The Counseling Psychologist

http://tcp.sagepub.com/

The Meta-Analysis of Clinical Judgment Project : Effects of Experience on Judgment Accuracy

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

The Counseling Psychologist 2009 37: 350 originally published online 4 October 2007

DOI: 10.1177/0011000006295149

Major Contribution

The Counseling Psychologist
Volume 37 Number 3
April 2009 350-399
© 2009 SAGE Publications
10.1177/0011000006295149
http://tcp.sagepub.com
hosted at
http://online.sagepub.com

The Meta-Analysis of Clinical Judgment Project

Effects of Experience on Judgment Accuracy

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

Clinical and educational experience is one of the most commonly studied variables in clinical judgment research. Contrary to clinicians' perceptions, clinical judgment researchers have generally concluded that accuracy does not improve with increased education, training, or clinical experience. In this meta-analysis, the authors synthesized results from 75 clinical judgment studies where the experience of 4,607 clinicians was assessed in relation to the accuracy of their judgments about mental health (e.g., diagnosis, prognosis, treatment) and psychological issues (e.g., vocational, personality). The authors found a small but reliable effect, d = .12, showing that experience, whether educational or clinical, is positively associated with judgment accuracy. This small effect was robust across several tested moderator models, indicating experienced counselors and clinicians acquire, in general, almost a 13% increase in their decision-making accuracy, regardless of other factors. Results are discussed in light of their implications for clinical judgment research and for counseling psychology training and practice.

Experience is one of the most commonly studied variables in the clinical judgment literature. Its effects are also one of the most hotly debated. Wiggins (1973) asserted more than 30 years ago, "There is little empirical evidence that justifies the granting of ‘expert’ status to the clinician on the basis of his [or her] training, experience, or information-processing ability” (p. 131). By contrast, Holt (1970) argued that clinicians “vary considerably in their ability to do the job, *but the best of them can do very well*” (italics added, p. 348).[1] One does not have to search far in the contemporary clinical judgment literature to uncover contentious debate about the worth or lack of worth of experience in clinical judgment (see Garb & Grove, 2005; Westen & Weinberger, 2005). For example, Dawes (1994) argued that judgmental capabilities cannot be expected to improve with experience given the ambiguous nature of tasks performed by psychologists. Several other writers have asserted that clinical judgment will not improve with clinical or educational experience and may, in fact, actually worsen (e.g., Brehmer, 1980; Brodsky, 1998; Faust, 1986, 1994; Faust et al., 1988; Faust & Ziskin, 1988; Garb, 1989; Wedding, 1991; Wiggins, 1973; Ziskin, 1995). By contrast, others have argued that decision making should improve with experience and that the decision processes used by experienced counselors, or experts, should even serve as the standard for measuring growth in the cognitive activity of novice counselors (e.g., Berven, 1985; Berven & Scofield, 1980; Falvey & Hebert, 1992; Gambrill, 2005; Shanteau, 1988).

To begin our account from what might be called a positive perspective, there are several findings in the counseling psychology literature that suggest that judgment accuracy improves with increased clinical experience.

The Division 17 logo denotes that this article has been approved for continuing education credit. To access the CE test, please visit http://per-ce.net/ce/tcp.php.

Authors’ Note: 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, academic year research grant, new faculty research grant); and by three Lyell Bussell Summer Graduate Research Assistant Grants, Ball State University, Muncie, Indiana. Preliminary findings from the experience meta-analysis were presented at the annual meetings of the American Psychological Association in Washington, D.C., August 2000; Boston, Massachusetts, August 1999; San Francisco, California, August 1998; Toronto, Canada, August 1996; and New York, New York, August 1995 and by invited address at the annual meeting of the Society for Personality Assessment, New Orleans, Louisiana, March 1999. We wish to extend our 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 their instrumental assistance obtaining articles, coding studies, and managing data for the Meta-Analysis of Clinical Judgment (MACJ) project. Correspondence concerning this article or the MACJ research project should be addressed to Paul M. Spengler, PhD, Department of Counseling Psychology, Teachers College 622, Ball State University, Muncie IN 47306; e-mail: pspengle@bsu.edu.

Experienced counselors have been found to differ from novice counselors on a number of cognitive dimensions, including (a) broader knowledge structures, (b) better short- and long-term memory for domain-specific information, (c) efficiency in terms of time spent on client conceptualizations, (d) number of concepts generated, and (e) the quality of the their cognitive schemata about case material (Cummings, Hallberg, Martin, Slemon, & Hiebert, 1990; Holloway & Wolleat, 1980; Kivlighan & Quigley, 1991; Martin, Slemon, Hiebert, Hallberg, Mayfield, Kardash, & Kivlighan, 1999; & Cummings, 1989; O'Byrne & Goodyear, 1997). Counseling psychologists have long proposed that cognitive differences between novices and experts are associated with increasingly nuanced and accurate classification systems (cf. Parker, 1958; Pepinsky & Pepinsky, 1954). Findings from other areas of psychology suggest that expert clinicians tend to be able to apply statistical heuristics more appropriately than nonexperts do, thereby avoiding common decision-making errors. This use is likely if it is apparent that statistical reasoning is appropriate in the domain (Nisbett, Krantz, Jepson, & Kunda, 1983). This may not be the case, however, for counselors and psychologists who must often make their judgments under uncertain conditions (Tversky & Kahneman, 1974) and in domains sufficiently unstructured so as to diminish the perceived utility of statistical heuristics (Kleinmuntz, 1990).

Garb (1989, 1998), among others, noted that there are a number of impediments that hinder clinicians from improving their assessment accuracy with clinical experience. Some of these have to do with generic problems with learning from experience. These problems include biases in retrospection, hindsight bias, and availability (cf. Hastie & Dawes, 2001). In addition to these obstacles hindering judgment accuracy, biases have been identified that may be especially likely to affect counseling and clinical psychologists. Counseling psychologists, for example, may unwittingly engage in biased hypothesis-testing strategies (Pfeiffer, Whelan, & Martin, 2000; Strohmer, Shivy, & Chiodo, 1990); unknowingly invoke stereotypes about race, ethnicity, and homosexuality (Casas, Brady, & Ponterotto, 1983; Ridley, Li, & Hill, 1998; Wampold, Casas, & Atkinson, 1981; Wisch & Mahalik, 1999); and unduly minimize clients' vocational problems when more interesting personal problems coexist (Spengler, 2000; Spengler, Blustein, & Strohmer, 1990). Debriefing of research participants suggests that these types of clinical biases occur out of conscious awareness (e.g., DiNardo, 1975). Because of this, psychologists may fail to appreciate the likelihood of error in their assessments of clients, cannot perceive errors in their models of clients, and thus may fail to learn from experience (Einhorn, 1986). More than a few writers have thus concluded that the positive perspective is inaccurate, that there is little relation between clinical experience and professional competence as decision makers (for further discussion, see Lichtenberg, 1997).

If clinical experience by itself cannot be shown to improve judgment, perhaps education will. McPherson, Piseco, Elman, Crosbie-Burnett, and Sayger (2000) noted that doctoral-level counseling psychologists usually receive more training in specific assessment techniques than do master's-level counselors and, therefore, should be "better prepared to identify problems and errors with various types of psychological assessment data" (p. 696). Likewise, developmental training and supervision models predict that judgment accuracy should improve with educational experience (see, e.g., the developmental theories described by Loganbill, Hardy, & Delworth, 1982; Stoltenberg, McNeill, & Delworth, 1998). In contrast to clinical experience, which may only provide the opportunity to repeat existing strategies over time, increased educational experience may improve the quality of decision-making strategies used by counseling psychologists. This improvement would be particularly true if decision making itself were taught. This intervention has considerable appeal, even among those who are pessimistic about the role of experience in professional competence (e.g., Swets, Dawes, & Monahan, 2000).

One may think of counseling psychologists as practicing scientists (Spengler, Strohmer, Dixon, & Shivy, 1995). The emphasis on training counseling psychologists in decision-making skills dates back to Pepinksy and Pepinsky's (1954) model of the counselor-as-scientist. If scientists must be trained in how to make effective inferences from their data, then similar and continuing training should be helpful for counseling skills and related decision-making processes. Indeed, counseling psychology educators have begun to teach trainees how to make judgments under uncertainty through standardized case simulations (e.g., Falvey & Hebert, 1992) and through the use of hypothesis-testing strategies (e.g., Kurpius, Benjamin, & Morran, 1985; Meier, 1999; Spengler & Strohmer, 2001). At present, however, there are no comprehensive quantitative analyses on whether these efforts are helpful or, for that matter, whether any form of educational experience is linked to clinical judgment accuracy.

Traditional narrative reviews of the clinical judgment literature have generally concluded that experience, however defined, does not improve judgment accuracy. Faust (1994) stated,

> Numerous studies have examined mental health professionals' experience and judgmental accuracy. Anyone who has a detailed familiarity with research

> should recognize that the great bulk of past and present studies run counter to the notion that more experienced clinicians reach more accurate diagnoses or predictions. (pp. 199-200)

Faust is not alone in reaching such a conclusion. Wedding and Faust (1989), in a review of clinical judgment research, stated, “There are virtually no data suggesting judgmental accuracy is related to experience” (p. 249). In a similar critique of psychologists serving as expert witnesses, Faust and Ziskin (1988) concluded that “virtually every available study shows that amount of clinical training and experience are unrelated to judgmental accuracy” (p. 32). In what is considered to be the most comprehensive review of the clinical judgment literature, Garb (1998) concluded, “Clinical experience has generally not been related to validity, both when experienced clinicians have been compared to inexperienced clinicians and when clinicians have been compared to graduate students” (p. 110). Likewise, Lichtenberg (1997) noted, “The fact that counselors’ accuracy of clinical judgment does not increase with experience is now generally acknowledged” (p. 231).

A recent exception comes from the *Report of the 2005 Presidential Task Force on Evidenced-Based Practice* (Levant, 2005) in which experience is afforded a central role in determining best clinical practice. While members of the task force recognized that clinical decision making could be negatively affected by common cognitive errors, they also concluded that clinical expertise affords the types of complex decisions that result in well-conceptualized evidenced-based practice. Expertise is defined as a multifaceted construct that arises primarily from clinical and educational experiences. They wrote, “Expertise develops from clinical and scientific training, theoretical understanding, experience, self-reflection, knowledge of research, and continuing professional education and training” (Levant, 2005, pp. 10-11). Indeed, counseling psychology as a field adheres to a developmental model in which counselor decision making and expertise is expected to improve with training, supervision, and clinical experience (e.g., D. R. Evans, Hearn, Uhlemann, & Ivey, 1998; Goodyear et al., 2000; Hill, 2004; Stoltenberg, 1981; Stoltenberg et al., 1998).

One reason that conclusions about experience and judgment competence may differ is that research on clinical judgment is very extensive and has not yet been organized by comprehensive reviews. The most extensive review of clinical judgment literature is Garb’s (1998) book, *Studying the Clinician*; nonetheless, we found the identification of extant research to be incomplete. For example, out of the 75 empirical studies on experience we retrieved using an aggressive search strategy, Garb identified only 12. A second reason conclusions may differ is that clinician attributes, such as experience, are often studied only incidentally in the clinical judgment literature. Studies that do so are frequently overlooked in the debate. Third, with the exception of two recent meta-analyses on clinical versus statistical prediction (Ægisdóttir et al., 2006; Grove, Zald, Lebox, Snitz, & Nelson, 2000), no other area of clinical judgment research has been synthesized by meta-analytic techniques. Researchers have derived their conclusions about the effects of experience from traditional narrative reviews, which are subject to impressionistic biases (Cline, 1985; Dawes, 1994; Dumont, 1993; Faust, 1986; Garb, 1989, 1992, 1997, 1998; Lopez, 1989; Rabinowitz, 1993; Turk & Salovey, 1985; Wedding & Faust, 1989; Widiger & Spitzer, 1991). Narrative reviews consist of the scholar's selecting relevant literature and writing a report. Hunt (1997) concluded that narrative reviews are "the classic—and inadequate—solution" (p. 6) for the scientific scrutiny of a collective body of research because of the inevitable subjectivity that biases reviewers' impressions.

Purpose of This Meta-Analysis

This meta-analysis seeks to quantitatively establish if there is a relation between educational or clinical experience and clinical judgment accuracy and to statistically test assumptions commonly raised in the experience debate. The omission in the literature of a quantitative review of the relation between experience and judgment accuracy has, in our opinion, been most problematic, leading to the varying interpretations of existing research. Methodological pluralism (e.g., narrative, box score, meta-analytic reviews) is necessary for the synthesis of a scholarly field of study and absolutely necessary when there are divergent interpretations of the same data. Meta-analysis allows for the empirical synthesis of a body of research and, while not foolproof, provides a view of the research forest without getting lost in its trees (for an articulate description of meta-analysis, see Hunt, 1997). Methods used for meta-analysis typically consist of collecting study effect sizes, usually one from each study, and conducting analyses of the existing analyses (i.e., effects). Such an analysis of analyses is thus a meta-analysis. In the field of counseling and psychotherapy, meta-analysis has played a significant role in quantitatively establishing the effectiveness of psychotherapy (e.g., Smith & Glass, 1977; Smith, Glass, & Miller, 1980). Subsequent psychotherapy and counseling meta-analyses were able to further research well beyond initial questions of whether it worked (e.g., common treatment factors, Norcross, 2002; prescriptive treatments, Nathan & Gorman, 2002; and dose–effect relationships, M. J. Lambert, Hansen, & Finch, 2001). This study similarly seeks to address the lack of a quantitative analysis of the relationship between educational experience, clinical experience, and clinical judgment accuracy.

This study, which is part of a larger effort we have called the Meta-Analysis of Clinical Judgment (MACJ) project (Spengler et al., 2000), synthesizes complex findings about experience and judgment accuracy through the use of a comprehensive archival literature search and subsequent meta-analysis. Both clinical experience and education are treated in this study as forms of experience. Because they both involve the cumulative effect of cognitive processes that may affect clinical judgments, they are treated as independent variables. Clinical experience reflects professional experience with the client population or with methods of client assessment under study. Educational experience is represented by level of graduate training or amount of supervision. Although clinical and educational experiences are conceptually independent constructs, in practice they are often interrelated. Persons with more training may have seen more clients, for example. Thus, these two broad classes of experience are initially combined into a unifying construct in this study, followed by moderator tests for differences.

Judgments, of course, come in many varieties. We focused on perhaps the most basic: the accuracy with which a client was assessed. Accuracy, the dependent measure, was assessed using clinical decisions commonly studied by judgment researchers (e.g., diagnosis, behavior prediction). Some authors went to great lengths to determine an objective standard for judgment accuracy (e.g., neuropsychological findings verified by autopsy; Wedding, 1983). In other instances, the standard for accuracy was more tentative but still could be established based on logic, research findings, or practice standards (e.g., referral for antidepressant medications is an accurate recommendation for clients with bona fide, severe major depression; Meyerson, Moss, Belville, & Smith, 1979). In many studies, judgment accuracy could not be coded. For example, research on client variables (see Lopez, 1989) evaluates the impact of people characteristics (e.g., race, gender, age) on various clinical judgments (e.g., attractiveness, willingness to treat) but frequently fails to establish a standard for judgment accuracy (see Spengler, 1998). Because biases were studied in these cases, but not accuracy issues (for further discussion, see Lopez, 1989), we were unable to derive conclusions about judgment accuracy related to multicultural issues. When decision-making accuracy in these studies could not be established, they were not included in our analyses.

