomplete Man Test. *Perceptual and Motor Skills*, *56*, 864-866.

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (1998). *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association.

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

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed). Newbury Park, CA: Sage.

Rosenthal, R., & Rubin, D. (1982). A simple general purpose display of magnitude of experimental effects. *Journal of Educational Psychology*, *74*, 166-169.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276-1284.

Rouse, S. V., Butcher, J. N., & Miller, K. B. (1999). Assessment of substance abuse in psychotherapy clients: The effectiveness of the MMPI-2 substance abuse scales. *Psychological Assessment*, *11*, 101-107.

*Russell, E. W. (1995). The accuracy of automated and clinical detection of brain damage and lateralization on neuropsychology. *Neuropsychology Review*, *5*, 1-68.

Sarbin, T. L. (1942). A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology*, *48*, 593-602.

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

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199-223.

Shaffer, J. W., Perlin, S., Schmidt, C. W., & Stephens, J. H. (1974). The prediction of suicide in schizophrenia. *Journal of Nervous and Mental Disease*, *150*, 349-355.

Shagoury, P., & Satz, P. (1969). The effect of statistical information on clinical prediction. *Proceedings of the 77th Annual Convention of the American Psychological Association*, *4*, 517-518.

Sines, J. O. (1970). Actuarial versus clinical prediction in psychopathology. *British Journal of Psychiatry*, *116*, 129-144.

Spengler, P. M. (1998). Multicultural assessment and a scientist-practitioner model of psychological assessment. *The Counseling Psychologist*, *26*, 930-938.

Spengler, P. M., & Strohmer, D. C. (2001, August). *Empirical analyses of a scientist-practitioner model of assessment*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Spengler, P. M., Strohmer, D. M., Dixon, D. N., & Shivy, V. A. (1995). A scientist-practitioner model of psychological assessment: Implications for training, practice, and research. *The Counseling Psychologist*, *23*, 506-534.

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2005). *The meta-analysis of clinical judgment project: Effects of experience on judgment accuracy*. Manuscript submitted for publication.

Steering Committee of the Physicians Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine*, *318*, 262-264.

Stein, L. A. R., Graham, J. R., Ben-Porath, Y. S., & McNulty, J. L. (1999). Using the MMPI-2 to detect substance abuse in an outpatient mental health setting. *Psychological Assessment*, *11*, 94-100.

Stricker, G. (1967). Actuarial, naïve clinical, and sophisticated clinical prediction of pathology from figure drawings. *Journal of Consulting Psychology*, *31,* 492-494.

Stricker, G., & Trieweiler, S. J. (1995). The local clinical scientist: A bridge between science and practice. *American Psychologist*, *50*, 995-1002.

Sullivan, E., Cirincione, C., Nelson, K., & Wallis, J. (2001). *Classifying inmates for strategic programming*. Washington, DC: National Criminal Justice Service, U.S. Department of Justice.

*Szuko, J. J., & Kleinmuntz, B. (1981). Statistical versus clinical lie detection. *American Psychologist*, *36*, 488-496.

*Taulbee, E. S., & Sisson, B. D. (1957). Configurational analysis of MMPI profiles of psychiatric groups. *Journal of Consulting Psychology*, *21*, 413-417.

*Thompson, R. E. (1952). A validation of the Glueck Social Prediction Scale for proneness to delinquency. *Journal of Criminal Law*, *Criminology, and Police Science*, *43*, 451-470.

Waehler, C. A., Kalodner, C. R., Wampold, B. E., & Lichtenberg, J. W. (2000). Empirically supported treatments (ESTs) in perspective: Implications for counseling psychology training. *The Counseling Psychologist*, *28*, 657-671.

*Walters, G. D., White, T. W., & Greene, R. L. (1987). The use of MMPI to identify malingering and exaggeration of psychiatric symptomatology in male prison inmates. *Journal of Consulting and Clinical Psychology*, *1*, 111-117.

Wampold, B. E. (1997). Methodological problems in identifying efficacious psychotherapies. *Psychotherapy Research*, *7*, 21-43.

Watley, D. J. (1966). Counselor variability in making accurate predictions. *Journal of Counseling Psychology*, *13*, 53-62.

*Watley, D. J., & Vance, F. L. (1964). *Clinical versus actuarial prediction of college achievement and leadership activity*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Webb, S. C., Hultgen, D. D., & Craddick, R. A. (1977). Predicting occupational choice by clinical and statistical methods. *Journal of Counseling Psychology*, *24*, 98-110.

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

*Weinberg, G. H. (1957). *Clinical versus statistical prediction with a method of evaluating a clinical tool*. Unpublished doctoral dissertation, Columbia University, New York.

Werner, P. D., Rose, T. L., Yesavage, J. A., & Seeman, K. (1984). Psychiatrists' judgment of dangerousness in patients on an acute care unit. *American Journal of Psychiatry*, *141*, 263-266.

Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist*, *59*, 595-613.

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: The forecasting of graduate school success. *Journal of Personality and Social Psychology*, *19*, 100-106.

*Wirt, R. D. (1956). Actuarial prediction. *Journal of Consulting Psychology*, *20*, 123-124.

*Wittman, M. P., & Steinberg, L. (1944). Follow-up of an objective evaluation of prognosis in dementia praecox and manic-depressive psychosis. *The Elgin Papers*, *5*, 216-227.

Zalewski, C. E., & Gottesman, I. I. (1991). (Hu)man versus mean revisited: MMPI group data and psychiatric diagnosis. *Journal of Abnormal Psychology*, *100*, 562-568.