Studies varied in how they were designed; therefore, moderator variables were tested to clarify the nature of the relationship between experience and clinical judgment accuracy. These moderators were studied because they

(a) had been suggested by prior research as having likely relevance to judgment accuracy, (b) had been empirically examined, and/or (c) are typical methodological issues assessed by meta-analysis (e.g., study quality). In conducting this study, we sought to answer questions posed in the literature by assessing the moderating effects of (a) experience type (clinical or educational), (b) experience breadth (specific vs. general), (c) ecological validity of the method of study (in vivo, archival, analogue), (d) ecological validity of the stimuli (directly vs. indirectly experienced), (e) relation of experience to the research design (primary vs. supplementary analyses), (f) whether experience was treated as a major variable in the study's conceptualization, (g) whether feedback was given about accuracy, (h) criterion validity for accuracy-dependent measure, (i) study quality, (j) publication source, (k) year of publication or completion, and (l) type of judgment made (e.g., problem type, hit rate).

Our meta-analysis addresses the most fundamental of questions in the clinical judgment literature: Do judgments improve with experience? Accurate assessment of clients is a primary pathway to providing effective interventions (Meyer et al., 1998). If counseling, clinical, school, and other psychologists' decisions are inaccurate, counseling is likely to be ineffective (e.g., treatment failure; Wolfgang et al., 2006), and clients are less likely to follow through with treatment (e.g., premature dropout; Epperson, Bushway, & Warman, 1983). Failure to accurately identify client problems and make accurate predictions about future client behavior could have significant detrimental effects for the work engaged in by counseling psychologists. According to developmental models of training, counselors and psychologists should acquire increasingly sophisticated and veridical decision-making skills with experience. A more prevalent perspective in the clinical judgment literature, however, is that experience will not be associated with improved decision-making accuracy. This conclusion, stated numerous times in reviews of empirical work over several decades, runs counter to the strong intuitions (and experiences) of clinicians. Our hope is that this meta-analysis will clarify these questions and provide future direction for clinical judgment research, training, and practice.

Method

Because this manuscript is intended to be the most encompassing report from the MACJ project, with others citing this for the full methods, a description of the methods for the whole MACJ project is provided first.

Many variables (e.g., experience) are embedded within studies and are not the primary variables of interest to clinical judgment researchers; therefore, the need existed for a thorough search and examination of the clinical judgment literature, which was achieved by the MACJ project and is described next. Methods specific to the experience meta-analysis are described second.

MACJ Project

Research Team

Paul M. Spengler, Michael J. White, Stefanía Ægisdóttir, Alan S. Maugherman, six doctoral students, and one master's student collected, evaluated appropriateness for inclusion, and coded published and unpublished clinical judgment studies appearing from 1970 to 1996 for the MACJ project. A second cohort of four undergraduate and seven graduate assistants secured identified articles and entered data. The first two authors participated 10 years on the MACJ project to develop the archival database, the second two authors participated approximately 8 of those 10 years, and additional team members rotated in and out of the project.

Study Selection

The MACJ search of electronic databases (PsycINFO, ERIC, Dissertation Abstracts, BRS, MEDLINE, and Social Science Index) used 207 search terms.[2] By way of comparison, Garb (1994) conducted a computer search using 3 terms in conjunction with the term *clinical*: (a) statistical, (b) actuarial, and (c) incremental validity. Grove et al. (2000) used 18 search terms. In a review of client race, social class, and gender, Garb (1997) searched PsycINFO using 4 terms: clinical judgment and (a) sex, or (b) gender, or (c) race, or (d) social class. The MACJ search strategy was developed through an iterative process of reading the literature, adding new search terms through team consensus, and repeating electronic searches until no new studies were found. We chose this open-ended strategy to create an archival database and maximize the number of studies reviewed. Records of identified articles were maintained using database software. To limit the retrieval of studies to a manageable yet representative sample, studies that appeared between 1970 and 1996 were included in the search. Unavailable dissertations and journal articles were purchased, and authors were contacted to obtain difficult-to-retrieve material. After identifying likely studies, forward and backward cross-referencing occurred until no new studies were obtained. Using this search strategy, more than 35,000 potentially relevant articles were identified. After evaluating these articles (titles and abstracts), 4,617 were found to be varying types of judgment studies (e.g., mental health, possible mental health) and were coded for their content. Of these studies, 1,135 met our inclusion criteria (see below) and formed the data set for the MACJ project from which specific meta-analyses (e.g., experience) could be constructed.

Included Studies

Standardized inclusion and exclusion criteria were applied to all potential studies.[3] Each study was reviewed for appropriateness by the lead author and, at minimum, one additional team member. Discrepancies were resolved by consensus between the first four authors. For a study to be accepted, it had to meet several criteria. First, the focus of the study had to be clinical judgment, clinical judgmental bias, or clinical versus statistical prediction. Clinical judgment studies investigated judgments about diagnosis, prognosis, problem severity, type of treatment, and so on, or the cognitive processes used to form these judgments. Clinical judgmental bias studies examined bias secondary to client characteristics (e.g., age, race, ethnicity, socioeconomic status, gender, level of intelligence), judgment heuristics (e.g., anchoring, availability, representativeness, saliency, confirmatory hypothesis testing, primacy/recency effects, illusory correlation), and judgment errors (e.g., diagnostic overshadowing, underdiagnosing, overdiagnosing). Studies of clinical versus statistical prediction compared clinical judgment with a statistical formula, regression equation, or normative data (see Ægisdóttir et al., 2006). Second, the judgments studied had to be concerned with mental health issues, psychological constructs (e.g., healthy personality, vocational issues related to career counseling), or treatment issues. Studies of financial, legal, or other nonpsychological predictions were not included.[4] Third, the judges had to be mental health professionals or graduate-level trainees. This included psychologists, psychiatrists, social workers, counselors (school, rehabilitation, mental health, community, pastoral), psychiatric nurses, or graduate students in any of these fields. Fourth, the studies had to present sufficient data so that calculation of effect sizes was possible. In those cases in which dissertations were subsequently published, the published version was used. Finally, unpublished master's theses and conference presentations were excluded because of retrieval problems (e.g., poor response from authors or institutions). Applying these criteria reduced the 4,617 initially identified judgment studies to the 1,135 studies that comprise the archival data set for the entire MACJ project.

Training and Coding Procedures

Three sequential steps were used to code all studies (see each step below): (a) coding 122 study characteristics, (b) coding statistical information needed to calculate effect sizes, and (c) for each specialized meta-analysis, coding additional study characteristics for questions specific to that meta-analysis (e.g., clinical vs. statistical prediction; Ægisdóttir et al., 2006). All coding procedures were standardized and outlined in training and reference manuals with operational definitions provided for all variables.[5] Prior to actual coding, extensive training, discussion, and practice sessions were provided to new team members by the first, third, and and/or fourth authors. Training typically involved several weekly meetings until coders achieved 90% or better agreement with trainers' article coding. New trainees were initially taught to code clear exemplars (i.e., studies previously coded with no disagreement) followed by increasingly more difficult-to-code studies (i.e., those previously coded that required resolution of coding differences).

After each phase of coding (i.e., study characteristics, statistical information, specialized study characteristics), a study was reviewed a minimum of two additional times for coding accuracy by the first author and by one of the following three authors (for study characteristics and specialized characteristics) or by the second author and one of the other first four authors (for effect size calculation). Thus, the first four authors served as the core research team and were responsible for maintaining the standards for the selecting and coding of articles in the MACJ project. Coding discrepancies were resolved through additional reading and coding by members of the research team followed by discussion to achieve consensus. We chose a team consensus strategy, as this leads to "highly trustworthy codings" (H. Cooper, 1998, p. 96). The initial level of agreement for categorical coding, measured by Cohen's kappa, was moderate, $K = .54$. This level of agreement was not unexpected, given potential problems that exist when there are a large number of categories to code (e.g., low frequency use of some categories; Jones, Johnson, Butler, & Main, 1983).

Study characteristics coding. Each study was coded using a coding form with 122 characteristics. These were categorized under the conceptual groupings of *prediction processes* (clinical, model of clinical judgment, statistical), *judgment processes* (e.g., client variable bias, confirmatory hypothesis testing), *other manipulations* (e.g., amount of information, feedback), *stimulus material* (e.g., written case material, standardized case simulation), *client variables* (e.g., race, gender, age), *clinician individual differences* (e.g., clinical experience, educational experience), *judgment outcomes* (e.g., problem type, prognosis), *standard for accuracy* (e.g., standardized interview, a priori validation of clinical materials), *method of study* (analogue, archival, in vivo), and *type of design*. The authors, date, and journal outlet were also recorded. Data were entered into the database, which allowed for cross-tabulation of study characteristics and selection of those with variables of interest for a specific meta-analysis.

Effects coding. Statistics necessary for calculation of effect sizes, a global estimate of study quality (threats to validity), and when possible, determination of the direction of accurate judgment were collected in a second coding sequence. The team member who first recorded the effects made an initial coding on a written protocol. The second author then checked this coding prior to entering the values into the DSTAT meta-analysis program (Johnson, 1993). Values entered into DSTAT were themselves checked for accuracy by one of the first four authors. On some occasions, more information was obtained to establish what constituted an accurate judgment (e.g., detailed review of citations to establish construct validity; Morran, 1986).

Experience Meta-Analysis

Included and Excluded Studies

The MACJ project contained 316 studies coded for experience; of these, 106 established a standard for judgment accuracy. The sample was reduced to 75 studies with sufficient information to calculate effect sizes. These 75 studies yielded 91 effects for clinical or educational experience related to judgment accuracy by mental health professionals. Studies were excluded that did not provide enough information to make at least one accuracy comparison between two levels of experience or first-order correlation between experience and accuracy.

Specialized Study Characteristics

A coding form was constructed for the experience meta-analysis, based on additional variables of interest from the experience literature. Educational experience and clinical experience were coded as general or specific experience with operational definitions noted (e.g., year in graduate training, number of clients). Other codes included whether experience was part of the design or a planned major variable and whether feedback was given (see Table 1).

Independent Measure: Experience

As noted earlier, two types of experience were extracted from the literature.[6] Respondents varied on the first of these, clinical experience, according to (a) how many clients they had seen, (b) how long they had worked with them, (c) the number of tests administered, and (d) job setting. For example, Goldsmith and Schloss (1986) classified high-experienced school psychologists as having worked with 10 or more students with hearing impairments (over the past 3 years) and low-experienced school psychologists as having worked with 2 or fewer. Spengler, Strohmer, and Prout (1990) used a measure of the number of months worked with people who have mental retardation to assess rehabilitation counselors' experience. They also used a second measure of the number of clients worked with from this population. Related to number of tests administered, Carlin and Hewitt (1990) asked clinical psychology interns and clinical psychologists to rate their experience using the Minnesota Multiphasic Personality Inventory on a 5-point scale ranging from *no experience* (1) to *extensive experience* (5). Wedding (1991) similarly used an open-ended measure of the number of past administrations of the Halstead-Reitan Battery. In the case of job setting, Reiss and Szyszko (1983) classified experience with clients who have developmental disabilities as high (state developmental center), moderate (state mental health hospital), and low (clinical psychology trainees).

The second type of experience, educational, was measured according to (a) number of graduate courses taken, (b) year in graduate training (first, second, third, etc.), (c) level of training (master's, internship, doctoral, postdoctoral), (d) training intervention to improve clinical judgment, and (e) amount of clinical supervision. The most common method for measuring educational experience was by level of training. For example, Hillerbrand and Claiborn (1990) classified 15 novice (graduate counseling student) and 17 expert (licensed, employed) psychologists according to whether they were in training or licensed. Thompson and Hill (1991) assigned therapists to an experienced group if they were doctoral-level psychologists in an academic department or a counseling center and to a less experienced group if they were advanced doctoral students in counseling psychology. Others classified educational experience by how many courses had been taken (e.g., Batson & Marz, 1979) or graduate students' year in their training program (e.g., Rock & Bransford, 1992). Fewer studies investigated the impact of a training intervention as a form of educational experience. Lefkowitz (1973), for example, trained clinical judges to use a statistical formula to identify couples at risk for divorce and compared their predictive accuracy with that of an untrained group. Likewise, few studies investigated the impact of clinical supervision on judgment accuracy. Faust et al. (1988) used an open-ended question to assess hours of supervision in neuropsychology. Many studies provided more than one measure of experience. In these instances, we created a multiple measure of experience.

Dependent Measure: Judgment Accuracy

Several methods of defining judgment accuracy were used. Similar to our decision to initially combine educational and clinical experience into one independent variable, we initially combined studies with high- and low-criterion validity (representing the accuracy of a judgment) into one dependent measure. We identified high-criterion validity when an objective or highly valid criterion existed (e.g., a priori prediction of inpatient violence verified by subsequent patient behavior, Werner, Rose, & Yesavage, 1983; extensive a priori validation of written clinical vignettes, Spengler, Strohmer, et al., 1990; post hoc manipulation checks of clinical vignettes, Goldmith & Schloss, 1986). In other studies, judgment accuracy could be determined, but with less confidence (i.e., low-criterion validity), when the criteria were based on logic, professional consensus, or other less objective methods (e.g., referral to a physician to rule out medical causes before initiating psychological treatment of depression; American Psychiatric Association, 2000).

The most common judgment was about the client's problem(s), diagnosis, or symptom(s). The accuracy of these problem-type judgments was defined using valid criteria but not necessarily unequivocally valid criteria. For example, Walker and Lewine (1990) asked clinicians to determine, after viewing clients' childhood home movies, whether a client was preschizophrenic, using a forced-choice rating scale. All of the clients met diagnostic criteria for schizophrenia as validated by 100% agreement by two evaluators on the Schedule for Affective Disorders and Schizophrenia (Endicott & Spitzer, 1978). Hill, Thompson, and Corbett (1992) explored whether clinicians could accurately perceive clients' reactions (e.g., scared, worse, stuck) to therapy on the Client Reactions System (Hill, Helms, Spiegel, & Tichenor, 1988). A more precise, although less common, method of defining accuracy required judges to make repeated judgments with objective and reliable criteria for accuracy. In these judgments, where hit rates (e.g., percentage correct) could be calculated, there was always a high level of validity for classifying responses. Wedding (1983), for example, asked judges to classify responses on the Halsted-Reitan Test Battery as coming from either brain-damaged or psychiatric patients. Autopsy and other highly reliable medical data provided unequivocal evidence about the correct diagnosis.

A judgment related to problem type involved the severity or perceived magnitude of the client's problem or pathology. In a study by Benlifer and Kiesler (1972), judges viewed films of two children who were said to be in psychotherapy, even though this was true of only one. Judges then rated the psychiatric severity of the two children. Accurate responses assigned more severity to the child who was in therapy.

Accuracy of treatment was defined either by recommendation for treatment type or an evaluation of treatment need. Treatment types could involve psychological, psychiatric, or psychoeducational interventions that reflected accepted practices for particular conditions. As an example, Wilson and Gettinger (1989) asked school psychologists whether they would report a clearly described case of child abuse. The decision to report the case was considered to be more accurate. Alternatively, accuracy of treatment need examined the necessity of treatment, independent of modality chosen. The only study using this approach (Goldsmith & Schloss, 1986) asked judges to consider a phobic client who either had or did not have a hearing impairment. Accurate responses involved recommending treatment regardless of the client's hearing status.

Accuracy of prognosis was investigated in four studies. In one of these (Berman & Berman, 1984), judges rated the prognosis of a client who was labeled as either psychotic or normal. Poorer prognoses for the psychotic-labeled client were considered more accurate. Berman and Berman (1984) also assessed accuracy of adjustment. Here, accurate responses considered the psychotic-labeled client to be more poorly adjusted. There were two studies that assessed whether judges could accurately recall a client's problem. One of these, Loewy (1994), considered whether judges would be able to accurately recall information about clients identified as overweight. There were several measures of accuracy that could not be easily categorized. These included quality of decision reasoning (Hillerbrand & Claiborn, 1990), adherence to test interpretation standards (DeHaven, 1992), and amount of therapeutically useful information identified (Butcher, 1983). These disparate measures were classified as "other."

Effect Sizes

Meta-analyses capture a numerical index (i.e., an effect size) from every study and then combine these estimates to form a summary. Each effect size (such as a correlation coefficient, a standardized mean difference, or an odds ratio) measures the degree or magnitude of difference between two groups. In this study, an effect size was calculated (g, the difference between two means divided by the pooled standard deviation; Glass, McGaw, & Smith, 1981) for each of the studies that considered the influence of experience (clinical or educational) on clinical judgment accuracy. Where possible, the means and standard deviations for groups that were higher and lower in experience were used to make these computations using the DSTAT program. If these were unavailable, g was calculated using F, t, frequencies, or other statistics allowing nonconfounded comparisons for groups high and low in experience. Simple correlations (*r*) were used to calculate *g* in the case of studies using continuous measures of experience. The effect size *g* was then corrected for sample size bias by converting it to *d* (Hedges & Olkin, 1985). Many studies used multiple measures of judgment accuracy. In these cases, separate effects for each dependent variable were calculated and then averaged to form a single overall effect size (Marín-Martínez & Sánchez-Meca, 1999).[7] In some cases, this study effect size was composed of one type of accuracy, while in others it was based on multiple types. In the latter case, the overall effect was the average of the effect sizes across different measures. Multiple effect sizes were calculated only for studies using more than one sample of participants.

Study characteristics and effect sizes are shown in Table 1. Effect sizes are positive if greater experience is associated with enhanced accuracy and negative if greater experience is associated with reduced accuracy. A true zero effect indicates no difference between high and low levels of experience. As noted in Table 1, however, all zero effects from our sample of studies are inferred from results reported as statistically nonsignificant. Zero effects were used only when statistical information for computing effect sizes was unavailable. It should be noted that this strategy is likely to produce a conservative estimate of the effect sizes (i.e., to produce a bias in favor of the null hypothesis). In addition to the overall accuracy values, Table 1 contains effect sizes for all of the dependent variable measures of accuracy. The weighted mean effects for each of these variables, computed by weighting each study's effect by the reciprocal of its variance, are also shown (Hedges & Olkin, 1985).

Results

Overall Accuracy Analysis

An initial overall analysis was conducted using 75 effects. Note that this is fewer than the original 91 effects. Fifteen studies reported effects using multiple levels of one or more variables. For example, several studies used the same sample to assess both clinical experience's and educational experience's effect on accuracy. Because we considered each study the unit of analysis (H. Cooper, 1998), we did not treat the effects for clinical experience as separate from educational experience. Thus, a study that originally reported effects for clinical and educational experience would have a single effect size for a new variable labeled *both*. These combinations are noted in

(text continues on p. 371)

Table 1
Corrected Effect Sizes Between Experience and Accuracy (Combined Across All Categorical Variables) and Study Characteristics

		Study Effect Size	95% Confidence Limits for *d*				Judgment Accuracy Type							
Study	*n*	(*d*)[c]	Lower	Upper	Categorical Variables	Study Age	Problem	Hit Type	Treatment Rate	Severity	Prognosis	Adjustment	Other	Problem Recall
Alford & Locke (1984)	372	–0.146	–0.35	0.06	1,2,1,2,2,1,2,1,2,2,3,0	19				–0.146				
Amira, Abramowitz, & Gomes-Schwartz (1977)	217	0.229	–0.04	0.50	1,1,1,2,2,2,2,2,1,2,3,0	26			0.229					
Arkell (1976)	20	0.374	–0.51	1.26	2,1,1,2,2,1,2,2,2,2,0,6	27	0.374							
Barkin (1991)	91	–0.065	–0.48	0.35	1,2,1,1,4,2,2,2,3,4,1,0	12	–0.030					–0.099		
Batson & Marz (1979)	28	0.764	–0.01	1.54	2,1,1,1,2,1,2,2,1,2,0,1	24	0.764							
Beck & Ogloff (1995)	181	–0.178	–0.48	0.12	2,3,1,2,4,2,2,2,1,1,1,3	8			–0.177					
Benlifer & Kiesler (1972)	50	0.021	–0.53	0.58	1,1,1,1,2,1,2,1,1,2,5,0	31				0.359		–0.317		
Berman & Berman (1984)	30	–0.722	–1.46	0.02	2,1,1,2,1,1,2,2,1,2,3,0	19					–0.757	–1.037		
Berven (1985)	30	0.574	–0.16	1.30	2,1,1,2,2,1,2,1,1,1,0,6	18							0.574	
Blashfield, Sprock, Pinkston, & Hodgin (1985)	20	0.000	–0.88	0.88	2,1,2,2,2,1,2,1,1,3,0,3	18	0.000[a]							
Blumetti (1972)	14	–0.602	–1.67	0.47	1,1,2,2,2,1,2,2,2,4,3,0	31					–0.602			
Brenneis (1971)[b]	34	0.029	–0.65	0.71	3,2,1,2,2,1,2,1,1,2,3,0	32	0.029							

(continued)

Table 1 (continued)

			95% Confidence Limits for d				Judgment Accuracy Type							
Study	n	Study Effect Size (d)[c]	Lower	Upper	Categorical Variables	Study Age	Problem	Hit Type	Treatment Rate	Severity	Prognosis	Adjustment	Other	Problem Recall
ABruhn & Reed (1975)	106	–0.032	–0.41	0.35	2,1,1,2,1,2,2,1,2,2,0,3	28		–0.032						
Butcher (1983)	68	0.438	–0.04	0.92	2,1,1,3,2,1,2,1,1,4,0,6	20							0.438	
Caetano (1974)	113	–0.041	–0.44	0.35	2,1,3,1,2,1,2,2,1,2,5,0	29				–0.041				
Carlin & Hewitt (1990)	6	–0.829	–2.50	0.84	1,2,2,2,2,1,2,3,2,2,2,0	13		–0.829						
Carroll, Rosenberg, & Funke (1988)	20	0.032	–0.85	0.91	1,2,1,1,2,1,2,3,2,2,3,0	15	0.032							
Chandler (1970)	20	0.000	–0.88	0.88	1,2,1,2,2,1,2,2,2,1,2,0	33	0.000[a]							
R. P. Cooper & Werner (1990)	21	–1.159	–2.08	–0.23	1,2,1,2,2,1,2,2,2,2,3,0	13		–1.160						
Cressen (1975)	40	0.165	–0.46	0.79	2,3,1,2,2,1,2,2,2,2,3,3	28	0.165							
Dawson, Zeitz, & Wright (1989)	31	0.426	–0.29	1.14	1,2,1,2,2,1,2,2,2,2,3,3	14	0.852							0.000[a]
DeHaven (1992)	27	–0.099	–0.85	0.66	1,1,2,2,2,1,2,2,1,4,3,0	11							–0.099	
deMesquita (1992)	24	0.024	–0.78	0.83	2,1,1,2,2,1,1,2,2,2,6,7	11	0.024							
C. Evans (1983)	120	–0.067	–0.42	0.29	2,2,1,2,2,1,2,1,2,4,0,5	20	–0.067							
Faust et al., (1988)[b]	155	–0.028	–0.34	0.29	3,2,1,2,2,1,2,3,2,2,6,7	15	–0.028							
Finlayson & Koocher (1991)	180	0.059	–0.23	0.35	2,1,1,2,2,1,2,2,2,1,0,5	12	0.059							
Garcia (1993)	25	3.081	1.92	4.24	2,2,1,3,2,1,2,2,2,4,0,7	10	3.081							
Gaudette (1992)	6	0.002	–1.60	1.60	1,2,1,2,4,1,2,3,2,4,6,0	11		0.002						
Goldsmith & Schloss (1986)	100	0.072	–0.32	0.46	1,1,1,2,2,1,2,2,1,4,1,0	17	–0.010		0.154					
Goldstein, Deysach, & Kleinknecht (1973)	10	–0.274	–1.52	0.97	1,1,1,2,2,1,2,1,2,1,3,0	30		–0.274						

(continued)

Table 1 (continued)

Study	*n*	Study Effect Size (*d*)^c	95% Confidence Limits for *d*		Categorical Variables	Study Age	Judgment Accuracy Type								
			Lower	Upper			Problem	Hit Type	Treatment Rate	Severity	Prognosis	Adjustment	Other	Problem Recall
Graham (1971)	42	0.271	–0.34	0.88	2,2,1,2,1,1,2,2,2,1,5,0	32		0.271						
Gregory (1980)	60	–0.073	–0.58	0.43	2,1,1,2,2,1,2,1,2,4,0,3	23		–0.073						
Heaton, Smith, Lehman, & Vogt (1978)	9	0.499	–0.84	1.83	1,2,1,2,4,2,2,2,2,1,3,0	25		0.499						
Hill, Thompson, & Corbett (1992)	27	–0.091	–0.85	0.66	1,1,3,1,1,2,2,2,1,2,3,0	11	–0.091							
Hillerbrand & Claiborn (1990)	32	0.335	–0.36	1.03	2,1,1,2,2,1,2,1,1,2,3,3	13	0.000^a						0.502	
Homant & Kennedy (1985)	231	0.150	–0.11	0.41	1,2,1,2,2,1,2,1,1,1,5,0	18	0.150							
Horner, Guyer, & Kalter (1993)	22	0.238	–0.60	1.08	1,1,1,1,2,1,1,2,1,4,3,3	10				0.238				
Horowitz, Inouye, & Siegelman (1979)	20	1.090	0.15	2.03	2,1,1,1,2,1,2,2,1,1,0,3	24	1.090							
Howitt (1984)	12	–0.121	–1.25	1.01	1,2,1,2,2,1,2,2,1,4,3,0	19	–0.121							
Kendell (1973)	21	–0.097	–0.95	0.76	1,1,1,3,4,1,2,2,1,3,5,0	30	–0.097							
Kennel & Agresti (1995)	337	0.568	0.34	0.79	1,2,1,2,1,2,2,2,1,1,1,0	8			0.568					
Lacks & Newport (1980)	8	–0.027	–1.41	1.36	1,2,1,2,2,1,2,3,1,2,6,0	23		–0.026						
L. Lambert & Wertheimer (1988)^b	51	1.706	1.04	2.38	3,1,1,2,2,1,2,1,1,1,3,3	15	1.706							
Lefkowitz (1973)	24	0.130	–0.67	0.93	2,3,2,2,2,1,2,1,2,4,0,7	30		0.130						

(continued)

Table 1 (continued)

		Study Effect Size	95% Confidence Limits for *d*				Judgment Accuracy Type							
Study	*n*	(*d*)[c]	Lower	Upper	Categorical Variables	Study Age	Problem	Hit Type	Treatment Rate	Severity	Prognosis	Adjustment	Other	Problem Recall
Leli & Filskov (1981)	12	0.021	–1.11	1.15	2,1,2,2,3,1,2,3,2,2,0,3	22		0.020						
Leli & Filskov (1984)	10	0.046	–1.19	1.29	2,1,2,2,2,1,2,3,2,2,0,3	19		0.046						
Levenberg (1975)	360	0.158	–0.05	0.36	2,1,1,2,2,1,2,2,2,2,0,3	28	0.158							
Loewy (1994)	52	0.588	0.03	1.14	1,2,1,2,2,1,2,2,2,4,5,0	9								0.588
Lowell (1995)	25	0.123	–0.66	0.91	1,3,1,2,2,1,2,2,1,4,6,0	8	0.123							
Meyerson, Moss, Belville, & Smith (1979)	45	0.355	–0.25	0.96	1,2,3,1,2,1,2,1,1,3,3,0	24			0.356					
Millard & Evans (1983)	24	0.049	–0.75	0.85	2,1,1,2,2,1,2,1,2,2,5,3	20				0.049				
Morran (1986)[b]	40	0.498	–0.13	1.13	3,1,3,1,2,1,2,2,1,1,3,3	17							0.498	
Moxley (1973)	12	0.487	–0.73	1.70	2,2,1,2,2,1,2,2,2,2,0,5	33		0.487						
Patterson (1982)	64	0.151	–0.34	0.64	2,1,1,1,2,1,2,2,2,4,0,3	21	0.151							
Perez (1976)	4	–0.005	–1.96	1.96	2,1,2,2,2,1,2,2,1,2,3,0	27	–0.005							
Quinsey & Cyr (1986)	48	0.267	–0.30	0.83	2,1,1,2,2,1,2,1,1,2,0,3	17	0.267							
Reidy (1987)[b]	105	0.087	–0.30	0.47	3,3,1,1,4,2,2,2,2,4,5,0	16	0.087							
Reiss & Szyszko (1983)[b]	81	–0.016	–0.49	0.46	3,3,1,1,2,1,2,2,3,2,4,0	20	–0.032		0.000[a]					
Rock & Bransford (1992)[b]	16	0.421	–0.57	1.41	3,3,1,3,4,1,2,2,2,3,6,2	11	0.421							
Sandell (1988)	7	0.845	–0.72	2.41	2,1,1,2,2,1,2,2,2,2,0,3	15					0.845			
Schinka & Sines (1974)[b]	25	–0.221	–1.01	0.57	3,1,1,3,2,1,2,2,2,2,3,3	29							–0.221	
Seay (1991)	116	0.349	–0.04	0.73	1,2,1,2,2,1,2,2,1,4,4,0	12	0.349							
Silverberg (1975)	72	0.054	–0.41	0.52	1,1,1,2,2,1,2,2,1,4,5,0	28	0.054							
Spengler, Strohmer, & Prout (1990)	12	–0.187	–1.32	0.95	1,2,1,2,2,1,2,2,2,2,3,0	13	0.345		–0.718					
Starr (1987)[b]	46	0.150	–0.44	0.74	3,3,1,1,2,1,2,2,2,2,3,0	16		–0.150						

(continued)

Table 1 (continued)

Study	*n*	Study Effect Size (*d*)[c]	95% Confidence Limits for *d*		Categorical Variables	Judgment Accuracy Type								
			Lower	Upper		Study Age	Problem	Hit Type	Treatment Rate	Severity	Prognosis	Adjustment	Other	Problem Recall
Steiner (1977)[b]	32	0.480	–0.23	1.19	3,3,1,2,2,1,2,2,2,4,5,3	26	0.331						0.621	
Temerlin (1970)	70	0.028	–0.46	0.52	2,1,1,2,2,1,2,2,2,2,0,3	33	0.028							
Thompson & Hill (1991)	32	0.098	–0.60	0.79	2,1,1,1,2,2,2,3,2,2,0,3	12							0.098	
Twaites (1974)[b]	17	0.254	–0.70	1.21	1,3,1,2,2,1,2,2,2,4,3,0	29		0.254						
Van Ijzendoorn & Bus (1993)	20	0.341	–0.54	1.22	2,1,1,2,1,1,2,2,2,2,0,3	10					0.341			
Walker & Lewine (1990)	12	0.000	–1.13	1.13	2,1,1,1,4,2,2,2,2,3,5,3	13	0.000[a]							
Walters, White, & Greene (1988)	40	–0.037	–0.66	0.58	2,2,1,2,2,1,2,2,1,1,0,3	16	–0.037							
Wedding (1983)[b]	13	–0.152	–0.1.24	0.94	1,3,1,2,2,1,2,3,2,2,2,3	20		–0.152						
Werner, Rose, & Yesavage (1983)[b]	30	–0.233	–0.95	0.49	1,3,1,2,1,1,2,3,2,1,3,0	20		–0.233						
Wilson & Gettinger (1989)[b]	110	–0.029	–0.41	0.35	3,3,1,2,2,1,2,2,1,2,3,7	14			–0.029					
Overall *n* or *d*	4,607	0.128	0.07	0.19			0.151	–0.054	0.208	–0.058	–0.255	–.313	0.327	0.363
n of studies	75						37	17	8	5	4	3	8	2

Note: Positive effects indicate higher accuracy associated with increased experience. Overall *d* is the average of *d*s for studies as units of judgment accuracy corrected for sample size. Categorical variables and codes: Experience type (1 = *clinical*, 2 = *educational*, 3 = *both*), experience breadth (1 = *general*, 2 = *specific*, 3 = *both*), method of study (1 = *analogue*, 2 = *archival*, 3 = *in vivo*), ecology of stimulus presentation (1 = *directly experience*, 2 = *indirectly experienced*, 3 = *both*), relation of experience to design (1 = *not in design*, 2 = *in primary design*, 3 = *supplementary analysis*, 4 = *multiple*), experience as a major variable (1 = *yes*, 2 = *no*), accuracy feedback (1 = *yes*, 2 = *no*), study quality (1 = *acceptable*, 2 = *good*, 3 = *excellent*), accuracy criterion validity (1 = *low*, 2 = *high*, 3 = *both*), publication source (1 = *American Psychological Association journal*, 2 = *other psychology journal*, 3 = *psychiatry or medicine journal*, 4 = *dissertation*), measure of clinical experience (0 = *not applicable*, 1 = *number of clients*, 2 = *number of test administrations*, 3 = *time of counseling*, 4 = *job setting*, 5 = *other*, 6 = *multiple measures*), and measure of educational experience (0 = *not applicable*, 1 = *number of graduate courses*, 2 = *year of graduate training*, 3 = *level of training* [master's, doctoral, internship, postdoctoral], 4 = *time of face-to-face supervision*, 5 = *training intervention*, 6 = *other*, 7 = *multiple measures*). Between-participants designs are used in all studies.

a. Zero effect is inferred. The study reports statistically nonsignificant results.

b. When effect size estimates were provided for multiple categories of a nominal variable and the same sample was used (e.g., clinical and educational experience), effect sizes were combined and reported as both or multiple categories.

c. Corrected for sample size.

Table 1. As shown in Table 1, the weighted mean effect size was 0.128, indicating that more experienced judges were more accurate. Furthermore, the 95% confidence interval (CI) did not include zero (CI = .07 to .19). The CI represents the range of effect sizes expected to occur 95% of the time given the variance of the overall effect size estimate. If the CI had included zero there would be a chance that the true effect is zero, and we would not be able to say with confidence there is meaningful difference. Because the CI does not include zero, however, we can conclude that there is a reliable difference between more and less experienced clinicians (Johnson, 1993). Increased experience is associated with better clinical judgment accuracy, although the size of the effect is not large by conventional standards (Cohen, 1988).

Homogeneity test for overall effect. The initial overall effect was not consistent across studies. Using Hedges's (1982) homogeneity statistic, Q, the assumption that the sample was homogeneous was rejected, $Q(74) = 119.58$, $p = .000$. When a collection of studies has heterogeneous variance then the traditional approach to meta-analysis is to test moderator variables that may better model the data. One study, however, produced effects dramatically different from the others. Garcia (1993) had an effect of 3.08, indicating a very large positive effect for experience. When Garcia's effect was eliminated from the analysis (roughly 1% of the total number of studies and 0.5% of the total participants), the overall effect decreased to .121 (CI = .06 to .18) but resulted in a homogeneous sample, $Q(73) = 94.62$, $p = .09$. Removing this outlier slightly reduced the overall effect but produced a more trustworthy effect size estimate because variability within the set of studies was no greater than would be expected by chance. Again, the CI did not include zero, supporting a positive relationship between increased experience and accuracy. Table 2 shows the effect size distribution using a stem and leaf plot. As may be seen in Table 2, the distribution is positively skewed with a majority (67%) of the studies producing positive effects and a tail extending into positive values.

As noted by Hedges and Olkin (1985), there are no definitive guidelines for eliminating outliers. Estimates vary regarding how many observations should be eliminated but range from as few as 10% to as many as 50%. It is noteworthy here that removing only 1% of the studies achieved homogeneity. It can be argued that homogeneous effects such as these should not be further analyzed. After all, the most parsimonious explanation for any remaining variance is that it may be attributable to random causes (Eagly & Wood, 1994). Obtaining homogeneous results in the present study means that experienced clinicians

Table 2
Stem-and-Leaf Plot of Combined Effect Sizes for Experience and Overall Accuracy Effects

	Effect Size (*d*)	Summary Statistics	
Stem	Leaf	(After Outlier Removed)	
+3.0	8 (removed outlier)	Maximum	1.71
.		Quartile 3	0.37
+1.7	1	Median	+0.05
.		Quartile 1	–0.04
.		Minimum	–1.16
+1.0	9	Standard deviation	0.39
0.9		Mean (weighted for sample *n*)	0.121
0.8	5	*N*	74
0.7	6	*Proportion with positive sign	0.67
0.6			
0.5	0 0 7 7 9		
0.4	2 3 4 8 9		
0.3	4 4 5 6 7		
0.2	3 4 5 7 7		
0.1	0 2 3 5 5 5 5 6 7		
+.0	0 0 0 0 2 2 2 3 3 3 5 5 5 6 7 9		
–.0	1 2 3 3 3 3 4 4 7 7 7 9		
–0.1	0 0 2 5 8 9		
–0.2	2 3 7		
–0.3			
–0.4			
–0.5			
–0.6	0		
–0.7	2		
–0.8	3		
–0.9			
–1.0			
–1.1	6		

Note: Positive effects indicate higher accuracy associated with increased experience.

are simply more accurate than less experienced clinicians across a wide range of situations. Even so, others have observed that moderators of an apparently uniform effect should still be tested after homogeneity has been obtained. H. Cooper (1998), for one, argues that “if the meta-analyst has good theoretical or practical reasons for choosing moderators” further analyses are appropriate (p. 146). Because this is the first meta-analysis of the experience–accuracy effect, we chose to conduct planned moderator analyses. These analyses are grounded in a historical debate and were conducted for practical reasons, so researchers and practitioners can assess the veridicality of basic assumptions posed in the clinical judgment literature. Results from Garcia (1993) are not included in these analyses.

Exploratory Moderator Analyses

The overall effect size for experience was analyzed as a function of moderator variables (i.e., independent variables hypothesized to influence the overall effect size; see Table 3). These study design characteristics were assessed using a strategy similar to analysis of variance (Hedges & Olkin, 1985).[8] The equivalent of a main effect in this case is a between-classes effect analysis (Q_B). Each variable has more than one level. The exception to this is the analysis of the continuous variable, age of the study. The 95% CI and Q_W, or the test of homogeneity of the effect size within each class, are also reported. As may be seen in Table 3, only three of the categorical variables were associated with significant differences in the overall accuracy effect. Furthermore, 27 of the 38 categorical models tested did not show within-class heterogeneity, meaning little variability remained not accounted for by the respective variables.

Experience type. There were three categories of experience: clinical, educational, or both. No specific hypothesis was made because of differences in the literature on the effects of experience on judgment accuracy. Garb and Boyle (2003), for example, summarized research showing why it should be difficult to learn from clinical experience (e.g., due to clinical judgment biases, pseudoscience feedback), whereas Westen and Weinberger (2004) argued that clinical experience should produce advantages in decision making because of the immediate feedback clinicians receive. Regarding educational experience, Garb and Grove (2005) contended that “the value of training has been consistently demonstrated” (p. 658), whereas Faust (1986) maintained that general educational experience has no benefit. Effect sizes for judgment accuracy were found to be no different across different types of experience, $Q_B(2) = 0.70$, $p > .05$.

Experience breadth. Some studies considered whether clinicians had general experience or training, while others looked at whether the experience was specific to a particular client problem, treatment, or psychological construct that served as the criterion. For example, it has been found that the specific type of experience neuropsychologists have may make them better than other

Table 3
Categorical Models for Overall Accuracy Effects With Outlier Removed

Variable and Levels	Between-Class Effect (Q_B)	k	Mean Weighted Effect Size (d_{i+})^a	95% Confidence Interval for d_{i+}		Homogeneity Within Class (Q_{wi})^b
				Lower	Upper	
Experience type	0.70					
Clinical		31	+0.14	+0.05	+0.22	42.54
Educational		32	+0.09	0.00	+0.18	24.73
Both		11	+0.16	0.00	+0.31	26.66**
Experience breadth	2.35					
General		38	+0.15	+0.06	+0.24	46.62
Specific		23	+0.13	+0.04	+0.22	40.20*
Both		13	+0.02	–0.13	+0.17	5.45
Judgment type	16.49*					
Problem type		29	$+0.15^{a,b}$	+0.06	+0.24	34.11
Hit rate		17	$–0.02^{b,c}$	–0.21	+0.17	10.23
Treatment		5	$+0.24^{a}$	+0.11	+0.37	17.80**
Severity		4	$–0.10^{c}$	–0.27	+0.08	1.03
Prognosis		3	$+0.10^{a,b,c}$	–0.52	+0.73	2.81
Problem recall		1	$+0.59^{a}$	+0.03	+1.14	0.00
Other		6	$+0.28^{a,b}$	+0.01	+0.54	4.28
Combined^{c}		9	$+0.04^{a,b,c}$	–0.15	+0.22	7.86
Criterion validity	9.36*					
Low		31	$+0.22^{a}$	+0.13	+0.31	58.39**
High		41	$+0.04^{b}$	–0.04	+0.13	26.85
Both		2	$–0.04^{a, b}$	–0.36	+0.27	0.02
Feedback	4.83					
Yes		2	+0.13	–0.45	+0.71	0.13
No		72	+0.13	+0.07	+0.19	119.45**
Publication source	9.75*					
American Psychological Association		14	$+0.27^{a}$	+0.15	+0.38	43.98**
Other psychology		36	$+0.04^{b}$	–0.05	+0.12	27.96
Medical		5	$+0.18^{a, b}$	–0.20	+0.55	1.21
Dissertation		19	$+0.12^{a, b}$	0.00	+0.25	11.72
Method of study	1.44					
Analogue		62	+.013	+0.07	+0.19	88.23*
Archival		8	–0.10	–0.46	+0.27	2.05
In vivo		4	+0.14	–0.14	+0.41	2.90
Ecology of stimulus	0.34					
Direct		16	+0.13	–0.02	+0.27	11.08
Indirect		54	+0.12	+0.05	+0.18	80.53*
Both		4	+0.22	–0.13	+0.57	2.67

(continued)

Table 3 (continued)

Variable and Levels	Between-Class Effect (Q_B)	k	Mean Weighted Effect Size (d_{i+})[a]	95% Confidence Interval for d_{i+}: Lower	Upper	Homogeneity Within Class (Q_{wi})[b]
Experience in design	6.69					
Not in design		7	+0.29	+0.12	+0.45	18.89**
In primary design		58	+0.12	+0.05	+0.18	66.32
Supplementary		1	+0.02	–1.11	+1.15	0.00
Multiple		8	–0.04	–0.23	+0.15	2.72
Experience: Major variable	1.45					
Yes		64	+0.10	+0.03	+0.17	72.41
No		10	+0.19	+0.07	+0.30	20.76*
Study quality	3.12					
Acceptable		17	+0.09	–0.02	+0.20	34.86**
Good		47	+0.15	+0.08	+0.23	55.25
Excellent		10	–0.05	–0.28	+0.18	1.39

Note: Positive effects indicate higher accuracy associated with increased experience.
a. Effect sizes for variable levels not sharing superscripts are significantly different, $p < .05$, from other levels of that variable based on post hoc contrasts using χ^2.
b. Significance indicates that a level is not homogeneous.
c. The judgment type *combined* represents the average of two or more effect sizes for judgment type.
$^*p < .05$. $^{**}p < .01$.

psychologists at detecting brain impairment (e.g., Goldstein, Deysach, & Kleinknecht, 1973). Likewise, psychiatrists who have more specific training in psychotropic medications may do a better job than general physicians at monitoring therapeutic doses (e.g., Fairman, Drevets, Kreisman, & Teitelbaum, 1998). Based on behavior decision-making theory (Nisbett & Ross, 1980*),* we assumed that clinicians with specific experience would give greater weight to relevant data and, therefore, would be more accurate than clinicians with general forms of experience. This hypothesis, however, was tempered by widely cited findings that experts may be no more accurate, just more confident, in their ratings, compared with novices (e.g., Goldberg, 1959). Contrary to our expectations, specific experience with a judgment task was unrelated to accuracy, $Q_B(2) = 2.35$, $p > .05$.

Type of judgment made. We tested the type of decision made as a potential moderator of the experience–accuracy effect size. This overall effect was significant, $Q_B(7) = 16.49$, $p < .05$.[9] More experienced, compared with less experienced, clinicians were more accurate at diagnosing (d_{i+} = .15) and more accurate at formulating treatment recommendations consistent with practice guidelines (d_{i+} = 0.24). Accurate recall of problems was highest for the more experienced clinicians (d_{i+} = 0.59); this effect, however, is based on only one study. The “other” category had a moderate effect in favor of more experienced clinicians (d_{i+} = 0.28). The remaining judgment effects had CIs that include a zero effect, meaning these are not interpretable.

Criterion validity. We assigned a rating of high- and low-criterion validity for how the accuracy of the judgments that formed the dependent measures was established. We coded a high level of validity when a highly valid criterion existed. We assumed that effect sizes would be higher when an accurate judgment could be established by higher criterion validity (Meehl, 1959; Meyer et al., 1998). Contrary to this expectation, studies with low-criterion validity had higher effect sizes, $Q_B(2) = 9.36$, $p < .05$, although they also had greater nonrandom variability.

Feedback. Meehl (1954) noted years ago that if clinicians do not receive feedback about their decisions they cannot improve their accuracy (Dawes, Faust, & Meehl, 1989; Garb & Boyle, 2003; Lichtenberg, 1997). Most reviewers of the clinical judgment literature, after addressing the potential pitfalls in decision making, likewise emphasize the importance of receiving feedback for scientific practice and increased accuracy. Yet others have noted problems with the ambiguity of feedback often obtained in clinical practice (e.g., Barnum effects; Garb & Boyle, 2003) and the unlikely benefits of feedback under these conditions. Only two studies were identified in which feedback about accuracy was assessed related to the experience–accuracy effect. No difference was found between them, $Q_B(1) = 3.66$, $p > .05$.

Publication source. Several commentators have discussed the presence of a publication bias in favor of statistically significant results (Rosnow & Rosenthal, 1989). The nature of this bias is that independent of a study’s quality in design and execution, reviewers prefer studies with significant results to those with nonsignificant results. Because of competition for publication in major journals, such as those published by the American Psychological Association (APA), effects may be larger in them. We tested this assumption and found that studies published in non-APA psychology journals (d_{i+} = 0.04) had much smaller effects than studies found in APA journals (d_{i+} = 0.27), $Q_B(3) = 9.75$, $p < .05$. This finding raises the possibility that focusing on only one publication source may present a skewed picture of the relationship between experience and accuracy.

Ecological validity of method of study. Studies differed considerably in the methods they used to elicit judgments. Some used in vivo clients, while most others presented client information in an analogue format. A third strategy was to study clinical judgments from archival data (e.g., case notes). Based on the ecological validity of clinical judgment—that is, how representative the judgment task is of clinical practice (Rock, Bransford, Maisto, & Morey, 1987)—we assumed that more ecologically valid approaches (in vivo, archival) would be associated with larger effects, but this was not the case, $Q_B(2) = 1.44$, $p > .05$.

Ecological validity of stimulus. Studies varied in how stimuli were presented to respondents. In some, judges directly experienced the stimulus client (i.e., through audiotape, videotape, live presentation, role-playing, or standardized case simulation). In others, the presentation mode was indirectly experienced (i.e., through written case material, test protocols, or other means). Although one might think that greater experience should result in more accuracy when paired with the perceptually richer information of direct stimuli (Holt, 1958, 1970), the alternative may be true (Garb, 1984). Experience's effects may be most powerful with the more abstract and less salient material found in indirect means. Neither type of stimuli was found to affect judgment accuracy, $Q_B(2) = 0.34$, $p > .05$.

Relation of experience to design. Although many studies included experience as a component of the primary research design, others addressed experience in supplementary analyses. Because unplanned analyses may be more likely to capitalize on chance, the size of the effect in these studies may be larger than when included as part of the original design. Contrary to our expectation, no difference was found for design issues, $Q_B(3) = 6.69$, $p > .05$.

Experience as a major variable. Some studies similarly treated experience as a major variable in the original conceptualization of the study, while others did not. We coded experience as a planned major variable if it was included in the rationale for the study in the introduction, if it was described as a theoretical variable of importance, or if an a priori hypothesis was provided. Effect sizes in studies that did not emphasize experience as a major variable may have larger, but more serendipitous, findings. We accordingly coded and tested studies for this bivariate characteristic. No difference was found, $Q_B(1) = 1.45$, $p > .05$.

Study quality. H. Cooper (1998) provides a good discussion of using global judgments of research quality in meta-analyses. We chose this strategy because global judgments of study quality, as compared with multidimensional judgments, have been found to have better interjudge agreement and similar heuristic value in meta-analyses. We assigned a general rating for each study's quality. The basis of this rating was our consensus regarding how well a study was conceived, executed, and analyzed. Ratings were coded as acceptable, good, and excellent. There was no difference in overall effect, $Q_B(2) = 3.12$, $p > .05$.

Continuous Test for Overall Effect

Study age. Since the early 1970s, an expanding literature has addressed the role of social cognition in decision making (e.g., Kahneman & Tversky, 1973). Several suggestions have appeared describing how these findings might improve the process of clinical decision making (e.g., Arkes, 1981, 1991; Dawes et al., 1989; Faust, 1986; Garb, 1989, 1998; Spengler et al., 1995). To the degree that these suggestions have had an impact, we assumed that the link between experience and judgment accuracy should improve as studies become more recent. The continuous variable, age of study, was tested using Rosenthal's (1991) focused comparison of effect size. There was no relation between a study's age and effect size, $z = -.158$, $p = .875$.

Additional Analyses

A fail-safe analysis was conducted on the overall accuracy effects (Rosenthal, 1991). Using this analysis, 469 undiscovered zero-effect studies would be needed to reduce the statistically significant d_{overall} ($d = 0.12$, $p < .001$) to statistical nonsignificance. Given the efforts we made to secure a broad and comprehensive archival database of studies, and our use of the conservative meta-analysis strategy of coding zero effects when researchers provided no data but reported statistical nonsignificance, this seems unlikely. To put the size of the effect into perspective, we calculated the values for a binomial effect size display (BESD; Rosenthal & Rubin, 1982). As the name implies, a BESD converts the effect size into a clinical judgment with binomial predictor and criterion variables. In this study, it represents the expected accuracy for clinicians with two levels of experience (e.g., novice or experienced) who are classifying clients on a dichotomous outcome variable (e.g., correct or incorrect diagnosis). An assumption is also made that 50% of the cases fall into each category on both the independent variable and the dependent variable. Based on an effect size of $d = 0.12$ ($r = .06$), it is expected that novice clinicians will accurately classify 47% of the clients, $(.50 - r/2) \times 100$, whereas expert clinicians will correctly classify 53% of the clients, $(.50 + r/2) \times 100$. Thus, the overall effect represents an increase in accuracy of almost 13%, $(53 - 47)/47$.

Discussion

While the overall experience–accuracy effect of $d = 0.12$ is small using Cohen's (1988) convention, it is not trivial. First, this is a reliable effect because the CI does not cross zero. Second, it is homogeneous, indicating that the additional variables we studied are not necessary to explain the relation between experience and accuracy (cf. Eagly & Wood, 1994). Third, few differences were found for any of the moderator analyses; greater clinical or educational experience leads to almost 13% more accurate decisions regardless of most other factors. Finally, this effect is conceptually significant when viewed contextually within the decades of debate about the relation between experience and clinical judgment accuracy (e.g., Berven, 1985; Berven & Scofield, 1980; Brehmer, 1980; Brodsky, 1998; Dawes, 1994; Falvey & Hebert, 1992; Faust, 1986, 1994; Faust et al., 1988; Faust & Ziskin, 1988; Gambrill, 2005; Garb, 1989, 1998; Holt, 1970; Lichtenberg, 1997; Shanteau, 1988; Wedding, 1991; Wiggins, 1973; Ziskin, 1995). These findings may well chagrin experienced counseling psychologists who perceive their decision-making skills as significantly improving over the years, when the actual association is only modest (e.g., see Locke & Covell, 1997; Sakai & Nasserbakht, 1997; Stoltenberg, McNeill, & Crethar, 1994; Watkins, 1995). On the other hand, they should be somewhat mollified because there is a demonstrable relationship between experience, whether clinical or educational, and judgment accuracy. Furthermore, it is a reliable and positive relationship.

Relative Importance of the Experience–Accuracy Effect

The relative importance of the overall effect of $d = 0.12$ can be clarified by comparing it with effect sizes from related areas of research, by assessing its practical meaning, and by contextually placing these findings within the scholarly debate about experience and clinical judgment. By way of comparison, meta-analytic reviews of psychotherapy find little evidence for a relation between experience and client outcome (M. J. Lambert & Ogles, 2004). In many psychotherapy meta-analyses, no reliable difference is found in therapeutic effectiveness between the outcome of clients treated by trained and untrained therapists. M. J. Lambert and Ogles (2004) qualify these findings by noting several methodological weaknesses in the psychotherapy studies, including a limited range in therapist experience (e.g., $M = 2.9$ years; Stein & Lambert, 1984). We found that with increasing experience the decision abilities of counselors appear to modestly improve, something that may or may not be true for their therapeutic effectiveness.

Rosenthal (1990) noted that the practical significance of a small effect must also be considered when determining its overall meaning. For example, Rosenthal notes that an effect as small as $r^2 = .0011$ ($d = 0.066$, after conversion) is responsible for the medical advice commonly given at middle age to take a daily aspirin to reduce the risk of heart attacks. There are several examples from the present studies where mental health judgments could have arguably similar levels of consequence, for example, predictions of future dangerousness (e.g., R. P. Cooper & Werner, 1990; Werner et al., 1983), treatment decisions for complex cases (e.g., Sandell, 1988; Spengler, Strohmer, & Prout 1990), and decisions related to parental fitness for child custody (e.g., Horner, Guyer, & Kalter, 1993; Starr, 1987). Where decisions have a higher degree of importance, consumers of mental health services (e.g., clients, judges, hospital administrators, and custody litigants) may correctly assume that there is a practical gain achieved by having more experienced clinicians making these judgments. The modest gains in decision-making accuracy found in our study may not be unimportant when the stakes related to accuracy are at their highest.

Prentice and Miller (1992) also proposed that the practical meaning of small effects in the social sciences is determined by the impact they have on thinking in the field. The impact of the widely accepted belief that experience does not improve mental health clinical judgment accuracy has been significant. Ziskin (1995) wrote a three-volume guide for attorneys on how to cope with psychiatric and psychological testimony. Research on the lack of benefits of experience is selectively cited for attorneys so they can challenge the credibility of psychologists as expert witnesses. Brodsky (1999), in a guide for psychologists as expert witnesses, retracted his earlier position (Brodsky, 1991) that experience improves accuracy and concluded that "experience does not count" (p. 48). In his book for consumers, *House of Cards: Psychology and Psychotherapy Built on Myth*, Robyn Dawes (1994) based much of his argument that psychotherapy is ineffective on the premise that clinical judgment is significantly flawed. Given the nature of the tasks performed by professional psychologists, Dawes asserted that judgmental capabilities really cannot be expected to improve with experience. Our meta-analysis leads us to a different conclusion, namely, that experience marginally but reliably improves judgment accuracy. While the effect is small in size, it may—once again—not be unimportant when thinking about the potential impact on these discussions and future judgment research.

Implications for Clinical Judgment Research

Given the small experience–accuracy effect size, studies in this area (unless they use very large sample sizes) will tend to have low power, and a majority will likely report statistically nonsignificant associations between experience and accuracy. For example, studies with $n = 200$ per group (i.e., 200 expert clinicians and 200 novice clinicians) and an alpha of .05 will have a power of .22 to reject the null hypothesis with the presumed population effect of 0.12. This means that even for studies conducted with this atypically large *N*, only one study in four will detect a statistically significant association between experience and judgment accuracy. Even though the APA (2001) now encourages researchers to include effect sizes in their reports, this was not previously the case. In light of the historical overemphasis on statistical significance, it is not difficult to understand why scholars have concluded that experience does not count.

An important consideration, then, for future researchers is why the experience–accuracy effect is so small. Does experience really lead to such small improvement? Or is there some flaw in the design of these studies that is masking a true (and commonsense) association between experience and accuracy? One possibility may be that experience improves judgment accuracy but only under certain optimal conditions not investigated in these studies. Most of the studies evaluated experience as a main effect, and we tested the impact of moderator variables. It could be that a larger experience–accuracy effect requires essential mediating variables. For example, the most common reason proposed for why psychologists' judgment capabilities may not show larger magnitudes of improvement with experience is the lack of useful feedback obtained by practitioners about their judgments (e.g., Ericsson & Lehmann, 1996; Faust 1991; Garb, 1998; Lichtenberg, 1997). Spengler (1998) stated, "To ensure judgment accuracy as a local clinical scientist, some form of feedback mechanism is needed" (p. 932). Ericsson and Lehmann (1996) addressed the need for deliberate practice with feedback in order to benefit from experience. We found no differences in the experience–accuracy effect related to whether feedback was given. But only two studies investigated this moderator variable. The reason we speculate this occurred is because deliberate practice with feedback is a mediating (essential) rather than a moderating (augmenting) condition for learning to occur (an assumption we are unable to test by the current available data; see Frazier, Tix, & Barron, 2004).

Moreover, as Kenny, Kashy, and Bolger (1998) note, the absence of a treatment effect (i.e., the small experience–accuracy effect) may occur under conditions where there are multiple naturally occurring mediators producing inconsistent effects. Continuing with the above illustration, the use of feedback to improve accuracy may not be without challenges. Garb (1998) describes several ways in which feedback can be misleading. It may be, for example, that experience leads to clinicians' using more feedback, but experience may also lead to clinicians' becoming more prone to confirmatory hypothesis testing (e.g., Strohmer et al., 1990) or the tendency to seek selective evidence to support rather than refute their assumptions. These two mediators would thus negate each other and reduce the overall experience–accuracy effect. This is exactly the type of scenario that has been hypothesized by clinical judgment scholars (e.g., see Arkes, 1981, 1991) and leads to our recommendation that researchers abandon simple predictor–outcome relations in favor of path and mediation analyses of the experience–accuracy effect (see Frazier et al., 2004).

One of the greatest challenges to future experience–accuracy research is the inherent problem in defining experts in the clinical judgment literature (Strohmer & Spengler, 1993). Better operational definitions of experts, based on repeated performance related to outcomes rather than on peer nomination or reputation (e.g., Jennings & Skovholt, 1999; O'Byrne & Goodyear, 1997), the most common methods used, are needed. The novice–expert distinction is widely studied in counseling psychology investigations of clinical judgment processes (e.g., Cummings et al., 1990; Falvey & Hebert, 1992; Kivlighan & Quigley, 1991; Martin et al., 1989). Yet it is used all too often without connecting these judgment processes to accuracy outcomes (Orlinsky, 1999). Promising case simulation models used for teaching purposes have been formulated where the judgment processes of expert counselors are used as a benchmark for accuracy (e.g., Falvey & Hebert, 1992). Once again, the problem with this line of research is the inherent assumption, made without research support, that the judgment processes of expert counselors are better than those of novice counselors (cf. Lichtenberg, 1997).

In addition to our call for additional research on mediational models and for researchers to link decision-making processes to judgment outcomes, future research should study wider ranges of experience as well as developmental models of clinical decision making. Given counseling psychology's emphasis on counselor training and developmental models, the field is well positioned to take a leadership role in this area of research. Part of the reason for the small size of the experience effect as manifested in these studies may be that there is a ceiling effect for experience's influence on accuracy. Most studies compared groups of persons who had some experience, even though respondents had different amounts of it (e.g., master's students and PhDs).[10] Yet one study, L. Lambert and Wertheimer's (1988), compared persons with no clinical experience with those who had considerable experience. This study had one of the largest effects ($d = 1.71$). The issue of the limited range of experience for these cross-sectional comparisons has been a frequently stated limitation (Skovholt, Rønnestad, & Jennings, 1997) and warrants consideration in future research.

A more humbling possibility exists: Training and experience may only improve things modestly for the professional. To some degree, we are all experts on human behavior. Regardless of our professional training, we recognize and recall certain patterns of human behavior. Even in a game such as chess, where there is quick feedback and a restricted pattern to optimal solutions, it is only in recall for known patterns that experts do better (Newell & Simon, 1972). Psychologists deal with much more ambiguous behavior than chess players do, and therefore, the gap between expert and novice may be even smaller. Based on these observations, we speculate that the greatest magnitude of the experience–accuracy effect will be found in the no-experience to extensive-experience interval, not in the some-experience to more experience interval.

Skovholt et al. (1997) similarly argued that experience comparisons made in psychotherapy research studies are too limited in range to capture true novice–expert differences. They speculated that 10 to 15 years of clinical experience might be needed to develop expertise. After reviewing expert performance findings from chess, medicine, auditing, computer programming, bridge, physics, sports, typing, juggling, dance, and music, Ericsson and Lehmann (1996) concluded, “The highest levels of human performance in different domains can only be attained after around ten years of extended, daily amounts of deliberate practice activities” (p. 273). Despite extensive discussions on the role of cognition in the transition from novice to expert counselor (e.g., Etringer & Hillerbrand, 1995), there are no examples of longitudinal data on the development of clinical judgment to advance clinical judgment decision-making research. The Collaborative Research Network, a subgroup of the Society of Psychotherapy Research, might serve as a model for clinical judgment researchers (Rønnestad & Orlinsky, 2002). Both cross-sectional and longitudinal assessments have been used to study changes in therapist expertise. Measures of time, intensity, and variety have been sensitive in capturing anticipated relations between experience and psychotherapy expertise. We know of no similar studies of the development of clinical judgment expertise and make a call for such a research program to determine if and how clinicians develop expertise over time.

A final issue to consider is that experience may not be the best predictor of judgment accuracy. Psychotherapy researchers have found this to be so. Converging sources of psychotherapy research suggest that therapist individual effects (e.g., skill level) within treatments influence psychotherapy outcomes and that these effects are much larger than the experience–accuracy effect (APA, 2002). For example, Crits-Christoph et al. (1991) reported the results of a psychotherapy meta-analysis and found that therapist individual differences account for 5% to 9% of final outcomes (the corresponding *r*s of .22 to .30 convert to *d*s of 0.45 to 0.63). Luborsky et al. (1986) reanalyzed data from four major meta-analyses on psychotherapy outcomes and concluded that the effect of therapist differences accounted for more variance than differences between treatments. Thus, the best therapists appear to be more efficacious in the provision of counseling and psychotherapy, although they may not be the most experienced. In the clinical judgment literature, the same question has not been addressed: Are there individual differences among clinicians, aside from being a novice or expert, in their consistent level of accuracy over time and across clinical situations? Again, given counseling psychology's emphasis on process and outcome research, and in developing models for counselor training and supervision, the field is well positioned to take leadership in this aspect of judgment research.

Moderators of the Experience–Accuracy Effect

We sought tests of moderator variables based on our review of decades of debate in the literature. These findings warrant mention, after a caveat. The central finding, that the 74 effect sizes (excluding 1 outlier) are homogeneous, implies that all included studies estimate a common population effect size and that there is no extraneous variance to be explained (above and beyond that due to sampling error). Thus, an alternative explanation for the three significant moderator effects is that they result from capitalization on chance and any interpretation of these findings must be qualified by this possibility. One moderator that appears to have promise is the type of judgment made. Effect sizes were significantly higher for diagnoses ($d_{i+} = 0.15$) and treatment recommendations ($d_{i+} = 0.24$) made by experienced clinicians. A reasonable interpretation of these data is that experience improves the ability to predict behaviors and correctly classify client conditions. It also may lead to greater familiarity with best practice guidelines and increases the likelihood that clinicians will follow these guidelines (cf. Levant, 2005; Westen & Weinberger, 2004). A second finding that was contrary to our expectation is that experienced clinicians were better than inexperienced clinicians at forming judgments for accuracy measures with low-criterion validity ($d_{i+} = 0.22$), compared with a negligible effect for accuracy measures with high-criterion validity ($d_{i+} = 0.04$). This was unexpected because we anticipated that greater criterion validity would be associated with better predictions. This finding may reflect that experience improves accuracy the greatest where more refined and nuanced understanding is required (e.g., when the criterion is "fuzzy") or under conditions of greater uncertainty (cf. Nisbett & Ross, 1980).

We found several other potentially important moderators that were either understudied (e.g., feedback, problem recall) or not studied at all (e.g., developmental change, judgment processes linked to accuracy) so that we were unable to reach conclusions about them. More experienced clinicians, for example, may have better recall of problems ($d_{i+} = 0.59$), but this finding is based on just one study and requires replication. Another moderator variable we found understudied is the benefits of training interventions. Some studies referred to training, but they provided so little information about the method of training that we decided to not include this moderator variable in our analyses. Finlayson and Koocher (1991), for example, divided doctoral-level pediatric psychologists into an upper third with "extensive training" in sexual abuse and a lower third with "limited training" in this area of assessment. C. Evans (1983) classified 15 licensed clinical and school psychologists and 15 lay judges with no knowledge of the Draw-a-Person test as having high levels of training and no formal training.

A small number of more recent investigations by counseling psychologists suggest that graduate students can be taught effective clinical judgment strategies that lead to more accurate decisions (Meier, 1999; Spengler & Strohmer, 2001). This is one of the most important and yet underdeveloped aspects of the experience–accuracy judgment research. It is, again, an area that counseling psychologists are well positioned to address. Other counseling psychologists have evaluated the impact of training on decision-making process variables (e.g., Kurpius et al., 1985; Mayfield et al., 1999). Yet despite the burgeoning number of counseling psychology studies on judgment processes, we had no way of determining what constituted more accurate processes and could not analyze these studies. Counseling psychology's focus on process in clinical judgment and decision making has a long history (see Parker, 1958) that may become more informative with a shift in focus to which processes lead to more accurate clinical judgments.

Implications for Counseling Psychology Training and Practice

Our findings suggest that counseling psychologists should, until proven otherwise, exercise caution when attributing large gains in decision-making accuracy related to their educational and clinical experience. As a subspecialty that subscribes to growth-oriented developmental models, this admonition may not be one that is welcomed. Counseling and other psychologists usually engage in a rigorous training program that includes large amounts of clinical experience (e.g., practica, internship, and clinical supervision) before they can be licensed. Given the amount of time, money, effort, and training required for clinicians, these findings suggest that they do not receive much pay-off or benefit for their cost. From a training perspective, it may be more cost-effective to have students spend more time learning quantitative assessment methods (see Ægisdóttir et al., 2006), which can be learned relatively quickly (Spengler & Strohmer, 2001), rather than having them spend years engaged in supervised clinical experiences. However, we qualify this recommendation by our observations that additional research is needed on how to best educate and train experts in mental health decision making.

There are many recommendations for how to improve clinical judgment and decision making through the use of course curricula, practicum, and other aspects of graduate training in counseling psychology (e.g., Kurpius et al., 1985; Meier, 1999; Nurcombe & Fitzhenry-Coor, 1987; Spengler & Strohmer, 2001). Spengler et al. (1995) developed a scientist–practitioner model for assessment that incorporates methods of scientific hypothesis testing and debiasing techniques (methods to reduce bias) into counseling practice. Recent evaluations of teaching graduate-level students, following the Spengler et al. model, demonstrate promising changes in judgment accuracy for a wide variety of types of decisions (Meier, 1999; Spengler & Strohmer, 2001). While more pedagogical research is needed along these lines, counseling psychology's long adherence to the scientist–practitioner training model argues for teaching methods to improve clinical decision making in every training program.

Practicing counseling psychologists should also familiarize themselves with aspects of social cognition and behavior decision-making theories (e.g., Kahneman & Tversky, 1973; Nisbett & Ross, 1980) as well as the many specific recommendations for improving clinical decision making found throughout the clinical judgment literature (e.g., see Arkes, 1981, 1991; Faust 1986; Garb, 1989, 1998). In our estimation, one reason for clinical experience having no more impact on accuracy than it does is that the scientist–practitioner model has not been fully implemented into clinical practice. Much has been written about ways to function as a local clinical scientist (Stricker, 2000). Practicing counseling psychologists are urged to use methods that have been empirically demonstrated to increase judgment accuracy above and beyond experience. Two recent meta-analyses have shown that there is a benefit to using statistical over clinical prediction techniques (e.g., see Ægisdóttir et al., 2006; Grove et al., 2000) for some types of clinical decisions. Knowing that experience will increase accuracy only slightly should lead practicing counseling psychologists to more closely investigate the clinical judgment and scientist–practitioner literatures for ways to improve their decision making.

Conclusion

Meta-analysis is not a panacea for analyzing the clinical judgment literature. We had to make qualitative judgments while coding articles, which necessitated team discussion and resolution. Despite repeated cross-checking of coding and extensive training and checking for rater drift, coding is an inherently subjective process. Furthermore, our findings are only as good as the quality of the studies reviewed. We found that some important questions we had hoped to address in more depth could not be tested because of a lack of research (e.g., longitudinal change). Likewise, we could not analyze some studies because we could not establish what constituted an accurate judgment. The MACJ project took 10 years to collect, record, and code the studies that are now in the archival database. Our findings are limited to the time frame of 1970 to 1996 and, because of the use of fixed-effects analyses, to the MACJ archival data set.

What can be learned from this collection of cross-sectional studies? Contrary to the predominant viewpoint in the clinical judgment literature, we found that more experienced clinicians are reliably, but only modestly, more accurate than less experienced clinicians. The relatively small size of this effect argues that much needs to be done to understand how experience, along with other factors, influences the development of effective clinical decision making.

Notes

[1] Space does not permit a full exploration of the issues referred to in this frequently cited quotation by Holt (1970). Holt wrote extensively about the "unfair" comparisons of clinical prediction techniques with statistical prediction techniques. He argued (among other points) that experienced clinicians who are more familiar with a judgment task, and the setting from which the data are collected, would be the most accurate at clinical prediction. From Holt's perspective, clinicians had been pitted against the very best statistical formulas, so logically, the fairest comparison would be to have the very best clinicians compete, which he defined as those with these types of experience.

[2] Electronic search terms are available on request.

[3] Written criteria for inclusion and exclusion of studies are available on request.

[4] Our decision to exclude studies that focused on educational, financial, medical (nonpsychiatric), or other nonpsychological criteria distinguishes the Meta-Analysis of Clinical Judgment project from other reviews in which those findings have been generalized to clinical judgments about mental health and psychological issues (e.g., Grove et al., 2000).

[5] Training manuals for coding study characteristics and metrics are available on request.

[6] While our examples of experience measures include roughly 50% continuous and 50% categorical, in fact, it was much more common for researchers to use categorical measures ($k = 61$) than continuous measures ($k = 14$) of experience. Researchers categorized a naturally occurring continuous measure in only seven instances (noted here as categorical).

7. Marín-Martínez and Sánchez-Meca (1999) demonstrate that when effect sizes are homogeneous, or assumed to highly correlate, as is the case in the present meta-analysis, the arithmetic mean is equivalent to other procedures for aggregating effect sizes (e.g., Hedges & Olkin, 1985).

8. Fixed-effects analyses, rather than random-effects analyses, were used because of the archival nature of the data and the intent of our analyses to capture what has occurred in these studies. If the intent is to describe what was done in the sample of studies (i.e., to make a conditional inference), then a fixed-effects analysis is used. By contrast, if the intent is to make a statement about a population of studies from which the sample was drawn (i.e., to make an unconditional inference), then a random-effects analysis is used. Hedges and Vevea (1998) provide a pseudo–Monte Carlo analysis of the likelihood of either method rejecting the homogeneity hypothesis and find that the difference is not great.

9. Effect sizes for judgment-type accuracy reported in Table 1 represent mean weighted effect sizes for every identified judgment found in every study. These conceptually represent H. Cooper's (1998) concept of shifting unit of analysis (p. 100). The statistical analyses for judgment type reported in Table 3, however, are based on Cooper's more conservative strategy of assessing studies as units (pp. 98-99), which results in independent effect sizes (i.e., one per study).

10. The problem of quantifying range of experience in these studies became an intractable one because of the various units used to measure experience (e.g., master's vs. doctoral students, graduate students vs. professionals with varying years postdegree, fewer than 2 clients vs. 10 or more clients) and the lack of specificity (e.g., undefined amount of experience for novices vs. experts, median split of unspecified years of experience). For purposes of this discussion, we observed that the modal study focused on restricted ranges of experience.

References

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

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341-382.

*Alford, J. D., & Lock, B. J. (1984). Clinical responses to psychopathology of mentally retarded persons. *American Journal of Mental Deficiency*, *89*, 195-197.

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

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2002). Criteria for evaluating treatment guidelines. *American Psychologist, 57*, 1052-1059.

*Amira, S., Abramowitz, S. I., & Gomes-Schwartz, B. (1977). Socially-charged pupil and psychologist effects on psychoeducational decisions. *Journal of Special Education*, *11*, 433-440.

*Arkell, R. N. (1976). Naïve prediction of pathology from human figure drawings. *Journal of School Psychology*, *14*, 114-117.

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49*(3), 323-330.

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

*Barkin, S. L. (1991). *Judgments of professional psychologists about patients with the Acquired Immunodeficiency Disease (AIDS) virus and terminal cancer*. Unpublished doctoral dissertation, Hofstra University, Hempstead, NY.

*Batson, C. D., & Marz, B. (1979). Dispositional bias in trained therapists' diagnoses: Does it exist? *Journal of Applied Social Psychology*, *9*, 476-489.

*Beck, K. A., & Ogloff, R. P. (1995). Child abuse reporting in British Columbia: Psychologists' knowledge of and compliance with the reporting law. *Professional Psychology: Research and Practice, 26*, 245-251.

*Benlifer, V. E., & Kiesler, S. B. (1972). Psychotherapists' perceptions of adjustment and attraction toward children described as in therapy. *Journal of Experimental Research in Personality*, *6*, 169-177.

*Berman, G., & Berman, D. (1984). In the eyes of the beholder: Effects of psychiatric labels and training on clinical judgments. *Academic Psychology Bulletin*, *6*, 36-42.

*Berven, N. L. (1985). Reliability and validity of standardized case management simulations. *Journal of Counseling Psycholog*y, *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.

*Blashfield, R., Sprock, J., Pinkston, K., & Hodgin, J. (1985). Exemplar prototypes of personality disorder diagnoses. *Comprehensive Psychiatry, 26*, 11-21.

*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.

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

*Brenneis, B. (1971). Factors affecting diagnostic judgment of manifest dream content in schizophrenia. *Psychological Reports, 29*, 811-818.

Brodsky, S. L. (1991). *Testifying in court: Guidelines and maxims for testifying in court*. Washington, DC: American Psychological Association.

Brodsky, S. L. (1998). Forensic evaluation and testimony. In G. P. Koocher, J. C. Norcross, & S. S. Hill (Eds.), *Psychologists' desk reference*. New York: Oxford University Press.

Brodsky, S. L. (1999). *The expert expert witness: More maxims and guidelines for testifying in court*. Washington, DC: American Psychological Association.

*Bruhn, A. R., & Reed, M. R. (1975). Simulation of brain damage on the Bender-Gestalt Test by college students. *Journal of Personality Assessment, 39,* 244-255.

*Butcher, J. E. (1983). *Validation of a standardized simulation for the assessment of competence in mental health counselors.* Unpublished doctoral dissertation, Pennsylvania State University, University Park.

*Caetano, D. F. (1974). Labeling theory and the presumption of mental illness in diagnosis: An experimental design. *Journal of Health and Social Behavior*, *15*, 253-260.

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

*Carroll, N., Rosenberg, H., & Funke, S. (1988). Recognition of intoxication by alcohol counselors. *Journal of Substance Abuse Treatment, 5*, 239-246.

Casas, J. M., Brady, S., & Ponterotto, J. G. (1983). Sexual preference biases in counseling: An information processing approach. *Journal of Counseling Psychology, 30*, 139-145.

*Chandler, M. J. (1970). Self-awareness and its relation to other parameters of the clinical inference process. *Journal of Consulting and Clinical Psychology*, *35*, 258-264.

Cline, T. (1985). Clinical judgment in context: A review of situational factors in person perception during clinical interviews. *Journal of Child Psychology and Review, 26*, 369-380.

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

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

*Cooper, R. P., & Werner, P. D. (1990). Predicting violence in newly admitted inmates: A lens model analysis of staff decision making. *Criminal Justice and Behavior*, *17*, 431-447.

*Cressen, R. (1975). Artistic quality of drawings and judges' evaluations of the DAP. *Journal of Personality Assessment*, *39*, 132-137.

Crits-Christoph, P., Baranackie, K., Kurcias, J. S., Beck, A. T., Carroll, K., Pery, K., et al. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research, 2*, 81-91.

Cummings, A. L., Hallberg, E. T., Martin, J., Slemon, A., & Hiebert, B. (1990). Implications of counselor conceptualizations for counselor education. *Counselor Education and Supervision, 30*, 120-134.

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

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

*Dawson, V. L., Zeitz, C. M., & Wright, J. C. (1989). Expert-novice differences in person perception: Evidence of experts' sensitivities to the organization of behavior. *Social Cognition*, *7,* 1-30.

*DeHaven, C. L. (1992). *Educational diagnostic decision-making practices demonstrated by school psychologists, and a computer-based test interpretation system*. Unpublished doctoral dissertation, University of Southern California.

*deMesquita, P. B. (1992). Diagnostic problem solving of school psychologists: Scientific method or guesswork? *Journal of School Psychology, 30*, 269-291.

DiNardo, P. A. (1975). Social class and diagnostic suggestion as variables in clinical judgment. *Journal of Consulting and Clinical Psychology, 43*, 363-368.

Dumont, F. (1993). Inferential heuristics in clinical problem formulation: Selective review of their strengths and weaknesses. *Professional Psychology: Research and Practice, 24*, 196-205.

Eagly, A., & Wood, W. (1994). Using research syntheses to plan future research. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 485-500). New York: Russell Sage.

Einhorn, H. J. (1986). Accepting error to make less error. *Journal of Personality Assessment, 40,* 531-538.

Endicott, J., & Spitzer, R. L. (1978). A diagnostic interview: The Schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry, 25*, 837-844.

Epperson, D. L., Bushway, D. J., & Warman, R. E. (1983). Client self-terminations after one counseling session: Effects of problem recognition, counselor gender, and counselor experience. *Journal of Counseling Psychology, 30,* 307-315.

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.

Etringer, B. D., & Hillerbrand, E. (1995). The transition from novice to expert counselor. *Counselor Education and Supervision, 35*, 4-18.

*Evans, C. (1983). *"Draw a person . . . a whole person": Drawings from psychiatric patients and well-adjusted adults as judged by six traditional DAP indicators, licensed psychologists, and the general public*. Unpublished doctoral dissertation, Temple University, Philadelphia.

Evans, D. R., Hearn, M. T., Uhlemann, M. R., & Ivey, A. E. (1998). *Essential interviewing: A programmed approach to effective communication* (5th ed.). Belmont, CA: Thompson Books/Cole.

Fairman, K. A., Drevets, W. C., Kreisman, J. J., & Teitelbaum, F. (1998). Course of antidepressant treatment, drug type, and prescriber's specialty. *Psychiatric Services, 49*, 1180-1186.

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. Volume I: Matters of public interest* (pp. 185-217). Minneapolis: University of Minnesota Press.

Faust, D. (1994). Are there sufficient foundations for mental health experts to testify in court? No. In S. A. Kirk & S. D. Einbinder (Eds.), *Controversial issues in mental health* (pp. 196-201). Boston: Allyn & Bacon.

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

Faust, D., & Ziskin, J. (1988). The expert witness in psychology and psychiatry. *Science, 241*, 31-35.

*Finlayson, L. M., & Koocher, G. P. (1991). Professional judgment and child abuse reporting in sexual abuse cases. *Professional Psychology: Research and Practice, 22*, 464-472.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology, 51,* 115-134.

Gambrill, E. (2005). *Critical thinking in clinical practice: Improving the accuracy of judgments and decisions about clients* (2nd ed.). New York: John Wiley.

Garb, H. N. (1984). The incremental validity of information used in personality assessment. *Clinical Psychology Review, 4*, 641-655.

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

Garb, H. N. (1992). The *trained* psychologist as expert witness. *Clinical Psychology Review, 12*, 451-467.

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. (1997). Race bias, social class bias, and gender bias in clinical judgment. *Clinical Psychology: Science and Practice, 4*, 99-120.

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

Garb, H. H., & Boyle, P. A. (2003). Understanding why some clinicians use pseudoscientific methods: Findings from research on clinical judgment. In S. O. Lilienfeld, S. J. Lynn, & J. M. Lohr (Eds.), *Science and pseudoscience in clinical psychology* (pp. 17-38). New York: Guilford.

Garb, H. H., & Grove, W. M. (2005). On the merits of clinical judgment. *American Psychologist*, *60*, 658-659.

*Garcia, S. K. (1993). *Development of a methodology to differentiate between the physiological and psychological basis of panic attacks*. Unpublished doctoral dissertation, St. Mary's University, San Antonio, TX.

*Gaudette, M. D. (1992). *Clinical decision making in neuropsychology: Bootstrapping the neuropsychologist utilizing the Brunswik's Lens Model*. Unpublished doctoral dissertation, Indiana University of Pennsylvania, Indiana.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Goldberg, L. R. (1959). The effectiveness of clinicians' judgments: The diagnosis of organic brain damage from the Bender-Gestalt Test. *Journal of Consulting Psychology, 23*, 25-33.

*Goldsmith, L., & Schloss, P. J. (1986). Diagnostic overshadowing among school psychologists working with hearing-impaired learners. *American Annals of the Deaf*, *131*, 288-293.

*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.

Goodyear, R. K., Cortese, J. R., Guzzardo, C. R., Allison, R. D., Claiborn, C. D., & Packard, T. (2000). Factors, trends, and topics in the evolution of counseling psychology training. *The Counseling Psychologist, 28*(5), 603-621.

*Graham, J. R. (1971). Feedback and accuracy of clinical judgment from the MMPI. *Journal of Consulting and Clinical Psychology*, *36,* 286-291.

*Gregory, J. M. (1980). *Clinical judgment research tasks and the psychotherapist's ability to understand patients*. Unpublished doctoral dissertation, Duke University, Durham, NC.

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

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world* (2nd ed.). Thousand Oaks, CA: Sage.

*Heaton, R. K., Smith, H. H., Lehman, R. A. W., & Vogt, A. T. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology*, *46,* 892-900.

Hedges, L. V. (1982). Estimation of effect sizes from a series of independent studies. *Psychological Bulletin, 92,* 490-499.

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Hill, C. E. (2004). *Helping skills: Facilitating exploration, insight, and action* (2nd ed.). Washington, DC: American Psychological Association.

Hill, C. E., Helms, J. E., Spiegel, S. B., & Tichenor, V. (1988). Development of a system for categorizing client reactions to therapist interventions. *Journal of Counseling Psychology*, *35*, 27-36.

*Hill, C. E., Thompson, B. J., & Corbett, M. M. (1992). The impact of therapist ability to perceive displayed and hidden client reaction on immediate outcome in first sessions of brief therapy. *Psychotherapy Research, 2,* 143-155.

*Hillerbrand, E., & Claiborn, C. D. (1990). Examining reasoning skill differences between expert and novice counselors. *Journal of Counseling and Development, 68*, 684-691.

Holloway, E. L., & Wolleat, P. L. (1980). Relationship of counselor conceptual level to clinical hypothesis formation. *Journal of Counseling Psychology, 27*, 539-545.

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.

*Homant, R. J., & Kennedy, D. B. (1985). The effect of prior experience on expert witnesses' opinions. *Criminal Justice Review*, *10,* 18-21.

*Horner, T. M., Guyer, M. J., & Kalter, N. M. (1993). Clinical expertise and the assessment of child abuse. *Journal of the American Academy of Child & Adolescent Psychiatry*, *32,* 925-931.

*Horowitz, L. M., Inouye, D., & Siegelman, E. Y. (1979). On averaging judges' ratings to increase their correlation with an external criterion. *Journal of Consulting and Clinical Psychology, 47,* 453-458.

*Howitt, P. S. (1984). *Kinetic family drawings and clinical judgment: An evaluation of judges' ability to differentiate between the K-F-D's of abusing, control, and concerned mothers*. Unpublished doctoral dissertation, University of Windsor, Ontario, Canada.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York: Russell Sage.

Jennings, L., & Skovholt, T. M. (1999). The cognitive, emotional, and relational characteristics of master therapists. *Journal of Counseling Psychology, 46*, 3-11.

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

Jones, A. P., Johnson, L. A., Butler, M. C., & Main, D. S. (1983). Apples and oranges: An empirical comparison of commonly used indices of interrater agreement. *Academy of Management Journal, 26*, 507-519.

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

*Kendell, R. E. (1973). Psychiatric diagnoses: A study of how they are made. *British Journal of Psychiatry*, *122*, 437-445.

*Kennel, R. G., & Agresti, A. A. (1995). Effects of gender and age on psychologists' reporting of child sexual abuse. *Professional Psychology: Research and Practice*, *26*(6), 612-615.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 233-265). New York: Oxford University Press.

Kivlighan, D. M. Jr., & Quigley, S. T. (1991). Dimensions used by experienced and novice group therapists to conceptualize group processes. *Journal of Counseling Psychology, 38*, 415-423.

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

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.

*Lacks, P. B., & Newport, K. (1980). A comparison of scoring systems and level of scorer experience on the Bender-Gestalt test. *Journal of Personality Assessment, 44*, 351-357.

*Lambert, L., &Wertheimer, M. (1988). Is diagnostic ability related to relevant training and experience? *Professional Psychology: Research and Practice*, *19*, 50-52.

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., & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 139-193). New York: John Wiley.

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

*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*(3), 623-629.

*Leli, D. A., & Filskov, S. B. (1984). Clinical detection of intellectual deterioration associated with brain damage. *Journal of Clinical Psychology*, *40*(6), 1435-1441.

Levant, R. F. (2005, July 1). *Report of the 2005 Presidential Task Force on Evidence-Based Practice*. Retrieved September 15, 2005, from http://www.apa.org/about/president/initiatives.html

*Levenberg, S. B. (1975). Professional training, psychodiagnostic skill, and kinetic family drawings. *Journal of Personality Assessment*, *39*, 389-393.

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

Locke, T. F., & Covell, A. J. (1997). Characterizing expert psychologist behavior: Implications from selected expertise literature. *Educational Psychology Review, 9*, 239-249.

Loewy, M. I. (1994). Size bias by mental health professionals: Use of the illusory correlation paradigm*. Unpublished doctoral dissertation, University of California, Santa Barbara.

Loganbill, C., Hardy, E., & Delworth, U. (1982). Supervision: A conceptual model. *The Counseling Psychologist, 10*(11), 3-42.

Lopez, S. R. (1989). Patient variable biases in clinical judgment: Conceptual overview and methodological considerations. *Psychological Bulletin, 106*, 184-203.

*Lowell, E. S. (1995). *Cognitive strategies in psychodiagnosis.* Unpublished doctoral dissertation, University of Colorado at Boulder.

Luborsky, L., Crits-Christoph, P., McLellan, T., Woody, G., Piper, W., Liberman, B., et al. (1986). Do therapists vary much in their success? Findings from four outcome studies. *American Journal of Orthopsychiatry, 51*, 501-512.

Marín-Martínez, F., & Sánchez-Meca, J. (1999). Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *Spanish Journal of Psychology, 2,* 32-38.

Martin, J., Slemon, A. G., Hiebert, B., Hallberg, E. T., & Cummings, A. L. (1989). Conceptualizations of novice and experienced counselors. *Journal of Counseling Psychology, 36*, 395-400.

Mayfield, W. A., Kardash, C. M., & Kivlighan, D. M. (1999). Differences in experienced and novice counselors' knowledge structures about clients: Implications for case conceptualization. *Journal of Counseling Psychology, 46*, 504-514.

McPherson, R. H., Piseco, S., Elman, N. S., Crosbie-Burnett, M., & Sayger, T. V. (2000). Counseling psychology's ambivalent relationship with master's-level training. *The Counseling Psychologist, 28*(5), 687-700.

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

Meehl, P. E. (1959). Some ruminations on the validation of clinical procedures. *Canadian Journal of Psychology, 13*, 102-128.

Meier, S. T. (1999). Training the practitioner-scientist: Bridging case conceptualization, assessment, and intervention. *The Counseling Psychologist, 27*(6), 846-869.

Meyer, G. 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 1). Washington, DC: American Psychological Association.

*Meyerson, A. T., Moss, J. Z., Belville, R., & Smith, H. (1979). Influence of experience on major clinical decisions. *Archives of General Psychiatry, 36*, 423-427.

*Millard, R. W., & Evans, I. M. (1983). Clinical decisions processes and criteria for social validity. *Psychological Reports*, *53*, 775-778.

*Morran, D. K. (1986). Relationship of counselor self-talk and hypothesis formulation to performance level. *Journal of Counseling Psychology, 33,* 395-400.

*Moxley, A. W. (1973). Clinical judgment: The effects of statistical information. *Journal of Personality Assessment*, *37*, 86-91.

Nathan, P. E., & Gorman, J. M. (Eds.). (2002). *A guide to treatments that work* (2nd ed.). New York: Oxford University Press.

Newell, A., & Simon, P. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

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

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90,* 339-363.

Norcross, J. C. (Ed.). (2002). *Psychotherapy relationships that work: Therapist contributions and responsiveness to patient needs*. New York: Oxford University Press.

Nurcombe, B., & Fitzhenry-Coor, I. (1987). Diagnostic reasoning and treatment planning: I. Diagnosis. *Australian and New Zealand Journal of Psychiatry, 21*, 477-499.

O'Byrne, K. R., & Goodyear, R. K. (1997). Client assessment by novice and expert psychologists: A comparison of strategies. *Educational Psychology Review, 9*, 267-278.

Orlinsky, D. E. (1999).The master therapist: Ideal character or clinical fiction? Comments and questions on Jennings and Skovholt's "The cognitive, emotional, and relational characteristics of master therapists." *Journal of Counseling Psychology, 46*, 12-15.

Parker, C. A. (1958). As a clinician thinks. . . . *Journal of Counseling Psychology, 5*, 253-261.

*Patterson, D. R. (1982). *Social-cognitive biases in clinical judgment*. Unpublished doctoral dissertation, Florida State University, Tallahassee.

Pepinsky, J. 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.

Pfeiffer, A. M., Whelan, J. P., & Martin, J. M. (2000). Decision-making bias in psychotherapy: Effects of hypothesis source and accountability. *Journal of Counseling Psychology, 47*, 429-436.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*, 160-164.

*Quinsey, V. L., & Cyr, M. (1986). Perceived dangerousness and treatability of offenders: The effects of internal versus external attributions of crime casualty. *Journal of Interpersonal Violence*, *1*, 458-469.

Rabinowitz, J. (1993). Diagnostic reasoning and reliability: A review of the literature and a model of decision-making. *Journal of Mind and Behavior, 14*, 297-316.

*Reidy, K. (1987). *The effectiveness of the* DSM-3 *in reducing diagnostic errors with dual diagnosis cases.* Unpublished doctoral dissertation, University of New York, Buffalo.

*Reiss, S., & Szyszko, J. (1983). Diagnostic overshadowing and professional experience with mental retarded persons. *American Journal of Mental Deficiency, 87,* 396-402.

Ridley, C. R., Li, L. C., & Hill, C. L. (1998). Multicultural assessment: Reexamination, reconceptualization, and practical application. *The Counseling Psychologist, 26*(6), 827-910.

*Rock, D. L., & Bransford, J. D. (1992). An empirical evaluation of three components of the tetrahedron model of clinical judgment. *Journal of Nervous and Mental Disease, 180,* 560-565.

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

Rønnestad, M. H., & Orlinsky, D. (2002, June). The progression, struggle, stagnation and decline of psychotherapists: Trajectories of psychotherapist change and stability. In M. R. Rønnestad (Chair), *Research-based models of psychotherapists development*. Symposium presented at the meeting of the Society for Psychotherapy Research International Conference, Santa Barbara, CA.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*, 775-776.

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.

Sakai, P. S., & Nasserbakht, A. (1997). Counselor development and cognitive science models of expertise: Possible convergences and divergences. *Educational Psychology Review, 9*, 353-359.

*Sandell, R. (1988). A closer look at the ability to predict psychotherapeutic outcome. *Counselling Psychology Quarterly, 4,* 127-134.

*Schinka, J. A., & Sines, J. O. (1974). Correlates of accuracy in personality assessment. *Journal of Clinical Psychology, 30,* 374-377.

*Seay, O. J. (1991). *Major depression and mental retardation: Effects of psychologist workplace and level of mental retardation on diagnostic overshadowing*. Unpublished doctoral dissertation, University of Texas at Austin.

Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta Psychologica, 68*, 203-215.

*Silverberg, J. J. (1975). *Theoretical models of clinical decision-making: The discrimination of significant parameters of the judgmental process*. Unpublished doctoral dissertation, University of Iowa, Iowa City.

Skovholt, T. M., Rønnestad, M. H., & Jennings, L. (1997). Searching for expertise in counseling, psychotherapy, and professional psychology. *Educational Psychology Review, 9*, 361-369.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*, 752-760.

Smith, M. L., Glass, G. V., & Miller, T. L. (1980). *Benefits of psychotherapy*. Baltimore: Johns Hopkins University Press.

Spengler, P. M. (1998). Multicultural assessment and a scientist-practitioner model of psychological assessment. *The Counseling Psychologist, 26*(6), 930-938.

Spengler, P. M. (2000). Does vocational overshadowing even exist? A test of the robustness of the vocational overshadowing bias. *Journal of Counseling Psychology, 47*, 342-351.

Spengler, P. M., Blustein, D. L., & Strohmer, D. C. (1990). Diagnostic and treatment overshadowing of vocational problems by personal problems. *Journal of Counseling Psychology, 37*, 372-381.

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

Spengler, P. M., Strohmer, D. C., 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., Strohmer, D. C., & Prout, H. T. (1990). Testing the robustness of the diagnostic overshadowing bias. *American Journal on Mental Retardation, 95,* 204-214.

Spengler, P. M., White, M. J., Maugherman, A., Ægisdóttir, S., Anderson, L., Rush, J., et al. (2000, August). *Mental health clinical judgment meta-analytic project: Summary 1970-1996*. Paper presented at the meeting of the American Psychological Association, Washington, DC.

*Starr, R. H., Jr. (1987). Clinical judgment of abuse-proneness based on parent-child interactions. *Child Abuse and Neglect, 11*, 87-92.

Stein, D. M., & Lambert, M. J. (1984). On the relationship between therapist experience and psychotherapy outcome. *Clinical Psychology Review, 4*, 1-16.

Steiner, C. (1977). The effect of experience, training and cue-availability on clinical judgment.* Unpublished doctoral dissertation, Adelphi University, Garden City, NY.

Stoltenberg, C. D. (1981). Approaching supervision from a developmental perspective: The counselor complexity model. *Journal of Counseling Psychology, 28,* 59-65.

Stoltenberg, C. D., McNeill, B. W., & Crethar, H. C. (1994). Changes in supervision as counselors and therapists gain experience: A review. *Professional Psychology: Research and Practice, 25*, 416-449.

Stoltenberg, C. D., McNeill, B. W., & Delworth, U. (1998). *IDM supervision: An integrated developmental model for supervising counselors and therapists*. San Francisco: Jossey-Bass.

Stricker, G. (2000). The scientist-practitioner model: Gandhi was right again. *American Psychologist, 55,* 253-254.

Strohmer, D. C., Shivy, V. A., & Chiodo, A. L. (1990). Information processing strategies in counselor hypothesis testing: The role of selective memory and expectancy. *Journal of Counseling Psychology, 37*, 465-472.

Strohmer, D. C., & Spengler, P. M. (1993). Studying mental health counselor clinical judgment: A response to Falvey and colleagues. *Journal of Mental Health Counseling, 15,* 465-474.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Sciences in the Public Interest* (Suppl. to *Psychological Science*), *1*(1).

*Temerlin, M. K. (1970). Diagnostic bias in community mental health. *Community Mental Health Journal, 6*, 110-117.

Thompson, B. J., & Hill, C. E. (1991). Therapist perceptions of client reactions.* *Journal of Counseling and Development, 69*, 261-265.

Turk, C. T., & Salovey, P. (1985). Cognitive structures, cognitive processes, and cognitive-behavior modification: II. Judgments and inferences of the clinician. *Cognitive Therapy and Research, 9*, 19-33.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

Twaites, T. N. (1974). The relationship of confidence to accuracy in clinical prediction.* Unpublished doctoral dissertation, University of Minnesota at Minneapolis.

*Van Ijzendoorn, W. J. E., & Bus, A. G. (1993). How valid are experts' prognoses on children with learning problems? *Journal of School Psychology, 31*, 317-325.

*Walker, E., & Lewine, R. J. (1990). Prediction of adult onset schizophrenia from childhood home movies of the patients. *American Journal of Psychiatry*, *147*, 1052-1056.

*Walters, G. D., White, T. W., & Greene, R. L. (1988). Use of the MMPI to identify malingering and exaggeration of psychiatric symptomatology in male prison inmates. *Journal of Consulting and Clinical Psychology, 56,* 111-117.

Wampold, B. E., Casas, J. M., & Atkinson, D. R. (1981). Ethic bias in counseling: An information processing approach. *Journal of Counseling Psychology, 28*, 498-503.

Watkins, C. E., Jr. (1995). Psychotherapy supervisor and supervisee: Developmental models and research nine years later. *Clinical Psychology Review, 15*, 647-680.

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

Wedding, D. (1991). Clinical judgment in forensic neuropsychology: A comment on the risks of claiming more than can be delivered. *Neuropsychology Review, 2*, 233-239.

Wedding, D., & Faust, D. (1989). Clinical judgment and decision making in neuropsychology. *Archives of Clinical Neuropsychology, 4*, 233-265.

*Werner, P. D., Rose, T. L., & Yesavage, J. A. (1983). Reliability, accuracy, and decision-making strategy in clinical predictions of imminent dangerousness. *Journal of Consulting and Clinical Psychology*, *51*, 815-825.

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

Westen, D., & Weinberger, J. (2005). Clinical judgment in science. *American Psychologist, 60*, 659-661.

Widiger, T. A., & Spitzer, R. L. (1991). Sex bias in the diagnosis of personality disorders: Conceptual and methodological issues. *Clinical Psychology Review, 11*, 1-22.

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

*Wilson, C. A., & Gettinger, M. (1989). Determinants of child-abuse reporting among Wisconsin school psychologists. *Professional School Psychology, 4*, 91-102.

Wisch, A. F., & Mahalik, J. R. (1999). Male therapists' clinical bias: Influence of client gender roles and therapist gender role conflict. *Journal of Counseling Psychology, 46*, 51-60.

Wolfgang, L., Lambert, M. J., Harmon, S. C., Tschitsaz, A., Schurch, E., & Stulz, N. (2006). The probability of treatment success, failure and duration—What can be learned from empirical data to support decision making in clinical practice? *Clinical Psychology and Psychotherapy, 13,* 223-232.

Ziskin, J. (1995). *Coping with psychiatric and psychological testimony* (5th ed., Vols. 1-3). Los Angeles: Law and Psychology Press.

Paul M. Spengler is Associate Professor in the Department of Counseling Psychology and Guidance Services at Ball State University, Muncie, IN. He received his doctorate from The State University of New York, University at Albany, in 1991. He is the principal investigator on the federally funded Meta-Analysis of Clinical Judgment Project. He publishes primarily in the areas of psychological assessment and clinical judgment and teaches graduate courses in assessment and advanced counseling and psychotherapy. He maintains a private practice, has been on editorial boards for JCP, TCP, and JMHC, and currently serves as Associate Editor for The Counseling Psychologist.

Michael J. White received his PhD in 1975 from The Pennsylvania State University where he studied social psychology. He is currently a professor in the Department of Counseling Psychology & Guidance Services at Ball State University and is director of the MA program in social psychology. His current research interests include implicit attitudes and stereotyping.

Stefanía Ægisdóttir received her PhD in counseling psychology from Ball State University in 2000. Since 2002 she has been an assistant professor in the Department of Counseling Psychology and Guidance Services at the same university. Her research involves cross-cultural and international psychology, counseling attitudes and expectations, and measurement issues in psychology.

Alan S. Maugherman received his PhD in counseling psychology from Ball State University in 1999. He received the Ball State University Distinguished Dissertation award. Since graduation, he has been in private practice in Muncie, Indiana, where he specializes in psychotherapy and testing with adolescents. He also teaches classes at Ball State University and is the director of a summer program for adolescent rock musicians.

Linda A. Anderson is a licensed psychologist at Counseling and Psychological Services, Oregon State University, where she is also Coordinator of Sexual Assault Support Services. Her areas of specialization include provision of clinical services, training, and educational programming about sexual assault response and prevention. She received her PhD in counseling psychology from Indiana University Bloomington, and her MA in counseling from Ball State University.

Robert S. Cook, PhD, is Director of Clinical Services at the Center for Persons with Disabilities, Utah State University. He earned his PhD in Counseling Pyschology fom Ball State University in 1996. Dr. Cook specializes in the evaluation and treatment of children and adolescents.

Cassandra Nichols completed her doctorate in Counseling Psychology at Ball State University in 1996. She completed a pre-doctoral internship and postdoctoral training at the University of Utah Counseling Center. She currently is the Associate Director and Clinical Director at the Counseling Services at Washington State University. She is also the Sexual Assault Response Coordinator for the university and a faculty appoint with the Department of Educational Leadership and Counseling Psychology. She is also in private practice specializing in psychotherapy of adults, couples and adolescents.

Georgios K. Lampropoulos has a graduate degree in psychology from the University of Crete (Greece) and a PhD in counseling psychology from Ball State University (2006). He has guest-edited three special issues in psychotherapy journals and has over 25 publications in the area of psychotherapy. He was the recipient of a doctoral scholarship from the Alexander S. Onassis Public Benefit Foundation (Athens, Greece), and has received four student awards from Divisions 24, 29, and 32 of the American Psychological Association. He is currently a postdoctoral fellow at the Department of Psychology at Penn State University.

Blain S. Walker received his PhD from Ball State University in 1999. He is currently a licensed psychologist for the U.S. Army and is participating in a health psychology fellowship at Tripler Army Medical Center. At the time of publication, he was deployed to Iraq.

Genna Cohen is the Research Director for Louisiana Professional Academy, Inc. She received her MA in Counseling and Social Psychology from Ball State University in 1999. For the last six years, she has been part of a research team that is quantifying sex offender treatment and risk assessment in 27 sites across Louisiana. Her contributing effort has assisted the Louisiana Department of Public Safety and Corrections to earn the American Correctional Association's esteemed Exemplary Program Award in February, 2006.

Jeffrey Rush earned his PhD from Ball State University in 2004. He is currently a counselor at Logan River Academy in Logan Utah.